DEVANAGARI VIP TEAM ISSUES REPORT

A WHITE PAPER

DEVANAGARI VIP GROUP

Contents

0. PRELIMINARIES

a. BACKGROUND AND OVERVIEW[1]

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 the web 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 Indian (see Appendix I). The formulation of a policy document for providing 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 accepted.

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 (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 II) which embraced not only Hindi but other major languages using the Devanāgarī script (cf. Appendix I). 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 pre- and post-Singapore, 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.

This document is a report which attempts to lay down the background to writing system along with the various issues for the creation of Internationalized Domain Names in Indian 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.

[1] This section has been contributed by GIST Group. CDAC

b. STRUCTURE

The report, whose basic layout was finalized at the Pune meet, comprises the following sections:

Part I 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, legal and administrative, these issues are highlighted in a sequential order. Part II is an inventory of the major issues pertinent to the topic in question and examines the problems from all angles. Since the Registry plays an important role in IDN, a special section, Part III is devoted to this area. Finally since implementation of the policy will have fiscal, legal and societal implications, the administrative policy in the light of issues will be treated in Part IV.

A certain number of Appendices which provide ancillary information and also treat of the issues of definitions and questions raised at the Singapore meet, complete the report.

The document *IDN Variant TLDs - Lists of Issues - v06 redline* received on 24th August 2011 is under deliberation. However a majority of issues raised are handled in this draft report.

1. DEVANĀGARĪ : AN OVERVIEW[2]

This over-view of Devanāgarī is a linguistic introduction to Devanāgarī . It starts off with the historical evolution of Devanāgarī and in section 1.2 studies the structure of Devanāgarī . Section 3 develops the notion of the underlying nuclear element: the akshar and further proposes and akshar typology. IPA as well as simple transliteration has been used as a guide to the pronunciation of the examples.

1.1. Devanāgarī: A Historical Perspective

Devanāgarī (pronounced [d̪eːʋˈnaːɡriː])is the main script for the Indo-Aryan languages Hindi, Marathi, Maithili and Nepali recognized as official languages of the Republic of India. 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 Dogri, Kashmiri (by Hindu speakers), Sindhi and Santali. It is rising in use for speakers of tribal languages of Arunachal Pradesh, Bihar and Andaman & Nicobar Islands. Devanāgarī can be easily shown to be related to the modern scripts used for other Indian languages such as Gujarati, Gurumukhi (for Punjabi), and Assamese/ Bengali, as well as to the scripts used for Dravidian languages, such as Tamil, Telugu, Kannada and Malayalam.

It is now 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 Princep 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 spite of the diversified character of Brāhmī-derived scripts, they have a common structure. An understanding of the structure of Devanāgarī , or for that matter of any of the scripts derived from Brāhmī, is of general interest for this group of scripts of south and southeast Asia.

[2] This section has been contributed by Dr Pramod Pandey with additions by GIST Group. CDAC

1.2. The structure of written Devanāgarī

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

1.2.1. 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

1.2.2. The Implicit Vowel Killer: Halanta

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)

1.2.3. 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, Konkani also admit 2 vowels and their corresponding Mātrās as in

ऍ ◌ॅ ऑ ◌ॉ

ऍण्ड /and/ ऑर /or/

Marathi replaces the ऍ by अॅ

1.2.4. 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]

1.2.5. Nasalisation: Chandrabindu

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

1.2.6. Nukta ◌़[3]

Mainly used in Hindi, the Nukta sign is placed below a certain number of consonants to represent words borrowed from a Perso-Arabic loan. It can be adjuncted 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

1.2.7. Visarg ◌ः and Avagrah ऽ

The Visarg is frequently used in Sanskrit and represents a sound very close to /h/. दुःख

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

1.2.8. Nasalisation: Chandrabindu

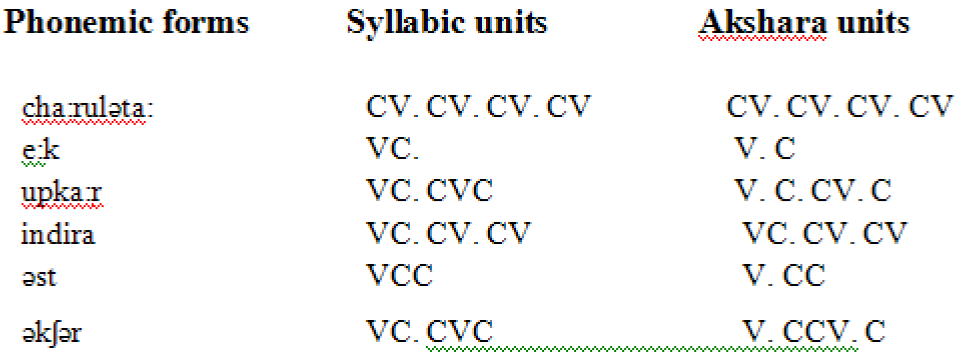
◌ँ Chandrabindu/Anunasika denotes nasalization of the preceding vowel as in आँख (eye). Present-day Hindi users tend to replace the chandrabindu by the anuswāra

In Parts 3 and 4, it will be shown how this classification of Devanāgarī characters can be reduced to a "compositional grammar" and reduced to a formalism which ensures the well-formedness of the akshar.

