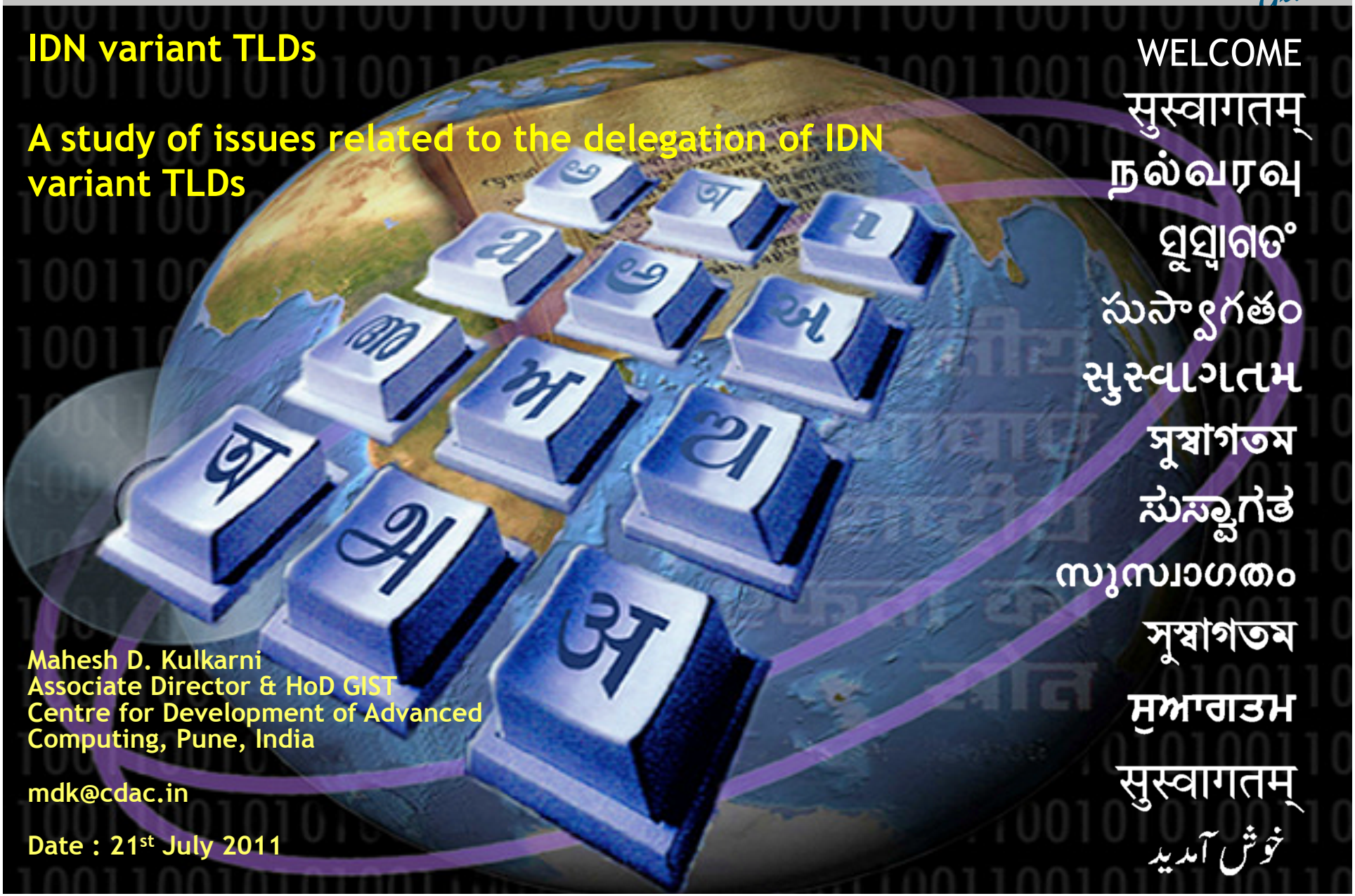
IDN variant TLDs

A study of issues related to the delegation of IDN variant TLDs



Mahesh D. Kulkarni
Associate Director & HoD GIST
Centre for Development of Advanced Computing, Pune, India

