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

INTERNATIONALIZED DOMAIN NAMES

IN

INDIAN LANGUAGES

A DRAFT POLICY DOCUMENT

Policy Framework

&

Implementation Plan

November, 2009
Department of Information Technology
Ministry of Communications & Information Technology
Government of India

Table of Contents

POLICY FRAMEWORK & IMPLEMENTATION PLAN

1. BACKGROUND

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

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

Today, the above mentioned addressing mechanism supports "Simple English Latin" alphabet of the 26 "Usual" un-accented Latin letters, the 10 "Arabic" Digits (i.e. 0-9), and hyphen (-) with dot (i.e. ".") as field separators. In addition, capitalization of letters (i.e. upper case) would be insignificant, so that the strings "Bharat", "BHARAT" and "BHArat" all resolve in the same manner.

However, for a number of crucial Customer-centric Applications (such as e-governance, e-learning, e-commerce) sole dependence on a single language (i.e. English) may not be sufficient to provide the requisite infrastructural support to all kinds of Internet usage in the present and in the future. Till recently, there was no standardized method for specifying Domain Names in any non- "Simple English Latin" Character Set. Thus, there is a dire need for the common Internet Infrastructure to support local languages using their respective written scripts for expressing the Domain Names; such Internationalization (termed as "Internationalized Domain Names (IDNs)) is being termed as the associated key process for supporting such requirements towards domain names in multilingual scripts with their respective linguistic as well as cultural sensitivities in the Internet Infrastructure.

One of the most significant innovations in the Internet since its inception will be the introduction of top level Internationalized Domain Names (IDNs). These will offer many new opportunities and benefits for Internet users around the world by allowing them to establish and use domains in their native languages and scripts.

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

The development of this policy culminating in this white paper has been the outcome of long discussions with linguists and experts. C-DAC GIST Pune in close collaboration with DIT and with the cooperation of C-DAC Kolkata and C-DAC Thiruvananathapuram has evolved the policy of implementation of Internationalized Domain Names in Indian Languages and also tried to institute safe-guards and safety measures through methods such as defining the Brahmi Syllable pertinent to IDN,

providing restriction rules as well as variant tables to make IDNs in Indian languages as secure as possible.

2. OBJECTIVES

The main objectives of this policy document are to demystify the IDN policy by providing information regarding the broad policy, furnishing details of the Backus-Naur Formalism which is the backbone of the Brahmi syllable and which has been suitably modified for IDN (termed hence hitherto as ABNF).

Given the complexity of Indian writing systems, pharming and spoofing can be a major threat. As a palliative measure, restriction rules have been introduced.

In a nutshell, the main objectives are as under:

a. To ensure that Indian languages can have their rightful place in Internationalized Domain Names and that one can have a URL in an Indian language.

b. To initially permit such URLs in the following major languages/scripts:

   Devanagari (Marathi, Hindi, Konkani, Sanskrit and Nepali), Gujarati, Oriya,

   Punjabi, Malayalam, Tamil, Telugu, Kannada, Assamese, and Bangla and

   subsequently to be adapted for all the 22 official languages including those using

   Perso-Arabic scripts: Urdu, Sindhi, Kashmiri.

c. To limit, at present the Indian language component to the Domain Name and localize the ccTLD1 i.e. .in as provided in the Appendix. To a large extent with some exceptions this will also allow language identification.

3. INTERNATIONALIZED DOMAIN NAMES (IDNs)

The development of Internet has mainly taken place in one language namely English leading to the language barriers for non-English speakers. The Internet was mostly designed based on the 'simple English alphabet of the 26 Latin letters, the 10 "Arabic" digits (0-9), and hyphen (plus, of course, the dots).

ITEF has brought out the implementation standards for non-Latin and non-Roman characters and how the domain name is converted into puny code and how it resolves in the root zone files. The process of supporting multilingual script and other linguistic and cultural needs on the Internet is generally known as Internationalization. Internationalized Domain Names are domain names or web addresses represented in local language characters.