[3] The Nukta will be treated at length in the section of Normalisation, since Unicode allows the characters mentioned above to be represented in two different ways: as a single character or a consonant+the nukta

1.3. The Nodal Unit: akshar

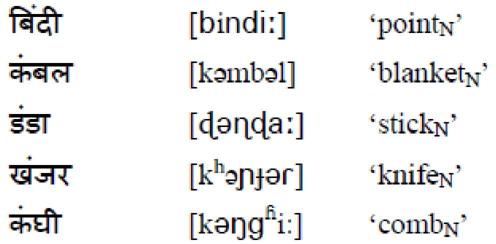
The *akshar* is the graphemic unit of Devanāgarī. The difference between the syllable and the akshar is that while the syllable includes one or more post-vocalic consonants, the akshar doesn't, as can be seen below:



Phonemic forms	Syllabic units	Akshara units
cha:ruləta:	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
əst	VCC	V. CC
əkʃər	VC. CVC	V. CCV. C

Table 1: Syllabic and akshara divisions of spoken forms

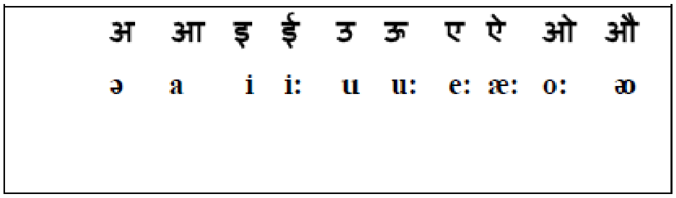
The only exception to the generalization about the post-vocalic consonants vis-à-vis akshars is the anuswāra, the underlying nasal consonant surfacing as homorganic with the following stop. The anuswāra is treated as a part of the grapheme. The orthographic and phonetic transcriptions of forms with the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



बिंदी	[bindi:]	'point$_N$'
कंबल	[kəmbəl]	'blanket$_N$'
डंडा	[ɖəɳɖa:]	'stick$_N$'
खंजर	[k^{h}əɲɟər]	'knife$_N$'
कंघी	[kəŋg^{ɦ}i:]	'comb$_N$'

Table 2: Representation of anuswāra in Devanāgarī

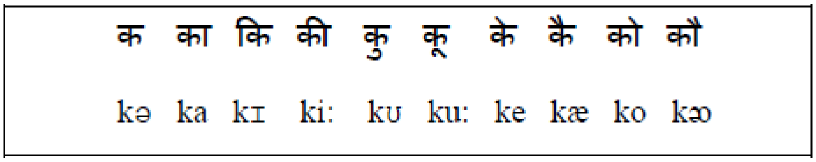
1. A vowel is an independent unit of *akshar* word-initially and post-vocalically.



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

Table 3: Independent vowel letters

a. 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 akshars 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	kɑʊ

Table 4: *Devanāgarī CV akshars*

2. As can be seen in the first grapheme in Table 3, the neutral vowel /★/ 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<db^{h}> उद् भव/udb^{h}əʋ/

6. For a C+r cluster, as noted above, the /r/ is specified as a subscript that looks like an inscript: क्र <kr>, ख्र <k^{h}r>, फ्र <p^{h}r>.

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 <kʂ>, 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>

2. ISSUES

From a typological point of view, issues arise because of the the following parameters:

a. ccTLD’s vs gTLD’s and geoTLD’s. While the former are under the control of a policy determined by a given country, the latter do not fall within the compliance of such a policy
b. Introduction of the notion of language tables, restriction rules and well-formedness constraints (in Brāhmī derived languages) and variant-hood to reduce spoofing, pharming and phishing. Thus for Brahmi based languages which are akshar driven, a formalism needs to be evolved to handle well-formedness.
c. Potential areas where such factors apply. These are:
    1. Issues arising out of the possible implementation of ZWJ/ZWNJ as prescribed in IDNA 2008
    2. Issues related to certain valid characters and combinations which will be protocol invalid
    3. Issues arising out of browser behavior and closely allied to the browser Font display issues
    4. Issues arising out of Registry Management
    5. Issues arising out of legal, administrative and financial implementation of the policy.
    6. Issues specific to gTLD’s, geoTLD’s

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

2.1. LANGUAGE vs. SCRIPT ISSUES

While the ccTLD for .भारत the dichotomy can be handled (with certain issues to be tackled at the registry level) , at the g(eo)TLD level, only script will dominate which implies adopting new strategies for handling variants under this Open Sky Policy.