mdk@cdac.in

Date : 21st July 2011

The Multilingual diversity of India - Some facts & Figures

- The Constitution recognizes 22 languages termed as Scheduled Languages.
- Two major script systems are used: Perso-Arabic based and Brahmi based
- Sindhi, Kashmiri, Urdu use the Perso-Arabic system with notational changes in Sindhi. The remaining 19 languages use 11 derivations of the Brahmi script.
- One to many and many to one relationship between language and script.
- Santali & Sindhi use more than one script.
- Devanagari script is used for Sanskrit, Hindi, Marathi, Nepali, Konkani, Maithili, Dogri, Bodo

Indian language complexities

Syllable formation level

ब्रह्मा

ब्र = ब + ् + र.

ह्मा = ह + ् + म + ा.

Alternate spellings

अन्न : अन्न्न अन्न

बिट्टु : बिट्ट्टु

Rendering order level

क ि त ा ब

क ि त ा ब

(179+219+194+218+202)

किताब

(ि+क+त+ा+ब)

Alternate forms

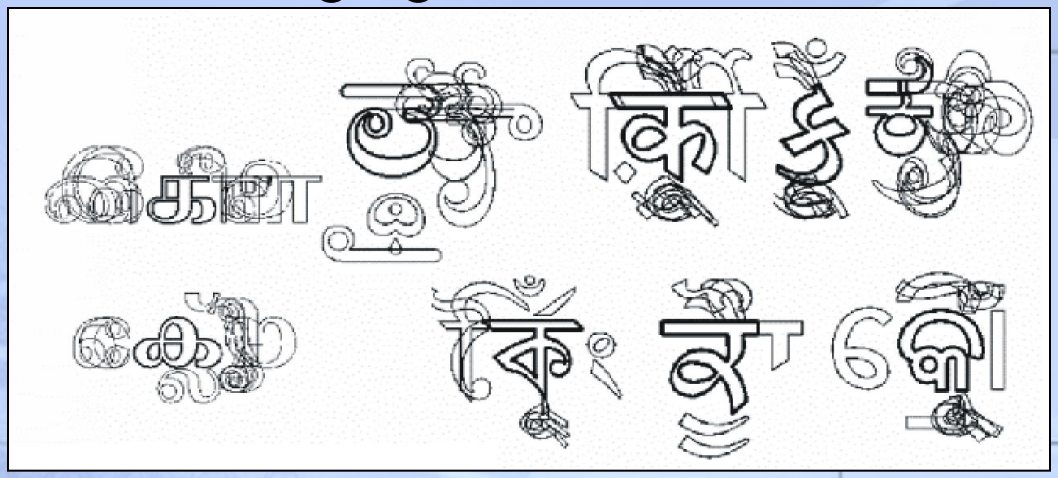
तसवीर तस्वीर

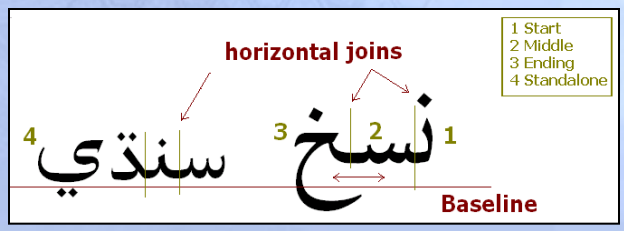
हिंदी हिन्दी

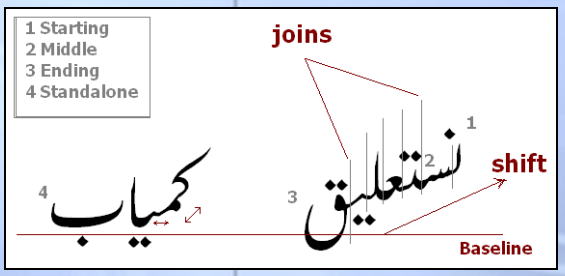
Different inputting mechanism in Indian languages

Need for Variant Identification : Indian Language Scenario

Most Indian languages are Multi-tier in nature







When conjuncts come in picture, resulting glyph shapes increase manifolds.