Internationalized Domain Names or Domain Names in Indian Languages is one of the effective ways of further promotion of Internet among the Indian Populace. For proliferation and preservation of heritage, culture and content creation in multiple languages, it is essential to have the domain names in their own scripts. The process

of Internationalization involves identification of the Character Set in the the respective Script for each language, identification of the Variant Set (similar looking characters within the Script) and a Language based implementation rule set. This is then followed by the process of normalization and coding for inclusion in the Domain Name System. An interface for Registrars for the issuance of Domain Names is then developed for the launch of Domain Name Registrations.

IDN is a societal issue as much as a technical challenge. Considering that a large number of users are not scholars of the language and hence can be easily cheated by homographs, spoofing, pharming as well as phishing will occur to a large extent in Indian languages. This calls for great care and caution in supporting local languages and scripts in the domain names.

4. LINGUISTIC SCENARIO OF INDIA

Official Indian Languages and their Scripts with the exception of those written in the Perso-Arabic Script such as Urdu, Sindhi and Kashmiri (Perso-Arabic) and Santali (Ol Chiki script), are Brahmi based. The grammatical and phonetic placement of Brahmi scripts for all the Languages is to a large extent identical such that any application developed using one language can be directly mapped to all other Indian Languages i.e. the rule set for application development in one language would generally hold true for all the other Indian Languages.

Thus Indian Languages with the exception of those mentioned above, have a unique identity because of the fact that they are all issued from a common writing system-Brahmi. They are extremely phonetic in nature and permit a large number of close homographs . They are very different from Latin based Languages, Chinese Japanese Korean (CJK) or Semitic Languages. While the Roman scripts are linear having as many glyphs as there are letters, Indic scripts which are Brahmi based are non linear and complex in their writing system. Simply put, the number of glyphs is much larger than the letters of the alphabet.

4.1 Characteristics of Indic scripts

With more than a billion inhabitants in the democratic republic of India, there are about 1650 dialects spoken by different communities. There are 22 constitutionally approved languages, used in different states for citizen interface but only 11 Scripts shared among them. Quite a few Indian languages share the same script, for example, the languages Bangla, Assamese, Manipuri have the same script Bengali. Similarly, Bodo, Marathi, Nepali, Hindi, Sanskrit, Konkani, Maithili and Dogri have the same script Devanagari. Conversely, there are some languages that can be written in many scripts. For example, Punjabi can be written in Shahmukhi and Gurmukhi. Sindhi in Perso-Arabic as well as Devanagari. In the case of the Indic scripts originated from the Brahmi script. Due to a similarity in origin, there is an essential cohesion between them. Indian scripts derived from Brahmi (unlike those derived from Semitic: Perso-Arabic) are syllabic in nature. Starting with this as the major premise, the SYLLABLE

within a Brahmi-based writing system can be visualized as VOWEL BASED or CONSONANT BASED.

4.2 Indian script typological sub-sets are as below

- The CONSONANT (C) contains an implicit VOWEL (schwa in the case of Devanagari), enabling it to be pronounced.
- VOWELS (V) can have two shapes: FULL VOWELS and DEPENDENT VOWELS or MATRAS (M) which can be appended to the CONSONANT to change the phonetic value of the consonant. When this adjunction occurs, the implicit vowel is dropped and is substituted by the corresponding vowel value .: k+schwa ी की ..: k+SCHWA+U =ku कु
- Vowel modifiers (D) which are the nasal markers e.g. Chandrabindu, Anusvara and Visarga .
- HALANTA (H). The Halanta is the implicit vowel stripper and is used for stripping off the implicit vowel when a ligature/conjunct has to be generated. Thus क ् क i.e. k+SCHWA+Halanta+ k resulting in kk क्क.
- NUKTA (N) used in Devanagari and some other scripts as to indicate either the flaps of Hindi ड़ ढ़ or loan-characters borrowed from another language (mainly Perso-Arabic languages क़, ख़, ग़, ज़, फ़ or in the case of Marathi the ऱ used to produce the eye-lash repha.
- The Vowel syllable has the following permissible pattern V [D] i.e. a Vowel can stand by itself to constitute a full syllable or be followed by a VOWEL MODIFIER i.e. Chandrabindu, Anusvara and Visarga. ANY OTHER COMBINATION IS ILLEGAL and is represented automatically by the intervention of a rounded shape termed in popular parlance as the “GOLU” (meaning: the round one) to indicate an “ILLEGAL” combination.
- As a general rule the Consonant Syllables follow the pattern *3 [CH] C [M] [D], meaning that upto 3 half characters followed by Consonant, which can be optionally followed by a Matra (independent vowel), and Vowel Modifier (Chandrabindu, Anusvara and Visarga). For certain consonants, Nukta can also follow. ANY OTHER COMBINATION IS ILLEGAL and is represented automatically by the intervention of a rounded shape termed in popular parlance as the “GOLU” (meaning: the round one) to indicate an “ILLEGAL” combination.