2.2. Variants in Devanagari Script

Variants in Indian Languages are based on visual look-alikes. Three types of variants can be identified. Of these the first two are because of Unicode issues and the last is a true set of variants based on visually confusing characters:

2.2.1. Variants generated from legacy inputting methods

Earlier versions of Unicode did not have certain characters. In order to generate these characters alternative methods such as the use of Halanta followed by a ZWJ were used.

e.g. Eye-lash ra

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

2.2.2. Variants generated because of normalization

These variants exist because Unicode allows for two ways of entering certain characters. In the case of Devanāgarī the “nukta” character is the candidate for Normalisation . e.g.

क+◌़ U+0915 U+093C	क़ U+0958
ख+◌़ U+0916 U+093C	ख़ U+0959

As per revised IDNA standard, “IDNA 2008” the atomic form of Nukta characters have been marked as “Disallowed”, still as a precautionary measure, they have been kept as variants

2.2.3. Confusingly similar shapes

2.2.3.1. Single characters

These are the characters which have confusingly similar shapes. However, this category of variants were not considered in the .in ccTLD policy as there was a possibility that this approach would result in barring many useful domain names from being registered.
e.g.

घ U+0918	ध U+0927
भ U+092D	म U+092E

Table 4

This table contains only a sample list.

2.2.3.2. Composite characters

These are conjuncts that look alike and can be easily confused in the small URL bar of the browser. These look-alikes have been identified for each language.

e.g.

<table>
<tr><td>द्ग
U+0926 U+094D
U+O917</td><td>द्र
U+0926 U+094D
U+0930</td><td>द्न
U+0926 U+094D
U+0928</td></tr>
<tr><td>द्ध
U+0926 U+094D U+0927</td><td colspan="2">द्घ
U+0926 U+094D U+0918</td></tr>
<tr><td>ष्ट
U+0937 U+094D U+091F</td><td colspan="2">ष्ठ
U+0937 U+094D U+0920</td></tr>
<tr><td>द्व
U+0926 U+094D U+ 0935</td><td colspan="2">द्ब
U+0926 U+094D U+092C</td></tr>
</table>

Table 5

This table contains only a sample list.

2.2.4 Cross-script character variants

Mixing scripts is extremely dangerous and could result in spoofing, phishing and scamming. Mixing is not advisable. Since the policy for .भारत will not allow code-mixing and assuming that code-mixing will be for g(eo)TLD's as an exercise, a list of cross-lingual visual similarities is provided and which includes also digits. It should be noted that such similarities are restricted to single characters and not to conjuncts. A sample list of such cross script resemblances is provided below.

DEVANAGARI SCRIPT	**COGNATE SCRIPT**	**CODEPOINT IN COGNATE SCRIPT**
VOWELS		
उ 0909	Bangla	ও 0993
उ 0909	Gurmukhi	ਤ 0A24
ऋ 090B	Gujarati	ૠ 0AE0

CONSONANTS		
क 0915	Bangla	ক 0995
ग 0917	Gujarati	ગ 0A97
ग 0917	Gurmukhi	ਗ 0A17
घ 0918	Gurmukhi	ਬ 0A2C
घ 0918	Gujarati	ઘ 0A98
ङ 0919	Gujarati	ઙ 0A99
छ 091B	Gujarati	છ 0A9B
ञ 091E	Gujarati	ઞ 0A9E
ट 091F	Gurmukhi	ਟ 0A17
ठ 0920	Gujarati	ઠ 0AA0
ठ 0920	Gurmukhi	ਠ 0A20
ड 0921	Gujarati	ડ 0AA1
ढ 0922	Gurmukhi	ਫ 0A2B
त 0924	Gujarati	ત 0AA4
ध 0927	Gujarati	ધ 0AA7
न 0928	Gujarati	ન 0AA8
न 0928	Bangla	ন 09A8

न 0928	Bangla	ণ 09A3
प 092A	Gujarati	પ 0AAA
प 092A	Gurmukhi	ਧ 0A17
प 092A	Gurmukhi	ਪ 0A2A
प 092A	Gurmukhi	੫ 0A6B
म 092E	Gurmukhi	ਸ 0A38
म 092E	Gujarati	મ 0AAE
य 092F	Gujarati	ચ 0A9A
र 0930	Gujarati	૨ 0AAE
र 0930	Gurmukhi	ਕ 0A15
ल 0932	Bangla	ল 09B2
व 0935	Gujarati	વ 0AB5
श 0936	Gujarati	શ 0AB6
श् 0936+094D	Bangla	ঽ 09BD
ष 0937	Gujarati	ષ 0AB7
स 0938	Gujarati	સ 0AB8
ह 0939	Gujarati	હ 0AB9
Nukta characters		

ग़ 095A or 0917+094D	Gurmukhi	ਗ਼ 0A5A
ढ़ 095D or 0922+094D	Gurmukhi	ਫ਼ 0A5E

Table 6

2.2.5 Homophonic Variants

In Devanagari based languages, homophonic variants which admit two homographs e.g. हिंदी and हिन्दी do occur but the rules for such variants are ill-defined and could increase the chances of malfeasance.

