Developing the root zone Label Generation Rules for Neo-Brahmi Scripts

By,
Akshat Joshi
C-DAC GIST
Venue: HICC, Hyderabad
7th Nov. 2016
@kshatj
What is Brahmi?

- An ancient script

- Most of the modern scripts in Indian subcontinent have been derived from Brahmi.
  - Geographically the scripts being used in Central Asia, South Asia and South-East Asia

- These scripts are used by multiple language families: Largely by Indo-Aryan and Dravidian
Why Brahmi?

• Despite their variations in the visual forms, the basic philosophy in their usage is common

• They all are “akshar” driven, and follow a specific syntax
  – Analogical reference can be made to Indian National standard, IS 13194:1991 – Section 8

• This syntax being the implicit foundation in representation of these scripts in the digital medium, adherence to the structure acts as a obligatory security consideration even in the case of Internationalized Domain Names.
Why Neo-Brahmi?

• Of all the scripts derived from “Brahmi”, not all are in modern usage

• Approach is in consonance with the “Conservatism Principle" of the LGR procedure.
Neo Brahmi GP - Current Status

- Mixed bag expertise like linguistic, Unicode

- **Chair** Udaya Narayana Singh - Bangla, Maithili, Hindi, English
- Anupam Agrawal - Hindi, Bangla
- Akshat Joshi - Hindi, Marathi
- Abhijit Dutta - Bengali, Hindi
- Mahesh Kulkarni - Marathi, Hindi
- Neha Gupta - Hindi
- Nishit Jain - Hindi
- Prabhakar Pandey - Hindi
- Raiomond Doctor - English, Hindi, Marathi, Gujarati
- N. Deiva Sundaram - Tamil
- Shantaram Walawalikar - Konkani
- Bal Krishna Bal - Nepali
- Ganesh Murmu – Santali
- Balaram Prasain - Nepali
- Rajib Chakraborty - Bangla
- Gurpreet Singh Lehal - Panjabi
- Saroja Bhate - Sanskrit
- Shambhu Kumar Singh - Maithili
- Swarna Prabha Chainary - Bodo
- Ghanashyam Nepal - Nepali
- Kalyan Vasudeo Kale - Marathi
- Shashi Pathania - Dogri
- Santhosh Thottingal - Malayalam, Saurashtra, Tamil
- Uma Maheshwar G - Telugu
- Girish Chandra Mishra - Odia
- K. C. Tikayat ray - Odia
- Debajit Sharma - Assamese
- Basanta Kumar Panda - Odia
- Arvind Bhandari - Gujrat
- Harish Chowdhary - Hindi
Neo Brahmi GP – Approach

• There are cases of
  – one script one language
  – one script multiple languages
    • In this case, multiple sub-groups may exist to ensure proper representation of each language

• Each sub-group ideally would comprise of
  – Language expert(s)
  – Community representative(s)
Neo Brahmi GP – Approach

Generation Panel

Chair

Policy Experts
Unicode/IDNA Experts
Registry/Registrar operation Experts

Subgroup
Language Experts
Community Representatives

Subgroup
Language Experts
Community Representatives

...
Neo-Brahmi GP Internal Composition

Integration Panel

Neo-Brahmi GP

Tamil Sub-Group

Telugu SG

Devanagari SG

Gujarati SG

Gurmukhi SG

Bengali SG

Bangla

Assamese

Manipuri

Hindi

Marathi

Konkani

Nepali

Bodo

Dogri

Maithili

Santhali

Bengali SG

SCRIPT SECTION

LANGUAGE SECTION
Neo Brahmi GP – Outreach Efforts

• Conducted a workshop in AprIGF-2014 for awareness and call for participation in LGR procedure.
  – Topic: “Bringing diverse linguistic communities together for a unified IDN ruleset”
  – The panel discussion touched upon the various aspects of creation of the LGR for the Neo-Brahmi scripts

• Participation and presentation in 49th ICANN Public meeting at Singapore

• Participation and presentation in 50th ICANN Public meeting at London
Root LGR procedure

• **Fundamental Blocks:**
  
  – Code point repertoire
  
  – Variant Rules
  
  – Whole Label Evaluation rules

ICANN String Similarity Assessment Tool
Previous similar work

• For IDN version of “.in” ccTLD, (.bharat) equivalent in 22 Official Indian Languages, similar exercise had been carried out

• Following things were finalized for each language
  – Permissible set of code points
  – Visually similar variant strings
  – Complex whole label evaluation rules
Revisiting the rules in context of LGR framework

• LGR work is different in following contexts
  – Wider stakeholder group
  – Overarching principles in the LGR procedure
    • Especially *Simplicity* and *Predictability* principles

• This revision however would not change
  – the need for the well-formedness of the label in terms of Akshar formalism
Status of current work

• In the process of finalizing the code-point repertoires:

![Devanagari](image1.png)  ![Bengali](image2.png)
Future undertakings: Cross-Script Similarities