5. RESOLVING IDN IN THE EXISTING DOMAIN NAME SERVER (DNS)

IDNA is a mechanism defined in 2003 for handling internationalized domain names containing non-ASCII characters. Rather than redesigning the existing DNS infrastructure, it was decided that non-ASCII domain names should be converted to

a suitable ASCII based form by web browsers and other user applications. IDNA was designed for the maximum backward compatibility with the existing DNS system, which was designed for use with names using only a subset of the ASCII character set.

ASCII Compatible Encoding: Punycode

To conform to the current DNS standard, the Unicode string needs to be converted to an intermediate ASCII string, called ASCII compatible encoding (ACE). This intermediary code is called the Punycode.

Punycode, is a Bootstring algorithm enabled by IDNA and defined in RFC 3492, which uniquely and reversibly converts Unicode string into an ASCII string, i.e., between the restricted ASCII (LDH) and non-ASCII representations of a domain. The algorithm consists of two components, viz. called ToASCII (Encoding) and ToUnicode (Decoding), both central to the working of IDNA. Each of these algorithms is not applied to a domain name as a whole but to the individual labels that compose a domain name.

ToASCII leaves unchanged any ASCII label. If given a label containing at least one non-ASCII character, ToASCII will apply the Nameprep algorithm which converts a label to lowercase and performs other normalizations. Then it translates the result to ASCII using Punycode before prepending the four character string "xn--". This four character string is called the **ACE prefix** and is used to distinguish between Punycode encoded labels from ordinary ASCII labels.

ToUnicode reverses the action of ToASCII, stripping off the ACE prefix and applying the Punycode decode algorithm. It does not have any effect on any string that does not begin with the ACE prefix.

6. MAJOR ADMINISTRATIVE POLICY ELEMENTS

6.1 POLICY IN BRIEF

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

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

6.2 BROAD POLICY

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

i. General Policies

ii. Reserve Name Policy

iii. Dispute Resolution Policy

iv. Pricing Policy

v. Launch Stages, including Sunrise Period

6.2.1 General Policies

1. The .IN IDN registrations may be open at top level, 2nd and 3rd levels similar to ASCII .IN domain names.
2. As is the case with ASCII domain name registrations, the zones .gov.in, .mil.in and .ac & .edu.in will be reserved for the Govt., Defence and Educational institutions respectively. The registrations at the 3rd level in these zones will be carried out by the Govt., or an institution identified by the Government, which presently are: .gov.in registration is handled by NIC, .ac.in & .edu.in by ERNET and .mil.in by a Defence organization suggested by the MOD.
3. For other zones (i.e. .IN, .co.in, .org.in, .firm.in, .net.in, .gen.in, .ind.in), existing Registrars of NIXI may do IDN registrations also, subject to their passing OT&E test for IDN services. The same condition will apply to any Registrars which NIXI may accredit in future.
4. The General Polices regarding Term of domain name, Auto Renewals, Transfers, Grace Period, Contact Information and Nameservers etc. will be the same as are being followed for ASCII domain names.

To enable easy user access and registration, the .IN Registry will make available the following utilities, in each official Indian language, on the .IN Registry Website:

- Floating (Soft) keyboards – one specially made to enable Domain Name Registration in that language, and another keyboard catering to all characters of that Indian Language according to latest version of Unicode (Unicode 5.1).
- One visually appealing Unicode compliant font, freely downloadable for that Indian Language.

6.2.2 Reserved Names

For Indian language .IN IDN reserved names would be held similar to the ASCII .IN reserved names policy. The .IN Registry will get the English Reserved Names list properly translated into Hindi for this purpose. Depending upon the experience with Hindi Reserve names, similar exercise may be undertaken for IDN in other languages, which will be launched later.