2.3. ISSUES PERTAINING TO UNICODE NORMALISATION

While Unicode does provide rules for normalization which are reflected in IDNA2008, two major issues arise:

Within Unicode itself a large number of normalizations are not defined. A good example is from the Arabic code-page:

ڈ U+0688	◌ؕ د U+062F U+0615

The similar case occurs in Malayalam which is written in Malayalam script. Though this report is only aimed at Devanagari, Malayalam belongs to the same family as Devanagari which is “Brahmi” and hence being discussed.

ന്‍ U+0D28 U+0D4D U+200D	ൻ U+0D7B

2.4. ZERO WIDTH JOINER (ZWJ) AND ZERO WIDTH NON-JOINER (ZWNJ) :

ZWJ and ZWNJ are the invisible characters that have been provided by the Unicode to generate out certain shapes which otherwise are not possible through normal rendering mechanism. This is mostly applicable to those forms which are alternatives of each others. In each case the use of ZWJ is specified and the issues arising out of the said use are provided next

2.4.1 Zero Width Joiner (ZWJ)

The ZWJ plays multiple roles.

2.4.1.1 Used to generate half form of base consonant in "Base-Cons+Halanta+Cons"

There are some cases of conjunct formation in Indian Languages in which the basic shapes of two characters being joined by Halanta are not retained. If in such cases if the conjunct form in which the basic shapes(in some cases as half forms) of the combining characters is to be retained, the ZWJ is used after Halanta.

e.g.

क (ka) + ◌् (halanta) + ष (ssha) -> क्ष (ksha)

क (ka) + ◌् (halanta) + ZWJ + ष (ssha) -> क्‍ष (ksha)

Issue :

The issue that arises in this usage of ZWJ is that, there are some conjuncts which by default are represented in the form where the basic shapes (in some cases as half forms) of the combining characters are retained. In such cases the use of ZWJ after Halanta character does not make any difference visually. Thus we eventually get two strings which have different storage but same visual appearance.

e.g.

क (ka) + ◌् (halanta) + न (na) -> क्न (kna)

क (ka) + ◌् (halanta) + ZWJ + न (na) -> क्‍न (kna)

Also a point to be noted here is that the shape which is formed by combining characters is highly dependent on font and/or underlying rendering engine. Though this behavior is largely governed by the language needs, there are still some cases where discrepancies are observed and thus such cases cannot be clearly identified and singled out.

2.4.1.2 To generate some special characters

To generate out some characters in Indian Languages including Devanāgarī based languages, Unicode provided a combination with the use of ZWJ. e.g. in Marathi which is a Devanāgarī based language to generate out "eyelash ra"

र (ra) + ◌् (halanta) + ZWJ -> र्‍ (eyelash ra)

In successive versions of Unicode, some of these characters were encoded atomically (e.g. case of "khanda ta" in Bengali script) or given an alternative combination without use of ZWJ. The latter case does exist in Devanāgarī based languages. The “eyelash ra” was given a new combination which is

ऱ (rra) + ◌् (halanta) -> ऱ् (eyelash ra)

Issue :

The issue that arises in this case is, two different combinations will result in same visual form. Including this kind of combination in variant table will solve this issue.

2.4..2 Zero Width Non-Joiner (ZWNJ)

ZWNJ on the other hand is used, to put in broad sense, to explicitly display virama between two characters which otherwise would have joined to form a conjunct. As per Unicode (Chapter 9) *“Explicit Virama (Halant). Normally a virama character serves to create dead consonants that are, in turn, combined with subsequent consonants to form conjuncts. This behavior usually results in a virama sign not being depicted visually. Occasionally, this default behavior is not desired when a dead consonant should be excluded from conjunct formation, in which case the virama sign is visibly rendered. To accomplish this goal, the Unicode Standard adopts the convention of placing the character U+200C zero width non-joiner immediately after the encoded dead consonant that is to be excluded from conjunct formation. In this case, the virama sign is always depicted as appropriate for the consonant to which it is attached.”*

e.g.

क (ka) + ◌् (halanta) + ष (ssha) -> क्ष (ksha)

क (ka) + ◌् (halanta) + ZWNJ + ष (ssha) -> क्‌ष (ksha)

In the latter case, we can see the combining characters retaining their forms, with the halanta which is a joining character, having explicit visual appearance.

Issue :

The issue that arises in this usage of ZWNJ is that, there are some conjuncts which by default are represented in the form where the halanta has explicit visual appearance even in the absence of ZWNJ. In such cases the use of ZWNJ after Halanta character does not make any difference visually. Thus we again eventually get two strings which have different storage but same visual appearance.

e.g.

ड (dda) + ◌् (halanta) + द (da) -> ड्द

ड (dda) + ◌् (halanta) + ZWNJ + द (da) -> ड्‌द

Similar to the case of ZWJ, the shape which is formed by combining characters is highly dependent on font and/or underlying rendering engine. Though this behavior is largely governed by the language needs, there are still some cases where discrepancies are observed and thus such cases cannot be clearly identified and singled out.