क ◌् ष = क्ष

0915 + 094D + 0937

Types of Variants :

1. **Homographic variants : Similar Looking**
   - 1 / l in Latin
   - द्र / द्न in Devanagari

2. **Homophonic variants** : Similar Sounding / alternate spelling
   - color / colour in Latin
   - हिंदी / हिन्दी in Devanagari

3. **Case variants :**
   - C / c in Latin (No such case in Indian Languages)

Homographic variants – confusingly similar:

- Most of the browsers and applications using IDNs display labels in minimal size.
- This results in maximum number of spoofing and phishing attacks.
- Multi-tier scripts such as used in Indian languages are less readable in the address bar.
- Unicode normalization rules have also been considered as variants

Homographic Variants :

Telugu Variants:

ಬ್ಧ 0CAC+0CCD+0CA6	ಬ್ಧ 0CAC+0CCD+0CA7	ಬ್ಧ 0CAC+0CCD+ 0CB2
ಧಿ 0CA5+0CBF	ದಿ 0CA6+0CBF	ಧಿ 0CA7+0CBF

Tamil Variants :

ஒள 0B92+0BB3	ஒள 0B94

Homophonic Variants & Alternate spellings:

- Valid Homophones : हिंदी versus हिन्दी
- Common Misspellings : इण्डिया versus इन्डिया

- While formulating the IDN policy for .in we have not considered these variants as historically other domains have always considered alternate spellings of www.color.com and www.colour.com as separate entities

Case Variants

- Case variants are not applicable in case of Indian Languages
- However Indian languages are rich in synonyms

Need for Variant Identification

- Invisible characters like ZWJ and ZWNJ can greatly amount to visual spoofing possibilities.
  - If permitted, their placement within the Domain Name/Label should be restricted to only most compulsory cases.
- In some cases, within the same script, two languages need different conjunct formation rules.
- Across the Operating systems, Rendering Engines and their versions, the rendering is not same.