6.2.3 Dispute Resolution Policy

The Dispute Resolution Policy and other mechanisms should be capable to handle the possibility of malicious intents. A comprehensive well defined dispute resolution policy has already been put in place for ASCII .IN names, in line with internationally accepted guidelines prescribed by WIPO, and Universal Dispute Resolution Policy (UDRP) adopted by ICANN. The same policy will be used for IDN. The .IN Registry may organize workshops for its arbitrators to familiarize them with Trademark issues in Indian languages.

While the .IN Dispute Resolution Policy (INDRP) has worked well, the rates prescribed under INDRP Rules of Procedure are on higher side, resulting in very few (less than fifty) complaints so far, thus indirectly giving an undue advantage to cyber squatters. It is proposed that to start with, the Fees for IDN .IN domain names may be fixed as follows:

Administration Fee	*Rs. 3000/-*
Arbitrator's Fee	*Rs. 12000/-*
For personal hearing	*Rs. 1000/- per hearing (Maximum of four hearings)*

It should also be ensured that arbitrators stick to the time frame specified in the INDRP.

6.2.4 Pricing for Internationalized Domain Names

The existing wholesale prices charged to Registrars for ASCII .IN domain names may be made applicable for IDN also, that is:

Rs. 300/- for Top Level (.भारत)

In order to protect the interests of .IN registrants, it should be ensured that while Registrars and their Resellers are free to charge any retail price above (or below) the wholesale price which .IN registry charges them, all other charges like RGP fees will NOT be more than the charges fixed by .IN Registry. It may also be ensured that the existing Registrar will NOT charge any Transfer Fees in case the Registrant desires to transfer the name to a new Registrar. Only the Gaining Registrar will charge the domain name per year fees as per its retail price. It should also be ensured that all domain names carry proper, correct details of Registrant, and that Registrars or their Resellers do not grab or hoard the .IN domain names in their own names or some dummy names.

6.2.5 Launch Stages

Domain Name Registration in Indian Languages will be started **with offering IDN in Hindi language followed by other Indian Languages in a phased manner**, with a three-stage launch:

- Soft Launch
- Sunrise Period
- Public Launch

Soft Launch Period of 60 days followed by 4-8 weeks of review and analysis for formal launch with a Sunrise period. For "Soft Launch", about 100 government websites would be identified and Hindi domain names will be made functional by directly entering them into the .IN registry database, and ensuring proper configurations of their nameservers. This will result in thorough testing of the domain name resolution methods, and would also give sufficient time to Technical Services Provider of NIXI to develop and test Registrar Toolkit for IDN. NIXI may also hire some Hindi website designers/design company to produce Hindi homepages for these chosen government websites in case they do not have existing Hindi homepage. The Domain-Language-Policy profile would also be published publicly and to the IANA IDN Language-Table register (http://www. iana.org/assignments/idnl).

Sunrise Period Trademark owners, registered companies and owners of intellectual property have a legitimate interest in protecting their valuable names online. In the Internet domain, it is achieved using a "Sunrise Period". A Sunrise period of 8 (eight) weeks from the opening of registration at 2nd/3rd level will be announced during which genuine registrations with proper verification will be allowed as per the policy for these registrations, with first preference

given to Indian interests. NIXI will adopt a similar procedure as it had adopted while launching .IN domain names in ASCII, taking care that sufficient legal staff/lawyers are appointed to complete the scrutiny of the applications in a time bound fashion. *In order that marketing costs and operational costs for Sunrise Period are met to some extent, an application price of Rupees 2000/- (Two thousand rupees) per application (i.e. per IDN domain) is proposed.*

Public Launch After the Sunrise period, public launch will be announced through advertisements in all media, and the registrations will be open to public on first come first served basis.

7. DIALOGUE WITH OTHER NATIONS USING THE SAME SCRIPT

Some Indian scripts are shared across geographical boundaries (i.e. Bangla is used in Bangladesh and India, Urdu in Pakistan and India, Tamil in Singapore and India, Nepali in Nepal and India) a collaborative effort can be made to avoid the confusion among the users of the corresponding language or script community.