2.5. ISSUES RELATED TO VALID CHARACTERS DECLARED "PROTOCOL INVALID"

1. Case of 02BC

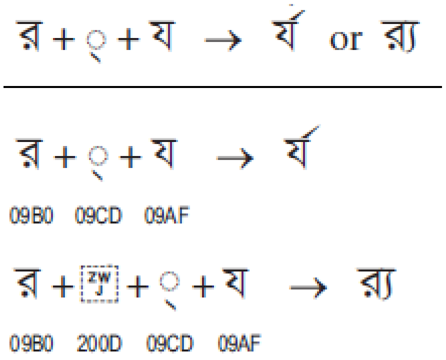
The character U+02BC *Modifier Letter Apostrophe* which acts as a tone mark or length mark is used very frequently in languages like Bodo, Dogri, Maithili which are Devanāgarī script based and Bangla which is Bengali script based.

Issue : U+02BC *Modifier Letter Apostrophe* character comes from the code space (code-page) 02B0-02FF. Whereas all the characters which belong to Devanāgarī script come from code space (code page) 0900-097F. If as a policy decision, script mixing is not allowed in IDNs, this character still be allowed as an exception because without this character the language representation will not be complete.

2. Use of ZWJ

As per IDNA 2008 protocol, the ZWJ has been permitted with the restriction that the preceding character must be a "virama". In Indian languages, ZWJ is used to display some combinations with same set of combining characters but different visual appearance. Though this case does not exist in Devanāgarī script, this case is found in other Indian language scripts.

The case of *"Interaction of Repha and Ya-phalaa"* which exists in Bengali script is a prime example. In general in Indian languages, the combination of "ra+halanta" when followed by a consonant generates a "reph". In case of Bengali, the combination "halanta+ya" is called as "ya-phala". When this combination is preceded by "ra" an ambiguous situation arises. Unicode[4] has proposed, that ZWJ be inserted after "ra" to generate ra with ya-phala.



2.6. ISSUES RELATED TO BROWSER BEHAVIOR[5] :

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

[4] Chapter 9. Unicode 6.0 http://unicode.org/

[5] Since the data dealing with browser behaviour under different Operating Systems is quite compendious, it is provided separately in the report as a set of PDF files.

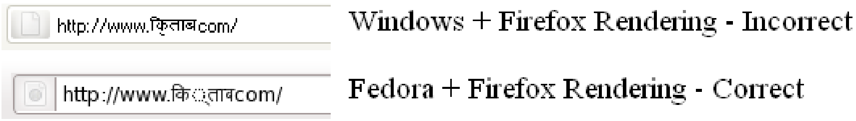
applied on the URL bar in IDNs are chosen by the browsers as per default font for the script of the domain name provided by the underlying OS.

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

1 Rendering Engine related issues
2 Font related issues

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 cases given below:

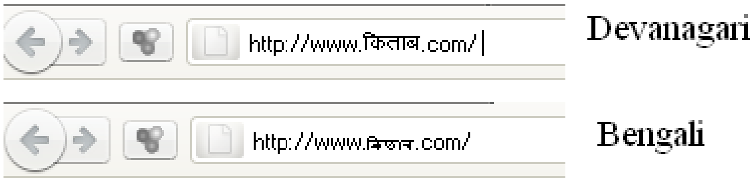

http://www.किताबcom/

http://www.कि◌ताबcom/
Windows + Firefox Rendering - Incorrect
Fedora + Firefox Rendering - Correct

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 act 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. There is a strong need for a central authority which will bring consensus in these implementations.


http://www.किताब.com/

http://www.কিতাব.com/

2.7. ISSUES ARISING OUT OF REGISTRY MANAGEMENT

Assuming that all other factors and conditions are satisfied, a major issue touches upon the registry. Registry issues can be divided in to the following parts. However a major caveat which overlies all issues is that of sub-sub domains. The policy developed for a given script/language stops being applicable once a domain is "acquired" by an individual or an entity.

e.g. An individual or an entity owns a domain www.कखग.भारत. The policies developed would apply only to the domain .कखग. The sub-domains would not be governed by the policy. Thus in the case of www.चछज.कखग.भारत, the policy which would apply to the first level domain name कखग will not apply to चछज. Appropriate mechanisms need to be evolved or a call to be taken as to the "depth" to which the policy evolved will apply.