<table>
<thead>
<tr>
<th>DEVANĀGARĪ SCRIPT</th>
<th>COGNATE SCRIPT</th>
<th>CODEPOINT IN COGNATE SCRIPT</th>
</tr>
</thead>
<tbody>
<tr>
<td>घU+0918</td>
<td>Gujarati</td>
<td>धU+0A98</td>
</tr>
<tr>
<td>उU+0909</td>
<td>Gurmukhi</td>
<td>उU+0A24</td>
</tr>
<tr>
<td>रU+0930</td>
<td>Gujarati</td>
<td>रU+0AAE</td>
</tr>
</tbody>
</table>

- Code point similarity across scripts

- Cases where Devanagari-Gujarati and Devanagari-Gurumukhi strings look similar.
Future undertakings: Whole Label Evaluation Rules

• Most crucial aspect of Neo-brahmi Label Generation Ruleset

• Details in the following slides.
Before starting with the Whole Label Evaluation Rules for LGR (The global approach)

let us take a look at Whole Label Evaluation Rules for .bharat policy (The Indian approach)
Why understanding the .bharat policy is important?

- It is founding work connecting IDNs and Indian languages
- It has been demonstrated and appreciated at various National and International forums
- It has **all the basic components that are required by the “Root LGR” work**, albeit in different forms.
Character classification

Components of the Syllable

- Consonants (C):
  कखगधङचछजझञटठडढणतथन्यऱलळऴवशषसह

- Vowels (V):
  अआइईउऊऋऐ-awesome

- Matras (M):
  ािंिािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािािं

- Vowel modifiers (D):
  ूूृॄॅॆेैॉौ

- Halant (H):
  ्

- Nukta (N):
  ्
Formalism at a glance ...
Formalism Illustrated...

• Variables:

Dash → Hyphen -
Digit → Indo-Arabic digits [0-9]
C → Consonant
V → Vowel
M → Matra
D → Anusvara/Bindi/Tippi/Sunna
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 possible Consonant Halanta Sequences
Formalism Illustrated...

• Formalism Operators:

  | → Alternative
  [ ] → Optional
  * → Variable Repetition
  () → Sequence Group
Formalism Illustrated...

The Formalism:

Consonant-Syllable $\rightarrow$

*\(k(C[N]H)\) C[N] [H|D|B|X|BD|BX|M[D|B|X|BD|BX]]
| [CH]Z
| L[HC[D|H|M[D]]]
| A[C[D|X|M[D|X]]

Vowel-Syllable $\rightarrow$ V[D|B|X|BD|BX]

Syllable $\rightarrow$ Consonant-Syllable [Y] | Vowel-Syllable[Y]

IDN-Label $\rightarrow$ (Syllable | digit)*([dash](Syllable | digit))
Fundamental differences .bharat and Root LGR

.bharat zone
• It is a focused zone – only the domain names under .bharat TLD
• Restricted only to Indian languages
• Policies can be strict
• Can define our own categories

Root Zone
• It is the most generic zone on the Internet. The root zone.
• Cannot be restricted. Encompasses all the scripts/languages of the world
• Policies have to be simple, yet sufficiently tight
• Have to rely on the Unicode Character properties
## Character classes - Differences

### Bharat character classes

<table>
<thead>
<tr>
<th>Character</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>Consonant</td>
</tr>
<tr>
<td>V</td>
<td>Vowel</td>
</tr>
<tr>
<td>M</td>
<td>Matra</td>
</tr>
<tr>
<td>D</td>
<td>Anusvara/BindiSunna</td>
</tr>
<tr>
<td>B</td>
<td>Chandrabindu/Arasunna</td>
</tr>
<tr>
<td>X</td>
<td>Visarga/Aytham</td>
</tr>
<tr>
<td>H</td>
<td>Halant/Chandrakala</td>
</tr>
<tr>
<td>A</td>
<td>Addak</td>
</tr>
<tr>
<td>N</td>
<td>Nukta</td>
</tr>
<tr>
<td>Y</td>
<td>Avagraha/Praslesham</td>
</tr>
<tr>
<td>L</td>
<td>Chillu</td>
</tr>
<tr>
<td>Z</td>
<td>Khanda Ta</td>
</tr>
</tbody>
</table>

### Unicode character classes

- **Mn - Mark, Non-Spacing**
  - 0901;DEVANAGARI SIGN CANDRABINDU
  - 093A;DEVANAGARI VOWEL SIGN OE
  - 093C;DEVANAGARI SIGN NUKTA
  - 094D;DEVANAGARI SIGN VIRAMA
- **Mc - Mark, Spacing Combining**
  - 0903;DEVANAGARI SIGN VISARGA
  - 093E;DEVANAGARI VOWEL SIGN AA
- **Lo - Letter, Other**
  - 0905;DEVANAGARI LETTER A
  - 0915;DEVANAGARI LETTER KA
  - 093D;DEVANAGARI SIGN AVAGRAHA
thank you