8. PUBLIC REVIEW

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

9. MAJOR IMPLEMENTATION POLICY ELEMENTS

1. CODE SET UNICODE COMPLIANCY

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

2. SCRIPT AND LANGUAGE:

DIFFERENTIATION OF SCRIPTS AND LANGUAGES

A major decision taken is that Scripts and Languages will be differentiated at the registrars level and that the user will be provided with keyboards which will allow

him to enter an IDN in the language of his choice. Although this does not affect languages like Gujarati or Tamil where there is the relationship of One script <-> One language, it does make a difference in scripts such as Devanagari or Bengali, where one script caters to many languages e.g. Hindi, Marathi, Konkani, Nepali, Sanskrit, all use Devanagari.

3. BASIC STRUCTURE OR CANONICAL FORM

The basic structure of the IDN which will be discussed at length in Part II is based on the notion of the syllable, which in turn is defined by a Backus-Naur Formalism. Within this formalism, certain entities shall be permitted and others disallowed:

A. PERMISSIBLE ENTITIES

Letter-Hyphen-Digit shall be the only entities permitted. Hyphen and Digits shall belong to the Latin set. Letters will be of the language in question.

e.g. ***1 2 3 4 5 6 7 8 9 0*** and – will be of the Latin set.

All letters (characters) will be of the pertinent script.

B. NOT PERMISSIBLE

1. CODE-PAGE MIXING

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

e.g. www.सोफ्ट-वेर .in or www.हिन्दी-Hindi.in is not permissible since Hindi and Gujarati are mixed together and Hindi and English are mixed together respectively.

2. DIGITS

Digits in Indian languages will NOT be allowed.

૦૧૨૩૪૫૬૭૮૯.

3. PUNCTUATION MARKERS

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

4. OTHER SYMBOLS AND ABBREVIATIONS

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

5. **RARE AND OBSOLETE CHARACTERS**

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

6. **STRESS MARKERS OF CLASSICAL SANSKRIT AND VEDIC**

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

7. **SINGLE DIGIT AND COMBINATION OF TWO DIGIT**

   Single digit (e.g 1,2,3,4 etc.) and Combination of two digits (e.g 12, 23, 34 etc) will NOT be allowed as per the .in registry.

   Similarly according to .in registry ".IN domain names may be between 3 and 63 characters in length".

   All other rules pertaining to .in will be followed.

4. SAFEGUARDS

To protect as far as possible against spoofing, phishing and pharming attacks the following safeguards have been introduced. These attacks take place by substituting an address which looks visually alike but which in fact is a fake URL.

It should be remembered that the browser window allows for a font size which is relatively small and hence can lead to visual spoofing.

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

A. DISALLOWING ZERO WIDTH JOINER AND NON-JOINER

The use of Zero width Joiner / Zero width non Joiner (vide RFC 3454)

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

महाराष्ट्र without ZWJ and ZWNJ

महाराष्ट्र with zero width joiner after हा

महाराष्ट्र with zero width non-joiner after म

B. VARIANT TABLE

The aim of the Variant table is to identify visual look-alikes or homographs and ensure that such homographs shall not be permitted. A common case of visual look-alike is the case of द्ध and ध्द It is precisely to protect against such visual identity or homographs that a variant table has been instituted. The function of the variant table is to allow one of the homographs, debarring the other one. First use of either one of the characters shall automatically disallow the other in the case of a given word.

Thus if a user chooses समरुद्धी, the choice will automatically debar समरुध्दीprotecting against possible spoofing.

The following rules determine variant tables:

1. Since exclusion tables based on variants can debar a large number of words commonly used, the variant table shall be used sparingly and only when absolutely necessary.

2. Further the variant table shall apply only to ligatures or conjuncts or combination of two or more consonants and single characters that have homographic identity shall not be part of the variant table, the logic being that a native speaker can easily disambiguate single characters. It is the conjunct forms that pose the maximum problems

3. Normalization shall be part of the variant table and shall be provided as a safety measure. Therefore characters which need normalization should automatically normalize if they are a part of a variant table.

e.g. in case of Hindi: क(0915) + ◌093C) = क़(0958)

5. LANGUAGES

Although India has 22 official languages, this document only caters to the following languages:

Assamese

Bangala

Gujarati

Punjabi

Hindi

Kannada

Konkani

Malayalam

Marathi

Nepali

Oriya

Sanskrit (Laukik)

Tamil

Telugu

Languages using Perso-Arabic script: Urdu, Sindhi, Kashmiri shall be handled later, because of intrinsic complexities of the diacritics. Other Brahmi-based languages will be eventually supported, since in some languages, Unicode support remains a major issue.