Given this major caveat, the following issues arise out of registry management

2.7.1. Delegation of variant TLD's

2.7.2.Registry Management of ABNF, Restriction rules, Language Tables and Variant Tables

2.7.3. "Localisation" of WHOIS

2.7.4. Email Addresses resolution

2.7.1. Delegation of variant TLD's

There is a strong possibility that the zone generation process might be affected when variants of a given TLD label are supplied to it. This eventuality raises certain issues which need serious consideration:

1. Identification of variant type. In the case of Brāhmī based scripts, three different variant types have been identified (cf. 3 above). The Registry will have to interact differently with each variant type. Variant which require normalization and legacy driven variants will need to be handled differently from look-alikes.
2. Corollary to the above is the question of how the zone file for a given TLD variant be handled ? Will it share the same zone file or will allocation be made in the registry for each variant? Basically the registry will have to take a call and as mentioned above, accommodate the variant as per its variant type.
3. The final issue is that of language and script. Given that a script supports more than one language (Devanāgarī and Bengali in the case of Brāhmī based languages) how should the registry handle this problem in terms of resource records ? Should for example a TLD admitting a variant in Nepali be pertinent to domains appertaining to that language alone ?
4. Finally outside the ambit of a ccTLD i.e. gTLD's and geoTLD's how will the disambiguation function across language and script? In the present state, script seems favoured over language.

2.7.2. Registry Management of ABNF, 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.

Some of the major issues that arise are as under:

1. In the case of Devanāgarī, a large number of languages use the code-page 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 3 supra)
3. In the case of GeoTLD's and gTLD's which are not under the control of .भारत registry, which rules will apply? It is suggested that in this case ICANN should deploy the rules and variant tables defined for each script/language

2.7.3. "Localisation" of WHOIS

The term "Localisation" has been used for the 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. Localisation 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
    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 ?

2.7.4. Email Addresses resolution

1. The problems raised in 5.4.3. have a marked resolution for resolution of email addresses. Given an email such as

   वित्त-मंत्रालय.भारत: Ministry of Finance

   Will the owner of the address also inherit the variant

   वित-मंत्रालय.भारत
2. In case both emails are valid, will there be an aliasing mechanism ?
3. The issue is also closely tied with that of the body resolving the email.

2.8. ADMINISTRATIVE ISSUES

Section 6 above spells out 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.

2.8.1. RESERVED NAMES LIST

Reserved names Lists are deployed for sensitive names which need to be protected by a given country. In the case of .भारत, the following issues could arise, especially with regard to geoTLD's and gTLD's:

1. Would geoTLD's and gTLD's need a reserved list? Will the Government send a list of reserved names of political sensitivity ? If so are payment issues
involved?
2. Should all variants of a given gTLD or geoTLD be also requested for blocking ?

2.8.2. 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 (cf. 4. Infra) multi-lingualism brings in its own issues:

Multi-lingual dispute claims. These 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 ? Ramifications of Automatic Bundling: What happens if a variant generated blocks out a valid 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 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. A trademark in Tamil for India resembles a similar one in Sri Lanka.
   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 ?

2.8. 3. PAYMENT ISSUES
With the creation of multi-lingual labels and also variants generated from each, certain issues of payment arise:
Will there be a fee for providing and registering Variants
Will there be a fee for a registrant desirous of removing a variant granted to him (issues of cyber-squatting)
Will there be a concession for providing the registrant a label in multiple languages ?

2.9. MANAGEMENT OF MULTI-LINGUAL geoTLD'S AND gTLD'S

The issues raised here are specific to geoTLD's and gTLD's where a given country's policies do not apply. Certain issues need to be discussed in this area:

1. How are these to be allocated?
2. Will the g(eo)TLD's permit code-mixing i.e. permitting more than one script to be used within a given g(eo)TLD ?
3. Will there be a specific reservation for a country to register its societal and politically sensitive names?
4. Which policy will apply for generation of variants ?
5. If a given corporate body is desirous of registering a geoTLD 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. Thus for Perso-Arabic scripts, the policy adopted by the Arabic study group be applied
6. If the above suggestion is accepted, what measures are taken in the case of a script shared by more than one country, in case the given countries have different policies

7. Legal and societal issues arising out of gTLD's:[6]

1. *Terms and Conditions for usage of gTLD are determined by Allottee*

This could lead to Monopolistic and Anti Competitive business advantage. The 'terms and conditions' should have a caveat of approaching relevant Anti Trust Bodies/ Competition Bodies/ Judiciary/ ICANN for misusing of 'Dominant Position' by Allottee if there are any cases. ICANN should also keep an eye open for such issue with proper remedial measures.

2. *Raising of objection*

The duration of two months is a small period. There is a possibility that within the short window open from Jan to April, 12 many organizations don't understand the value and the process of objection effectively.

Objection raising should also be extended after the allottee has been given the gTLD. Since this window is opening for the first time and has a small period, people may not be aware. Hence, this process should be extended for this particular opening.

3. *Registrar's role in domain name*

A private individual/organization should not have complete authority to decide over allotted gTLD for the remaining period. ICANN should be in a position to control any misuse of gTLD at any point of time. The message can reach ICANN through a notified national body like NIXI in India.

4. *Allottee determined 'Terms and conditions' of usage*

Allotee 'Terms and conditions' could prove unrealistic for some organizations/individuals and could create a digital divide.