Need for Variant Identification

Indian scripts introduce syllabic variants

श्व / श्व

Such homographs need to be considered while identifying variants

श ् र ् व = श्व

0936 094D 0930 094D 0935

श ् व = श्व

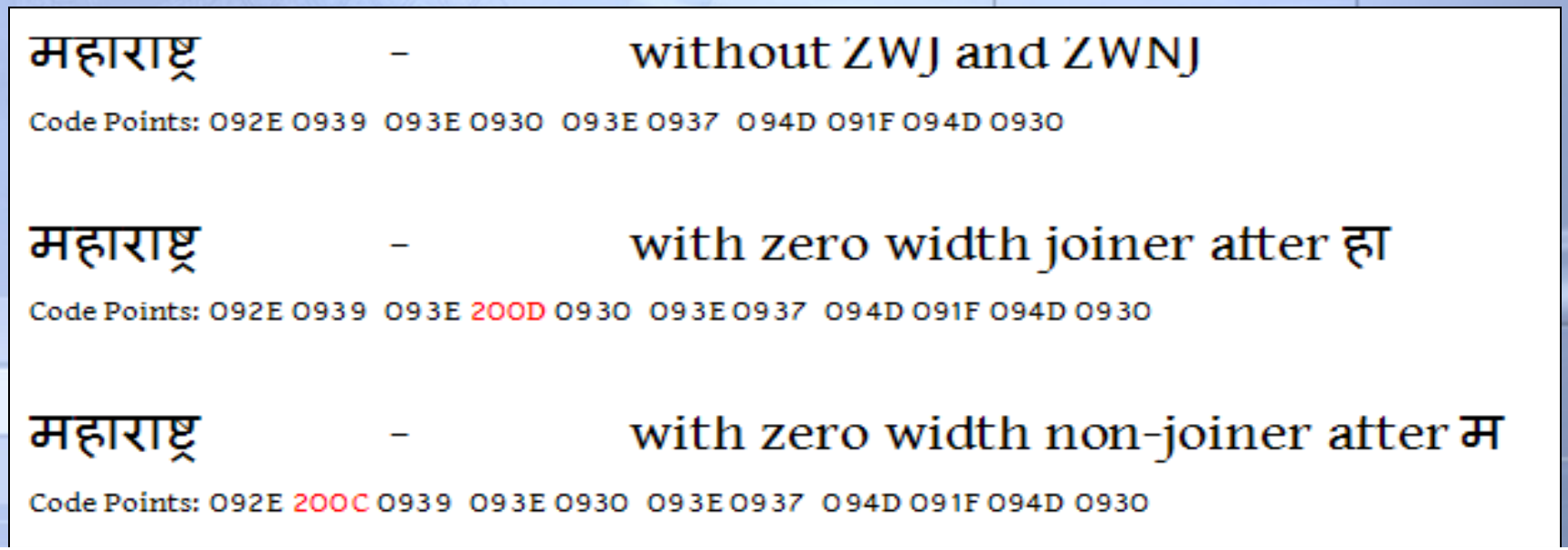
0936 094D 0935

Need for Variant Identification : Display Aspect

ZWJ and ZWNJ :

Invisible characters like ZWJ and ZWNJ can greatly amount to visual spoofing possibilities. A clear decision needs to be taken regarding their inclusion in TLDs and if included, their placement within the Domain Name/Label.

Examples :



IDN Variants TLDs

- .com is to commercial, since first three letter of English meaningful word
- In English one can easily correlate the short forms with the type of activity / content the domain may have.
- Transliteration can not always be acceptable for following reasons.
  - Some scripts may not have characters necessary to represent the sound of the words
    - E.g Tamil does not have “Bha” bharat will map to parat
  - Associating the transliterated IDNs with real world will be difficult
  - May convey entirely different meaning in other languages / region.
  - In Indian languages short form does not exists.

Examples

- Example word "PAL"
- In Tamil -> பால means Milk
- In Marathi PAL -> पाल means Lizard

IDN Variants TLDs

- Another solution is to translate the TLDs in different languages.
- However, since the TLD do not convey the language information, it is likely that a translation suitable for one region may not be suitable for other (because of regional translation requirements).
- This issue is more specific where the scripts / languages are shared across borders.

Need for checking well formed-ness of labels

Rendering Engine lacunae:

The well formed word “किताब” as seen in Address of Safari (Version - 3.1.2 (4525.22)) on MAC OS Version -10.4.11 (Tiger)

Actual display

Expected display

किताब

Bidi Algorithms needed for Urdu, Sindhi and Kashmiri are more complex.

Need for checking well formed-ness of labels

Rendering Engine lacunae:

An ill-formed word composed of sequence
0915 + 093F + 094D + 0924+ 093E + 092C
as seen in IE (Version 8) on Windows XP

Actual display

Expected display कि्ताब

Some applications are incapable of showing IDN labels and show punycode instead.

Indian IDNs in Browser Address Bar – rendering issues

Various Operating Systems and Browser combinations were considered during testing of IDN .in (.भारत)

www.বাংলা.cdac.in

In the URL, extra spaces are introduced after the Indian language domain text.

In the URL, font size of Indian language domain text is less; so proper rendering is not possible for Bengali text.

Implementation of IDNs in .in(.भारत) ccTLD

- A formalism based on ABNF has been put in place to validate desired domain name for each language based on syllabic structure.
- The applicable character sets for all official languages have been identified from the respective Unicode code charts for the script of the language.
- No intermixing of scripts is allowed
- Variant rules have been formalized for Domain Name label.
- Variants occurring syllables have been identified within each language.
- The variant set has been kept optimal ensuring safety of citizens without being too restrictive.

- Link : http://pune.cdac.in/html/gist/down/idn_d.asp

Best practices that can be carried forward in TLD

Suggested Qualification Criterion for IDN TLD :

- Validation as per Formalism
  - Proper length, proper character set, proper formation
- Non variant nature with any of the pre-registered tld
- Presence of symbols (Currency, logos, sentiment) should be avoided.
  - Tonal stress markers are needed for languages such as Bodo and should be permitted. Example code point 02BC is required for languages Bodo, Dogri, Assamese, Maithil is not part of the respective code pages.
- Political/Stakeholder opinions

Thank you

www.cdac.in

www.सीडैक.com

http://www.xn--11bx2e6a3b.com/