For each language a detailed layout will be presented as explained in 6. below:

6. LAYOUT

Each language will have the following layout :

1 The generic syllable structure provided by the ABN Formalism shall be suitably modified to respective language.

2 Restriction rules if such rules apply.

3 Sample Examples.

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

5 A map of the above code-chart specifying accurately the above code-chart shall be provided.

6 Finally to reduce the risk of spoofing a variant Table will be provided where the possible variants shall be listed. **As far as possible these variants shall not be individual characters but ligatures that are close homographs.**

Normalization shall be part of the variant table and shall be provided as a safety measure, although a large number of browsers automatically normalize.

7. VARIANT TABLES TO BE DETERMINED AS PER THE SCRIPT

The Variant tables will be determined by the Script. However for the sake of convenience and ease of reference, it is desirable that wherever more than one languages use a single script, separate variant tables be provided for each language, all the more so, since the variant table of Hindi involves normalization whereas that of Marathi, Nepali do not need any such device.

10. TECHNICAL DETAILS

1. BACKUS-NAUR FORMALISM EXPLAINED WITH EXAMPLES

Based on the behavior of the Brahmi syllable, the formalism with suitable emendations has bee successfully deployed both in ISCII and in UNICODE and has been adopted by both standards for explaining the behavior of the syllable. Suitably modified to suit the requirements of IDN norms of Letter-Hyphen-Digit, the BN formalism becomes the backbone on which the IDN structure relies.

2. Augmented Backus-Naur Formalism (ABNF):

The ABNF(Augmented Backus-Naur Formalism) is made generic to support all the languages coming under IDN Project currently. The restriction rules take the ABNF from generic form to language specific form so that it fully satisfies the language specific norms. When applied to IDN, the Backus-Naur formalism results in the following formalism, as shown step-wise:

a. Naming of Variables:

Dash → Hyphen -

Digit → Indo-Arabic digits [0-9]

C → Consonant

V → Vowel

M → Matra

D → Anusvara/Bindi/Tippi/Sunna/Bindu

B → Chandrabindu/Anunasika/Arasunna

X → Visarga/Aytham

H → Halant/Chandrakala/Virama

A → Addak

N → Nukta

Y → Avagraha/Praslesham

L → Chillu

Z → Khanda Ta

k → Number of Consonant Halanta Sequence

b. Comparison Chart

The following chart shows, whether the variable is present in the language or not.

	C	V	M	D	B	X	H	A	N	Y	L	Z
Hindi	✓	✓	✓	✓	✓	✓	✓		✓	✓		
Marathi	✓	✓	✓	✓	✓	✓	✓			✓		
Konkani	✓	✓	✓	✓	✓	✓	✓			✓		
Nepali	✓	✓	✓	✓	✓	✓	✓			✓		
Sanskrit	✓	✓	✓	✓	✓	✓	✓			✓		
Oriya	✓	✓	✓	✓	✓	✓	✓		✓	✓		
Punjabi	✓	✓	✓	✓		✓	✓	✓	✓			
Gujarati	✓	✓	✓	✓	✓	✓	✓			✓		
Bangla	✓	✓	✓	✓	✓	✓	✓		✓	✓		✓
Assamese	✓	✓	✓	✓	✓	✓	✓		✓	✓		✓
Tamil	✓	✓	✓			✓	✓					
Telugu	✓	✓	✓	✓	✓	✓	✓			✓		
Kannada	✓	✓	✓	✓		✓	✓			✓		
Malayalam	✓	✓	✓	✓		✓	✓			✓	✓	

c. ABNF Operators:

Sr. No.	Symbols	Functions
1	“ | ”	Alternative
2	“ [ ] ”	Optional
3	“ * ”	Variable Repetition
4	“ ( ) ”	Sequence Group

d. **ABNF Structure:**

consonant-syllable→

*k(C[N]H) C[N] [H|D|B|X|BD|BX|M[D|B|X|BD|BX]]

| [CH]Z

| L[HC[D|H|M[D]]]

| AC[D|X|M[D|X]]

vowel-syllable → V[D|B|X|BD|BX]

Syllable → consonant-syllable [Y] | vowel-syllable[Y]

IDN-Label → (Syllable | digit)*([dash](Syllable | digit))

e. Restriction Rules:

Restriction rules are the additional filters which when applied to generic ABNF, it results to a Language specific ABNF.