A Standard terms and condition of operating the gTLD should be issued by ICANN and it should be subject to National Jurisdiction.

5. *Allocation of gTLD can be revoked by Allottee*

Allottee should have enough reasonable and valid legal grounds for this extreme step. Could lead to a scenario where the gTLDs original allottee can force a user to stick to fixed roadway. Revocation powers should be utilized only after consultations with a neutral body, and ICANN should have a role in it.

6. *Licensing fees*

There is not much scope for public services/charitable organizations. Some consideration should be given to organizations from LDC/DC as per nomenclature of world bank. Also, charitable organizations should be given a hefty discount after proper verification.

[6] This is a summary of Issues raised during a Workshop held on IDN's gTLD's and geoTLD's held at Hyderabad on 19th August 2011. These issues were raised by Mr. Ankit Kumar, Legal and Corporate Affairs.,Deloitte Consulting India Private Limited,

2.10. SUMMING UP

The following table sums up the above discussion for easy reference:

ISSUES	SUB-ISSUES
Linguistic Issues	Language vs. Script. While the ccTLD for .भारत the dichotomy can be handled, at the g(eo)TLD level, only script will dominate which implies adopting new strategies for handling variants
Unicode normalisation issues	In the case of Brahmi-based Scripts as well as Scripts derived from the Arabic Code-page, there is an urgent need to study possible normalization rules not covered by Unicode and by IDNA2008.
Issues arising out of the possible Implementation of ZWJ/ZWNJ as prescribed in IDNA 2008	ZWJ can be handled by constraint rules. Such rules will need to be defined. ZWNJ for generating noun paradigms for languages like Nepali need to be discussed since there is no rule-governed behavior
Issues related to certain valid characters and combinations which will be protocol invalid	Case of Boro, Dogri which use a character from the Spacing Modifer letter set /ʼ/ and which cannot be accommodated in the present conditions laid down by IDNA2008
Issues arising out of browser behavior **Font display issues**	1. **Rendering Engine related issues** 2. **Font related issues**
Issues arising out of Registry Management	**Delegation of variant TLD's** 1. Identification of variant type. In the case of Brāhmī based scripts, three different variant types have been identified (cf. 3 above). The Registry will have to interact differently with each variant type. Variant which require normalization and legacy driven variants will need to be handled differently from look-alikes. 2. Corollary to the above is the question of how the zone file for a given TLD variant be handled ? Will it share the same zone file or will allocation be made in the registry for each variant? Basically the registry will have to take a call and as mentioned above, accommodate the variant as per its variant type. 3. The final issue is that of language and script. Given that a script supports more than one language (Devanāgarī and Bengali in the case of Brāhmī based languages) how should the registry handle this problem in terms of resource records ? Should for example a TLD admitting a variant in Nepali be pertinent to domains appertaining to that language alone ? 4. Finally outside the ambit of a ccTLD i.e. gTLD's and

	geoTLD’s how will the disambiguation function across language and script? In the present state, script seems favoured over language. **Registry Management of ABNF, Restriction rules, Language Tables and Variant Tables** 1. In the case of Devanāgarī, a large number of languages use the code-page 900. Given that the registry for .???? 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 3 supra) 3. In the case of GeoTLD’s and gTLD’s which are not under the control of .???? registry, which rules will apply? It is suggested that in this case ICANN should deploy the rules and variant tables defined for each script/language **“Localisation” of WHOIS** 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. Localisation 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: 3. 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 4. 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. 5. 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 ? 6. 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 ? **Email Addresses resolution** 1. Will the owner of an email address also address also inherit the variant 2.In case both emails are valid, will there be an aliasing mechanism ?
Issues arising out of legal, administrative and financial implementation of the policy.	**RESERVED NAMES LIST** Would geoTLD's and gTLD's need a reserved list? Will the Government send a list of reserved names of political sensitivity ? If so are payment issues involved? Should all variants of a given gTLD or geoTLD be also requested for blocking ? **DISPUTE RESOLUTION** Multi-lingual dispute claims. These are bundles containing labels in different languages. How does a complainant claim rights to a whole label ? What happens if a variant generated blocks out a valid label 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 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? Who will provide the expertise in the area of a multi-lingual dispute. Will the matter be referred to the State Government or to a competent language authority ? International Trademark resolution: Which procedure would be followed when a trademark or domain name is claimed by two countries ? Government vs. an Individual or a Corporate body . Who will take precedence ? **FISCAL ISSUES** Will there be a fee for providing and registering Variants ? Will there be a fee for a registrant desirous of removing a variant granted to him? Will there be a concession for providing the registrant a label in multiple languages ?
Issues specific to gTLD's, geoTLD's	How are these to be allocated? Will the g(eo)TLD's permit code-mixing i.e. permitting more than one script to be used within a given g(eo)TLD ?