Language-wise restrictions are as follows:

Hindi:

1. Maximum permissible number of consonants to form a syllable up to hence k = 3.
2. BD and BX combinations will be Non-existent.
3. Nukta will be allowed only after following characters:

   क (0915)

   ख (0916)

   ग (0917)

   ज (091C)

   ड (0921)

ढ (0922)

फ (092B)

Marathi:

1. Maximum permissible number of consonants to form a syllable up to 4 i.e k = 3.
2. BD and BX combinations will be Non-existent.
3. With consonant ऱ(0931), only following combinations will be allowed.

च्य → ऱ(0931) ◌् (094D) य (092F)

ऱ्ह → ऱ(0931) ◌् (094D) ह (0939)

Konkani:

1. Maximum permissible number of consonants to form a syllable up to 4 i.e k = 3.
2. BD and BX combinations will be Non-existent.
3. With consonant ऱ(0931), only following combinations will be allowed.

च्य → ऱ(0931) ◌् (094D) य (092F)

ऱ्ह → ऱ(0931) ◌् (094D) ह (0939)

Nepali:

1. Maximum permissible number of consonants to form a syllable up to 4 i.e k = 3.
2. BD and BX combinations will be Non-existent.
3. With consonant ऱ(0931), only following combinations will be allowed .

च्य → ऱ(0931) ◌् (094D) य (092F)

ੜ੍ਹ → ੜ(0931) ◌੍ (094D) ਹ (0939)

Sanskrit:

1. Maximum permissible number of consonants to form a syllable up to 5 i.e k = 4.
2. BD and BX combinations will be Non-existent.

Oriya:

1. Maximum permissible number of consonants to form a syllable up to 4 i.e k = 3.
2. BD and BX combinations will be Non-existent.
3. Nukta will be allowed only after following characters:

   ଡ (0B21) and ଢ (0B22)

Punjabi:

1. Maximum permissible number of consonants to form a syllable up to 2 i.e k = 1.
2. BD and BX combinations will be Non-existent.
3. Halant shall be permitted only with the following:

   Consonant + ◌੍(0A4D) + ਯ(0A2F)

   Consonant + ◌੍(0A4D) + ਰ(0A30)

   Consonant + ◌੍(0A4D) + ਵ(0A35)

   Consonant + ◌੍(0A4D) + ਹ(0A39)

4. Addak : It is used for germination

   ਅੱਕਾ अक्का

a. Addak shall not be permitted either at the beginning or at the end of the word :ੱਕ

b. Addak will not be permissible after halant, visarga or bindi/tippi.

5. Tippi will be used in place of bindi if the preceding character is one of the following:

a. A consonant

b. Short e matra ਿ (0A3F)

c. long and short u matra ੁ(0A41) ੂ(0A42)

d. Vowel ਅ (0A05) or ਇ (0A07)

6. Nukta will be allowed only after following characters:

ਖ (0A16)

ਗ (0A17)

ਜ (0A1C)

ਫ (0A2B)

ਲ (0A32)

ਸ (0A38)

Gujarati:

1. Maximum permissible number of consonants to form a syllable up to 4 i.e k = 3.

2. BD and BX combinations will be Non-existent.

Bengali:

1. Maximum permissible number of consonants to form a syllable up to 4 i.e k = 3.
2. CH will be permissible with Khanda Ta only if C is র (09B0).
3. Khanda Ta will not be allowed at the beginning of an IDN label.
4. Nukta will be allowed only after following characters:

   ড (09A1)

   ঢ (09A2)

   য (09AF)

Assamese:

1. Maximum permissible number of consonants to form a syllable up to 4 i.e k = 3.
2. CH will be permissible with Khanda Ta only if C is ৰ (09F0).
3. Khanda Ta will not be allowed at the beginning of an IDN label.
4. Nukta will be allowed only after following characters:

   ড (09A1)

   ঢ (09A2)

   য (09AF)

Tamil:

1. Maximum permissible number of consonants to form a syllable up to 3 i.e k = 2.
2. BD and BX combinations will be Non-existent.
3. Visarga/Aytham ஃ(0B83) will not be allowed after Matra.