	Will there be a specific reservation for a country to register its societal and politically sensitive names? Which policy will apply for generation of variants ? If a given corporate body is desirous of registering a geoTLD in a variety of scripts, which policy will apply ? What measures are taken in the case of a script shared by more than one country, in case the given countries have different policies ? Legal and societal issues arising out of gTLD's :

Table 7

3. REGISTRAR AND REGISTRY PERSPECTIVE[7]

3.1.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 can not 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?

3.2. 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?

[7] This section has been contributed by Afilias and Nixi

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.

3.3 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.

4. ROLE OF THE GOVERNMENT IN SHAPING THE IDN POLICY [8]

1. DISPUTE RESOLUTION

2. RESERVED NAMES LIST

3. FEES

THIS SECTION TO BE PROVIDED BY DIT

[8] This section has been contributed by DIT and NIXI

5. REFERENCES

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.

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
Dürst, Martin J. (December 10, 1996). "Internet Draft: Internationalization of Domain Names". The Internet Engineering Task Force (IETF), Internet Society (ISOC). Dürst, Martin J. (December 20, 1996). "URLs and internationalization". World Wide Web Consortium. IDN TABLES: http://www.iana.org/domains/idn-tables/

LANGUAGE SPECIFIC

3. *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:

4. *CHINESE*

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

5. *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 2690 A Proposal for an MOU-Based ICANN Protocol Support Organization September 1999
RFC 2870 Root Name Server Operational Requirements June 2000
RFC 3454 "Preparation of Internationalized Strings ('stringprep')"
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)"

http://tools.ietf.org/html/draft-iucg-idna2008-ietf-lc-00 Comments on the IDNA2008 documents set draft-iucg-idna2008-ietf-lc-00.txt September 30, 2009

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)

Allowing Special Characters in Identifiers http://unicode.org/review/pr-96.html

6. LIST OF APPENDICES

Appendix I: Devanāgarī Team Members.

Appendix II: List of Official Languages of India.

Appendix III: Discussion on the Definitions and Questions proposed at Singapore meet.

Appendix IV: IDN Variant TLDs - Lists of Issues - v06 redline.

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)

APPENDIX II: **List of Official Languages of India**[9]

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[10] (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 :

Language	ISO	Official Language	Family	Script
Assamese	asm	Assam	Indo-Aryan	Assamese
Bengali	ben	Tripura and West Bengal	Indo-Aryan	Bangla
Bodo	bod	Assam	Tibeto-Burman	Devanāgarī (modified)
Dogri	dgr	Jammu and Kashmir	Indo-Aryan	Devanāgarī (modified)
Gujarati	guj	Dadra and Nagar Haveli, Daman and Diu, and Gujarat	Indo-Aryan	Gujarati
Hindi	hin	Andaman and Nicobar Islands, Bihar, Chandigarh, Chhattisgarh, Delhi, Haryana, Himachal Pradesh, Jharkhand, Madhya Pradesh, Rajasthan, Uttar Pradesh and Uttaranchal	Indo-Aryan	Devanāgarī
Kannada	kan	Karnataka	Dravidian	Kannada
Kashmiri	kas		Indo-Aryan	Perso-Arabic Devanāgarī
Konkani	kok	Goa	Indo-Aryan	Devanāgarī Roman
Maithili	mai	Bihar	Indo-Aryan	Devanāgarī
Malayalam	mal	Kerala and Lakshadweep	Dravidian	Malayalam
Manipuri	mni	Meitei	Tibeto-	Bangla

[9] This section has been contributed by GIST Group. CDAC

[10] http://www.ethnologue.com/show_country.asp?name=in

			Burman	Meitei-Meyek
Marathi	mar	Maharashtra	Indo-Aryan	Devanāgarī
Nepali	nep	Sikkim	Indo-Aryan	Devanāgarī
Oriya	ori	Orissa	Indo-Aryan	Oriya
Punjabi	pan	Punjab	Indo-Aryan	Gurmukhi
Sanskrit	san	Pan-Indian	Indo-Aryan	Devanāgarī
Santali	sat	Jharkhand	Munda	Ol Ciki Devanāgarī (modified)
Sindhi	snd	Pan-Indian	Indo-Aryan	Perso-Arabic Devanāgarī Gujarati Roman
Tamil	tam	Tamil Nadu and Pondicherry	Dravidian	Tamil
Telugu	tel	Andhra Pradesh	Dravidian	Telugu
Urdu	urd	Jammu and Kashmir	Indo-Aryan	Perso-Arabic

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 Brahmi: 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, 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: Comments on the white paper on Definitions and Questions circulated at the ICANN meet in Singapore in June 2011[11]

PDF UNDER DELIBERATION. WILL BE CIRCULATED SEPARATELY

[11] This section has been contributed by GIST Group. CDAC with inputs from Dr N. Ostler and Mr. Andrew Sullivan.

Appendix IV:
IDN Variant TLDs - Lists of Issues - v06 redline

RECEIVED ON 24TH RESPONSES WILL BE MAILED SEPARATELY