Telugu:

1. Maximum permissible number of consonants to form a syllable up to 3 i.e k = 2.
2. BD and BX combinations will be Non-existent.

Kannada:

1. Maximum permissible number of consonants to form a syllable up to 4 i.e k = 3.
2. BD and BX combinations will be Non-existent.

Malayalam:

1. Maximum permissible number of consonants to form a syllable up to 4 i.e k = 3.
2. BD and BX combinations will be Non-existent.
3. “H” will be permitted after only one “L” ie.ൗ(0D7B) and the following consonant must be ഠ(0D31).

3. Some Examples to show valid ABNF Label:

Below are some sample IDN Labels which are permissible.

1. Generic combination permitted in all languages.

   a. Consonant+Matra)CM)

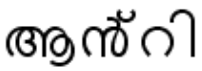
Language	IDN Label
Hindi	ताल
Gujarati	શિક્ષા
Punjabi	ਗੁਰੂ
Oriya	ଢିନ
Bengali	চাল
Tamil	கிடங்கு
Telugu	మాలా

Malayalam	കിണര്
Kannada	ಇಗ

2. Some special cases:

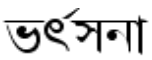
a. Malayalam :

Chillu+Halant+Consonant+Matra (LHCM)

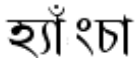


b. Bengali :

Letter Ra+Halant+Khando Ta (CHZ)

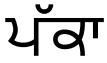


Consonant+Matra+Chandrabindu+Anusvar (CMBD)



c. Punjabi :

Consonant+Addak+Consonant+Matra (CHCM)



Appendix I

S.No	Script	Language	IDN ccTLD	English Transliteration
1.	Bengali Extended Bengali	Assamese	ভাৰত	**Bharat**
2.		Bangla	ভারত	**Bharat**
3.		Manipuri (*Meetei Mayek*)	ভারত	**Bharat**
4.	Devanagari	Hindi	भारत	**Bharat**
5.		Marathi	भारत	**Bharat**
6.		Bodo	भारत	**Bharat**
7.		Dogri	भारत	**Bharat**
8.		Nepali	भारत	**Bharat**
9.		Sanskrit	भारतम्	**Bharatam**
10.		Maithili	भारत	**Bharat**
11.		Santhali	**भारोत**	**Bharat**
12.		Konkani	**भारोत**	**Bharat**
13.	Gujarati	Gujarati	ભારત	**Bharat**
14	Gurmukhi	Punjabi	ਭਾਰਤ	**Bharat**
15	Kannada	Kannada	ಭಾರತ್	**Bharat**
16	Malayalam	Malayalam	ഭാരതം	**Bharat**

17	Perso-Arabic	Kashmiri	بارت	**Bharat**
18		Urdu	ہندوستان	**Hindostan**
19		Sindhi	ڀارت	**Bharat**
20	Oriya	Oriya	ଭାରତ	**Bharat**
21	Tamil	Tamil	இந்தியா	**Indiya**
22	Telugu	Telugu	భారత్	**Bharat**

11. REFERENCES

Bureau of Indian Standards., *Indian Standard. Indian Script Code for Information Interchange.* IS 13194:1991 New Delhi, 1993.

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

Hutchison. London. 1968.

ICANN : www.icann.org

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 Devanagari 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

Library of Congress. *Romanization Standards.*. USA. 2002

RFC 3454 : www.ietf.org/rfc/rfc3454.txt

RFC 3492 : RFC for PUNYCODE www.ietf.org/rfc/rfc3492.txt

Satpute. P. Open Type Font processing: some issues. *Proceedings of the NWCT*. 2007

Stone. Anthony., http://homepage.ntlworld.com/stone-catend/trind.htm*:*

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).