A UNICODE PRIMER

Neo-Brāhmī Generation Panel
16 July 2015

Agenda

1. From Stone to Chip
2. Fundamentals
3. Why Unicode: Need for Standards
4. What is Unicode ?
5. Basic Features of Unicode

FROM STONE TO CHIP

FROM STONE TO CHIP

Data Preservation has a fascinating history:

Text inscribed on Stones, Palm Leaves, Copper-plates, Clay Tablets gave way to Wooden Block printing by the 9th Century and by the 15th Century with Gutenberg's moveable metal type printing press created the Gutenberg Galaxy which has survived for over 5 centuries and is slowly being replaced by the Digital Galaxy.

Fundamentals:

Unlike the earlier means for data preservation, storage on chips has an important difference. A book can be replicated and shared and will be legible all those who share it. Sharing of electronic data which can ensure that anybody can access the data posed a major issue.

Fundamentals: Issues with digital data

¤ÌÛÂÚË ³á ´ÛÑÚÉ ÂÛÏ¢µá ³á ¤ÈÌÚÆ ³Ú ÌÚÌÑÚ, ÊÍÚÆ ÄÏè º

µÚºÛÍÚÊÚÄ, Ì¢µÑÔÚÏ, 14 ºÝÑÚ§ 2015 (08:06 IST)

µÚºÛÍÚÊÚÄê ÌÝ´èÍ ÆèÍÚÍÛ³ Ìº Û×è½èÏá½ ³Ü ¤ÄÚÑÂ Æá ³ÃÛÂ ÏÞÈ ×á ÂÛÏ¢µá ³á

¤ÈÌÚÆ ³á ÑÛ¬ ¤ËÛÆáÂÚ ¤ÌÛÂÚË Ê¸è¸Æ ±Ï ¤ËÛÖá³ Ê¸è¸Æ ³á ´ÛÑÚÉ

ÌÚÌÑÚ ÄÏè º ³ÏÚÆá ÔÚÑá Ôèí³èÂÛ ³Ú ×åÌÔÚÏ ³å ÍØÚ¢ ÊÍÚÆ ÄÏè º ³ÛÍÚê

अमिताभ के खिलाफ तिरंगे के अपमान का मामला, बयान दर्ज

गाजियाबाद, मंगलवार, 14 जुलाई 2015 (08:06 IST)

गाजियाबाद। मुख्य न्यायिक मजिस्ट्रेट की अदालत ने कथित रूप से तिरंगे के 'अपमान' के लिए अभिनेता अमिताभ बच्चन और अभिषेक बच्चन के खिलाफ मामला दर्ज कराने वाले व्यक्ति का सोमवार को यहां बयान दर्ज किया।

Fundamentals: Issues with Digital data

Before Unicode was invented, there were hundreds of different encoding systems for assigning these numbers. No single encoding could contain enough characters: for example, the European Union alone requires several different encodings to cover all its languages. Even for a single language like English no single encoding was adequate for all the letters, punctuation, and technical symbols in common use.

These encoding systems also conflict with one another. That is, two encodings can use the same number for two different characters, or use different numbers for the same character.

To resolve this confusion - in the late eighties, it was decided to have a single unified standard called Unicode. The aim of Unicode is to assign a unique number to each and every character of each and every "reasonable" writing system used on the earth, “no matter what the platform, no matter what the program, not matter what the language”.

This ensure that data entered with a keyboard using Unicode compliant fonts is shared seamlessly with another machine which may or may not necessarily have the font.

Fundamentals: Issues with Digital data

This involves standards. Adherence to standards ensures compatibility, safeguarded data, avoids vendor locking, proper exchange of data between various systems, applications, databases, devices, etc.
Standards exist for Input mechanisms: Keyboards, Display and Storage. GIST has been at the forefront in development of standards for all three.
However this presentation will focus only on Unicode and its relevance to Internationalised Domain Names.

UNICODE

- Storage standard
- Unicode is for all languages of the world: *the world speaks Unicode.*
- Enables seamless exchange of data – desktops, printers, databases, browsers, devices.

UNICODE

- *The Unicode consortium defines Unicode as :*

- *"Unicode is the universal character encoding, maintained by the Unicode consortium. This encoding standard provides the basis for processing, storage and interchange of text data in any language in all modern software and information technology protocols."*

- *It is the superset of all the languages in the world which also includes punctuation, special characters (shapes), currency symbols, mathematical symbols etc. Using Unicode, more than 65000 different characters can be represented. Unicode comprises of many code charts.*
- *The Unicode code charts can be referred at: http://www.unicode.org/charts*

Unicode Features

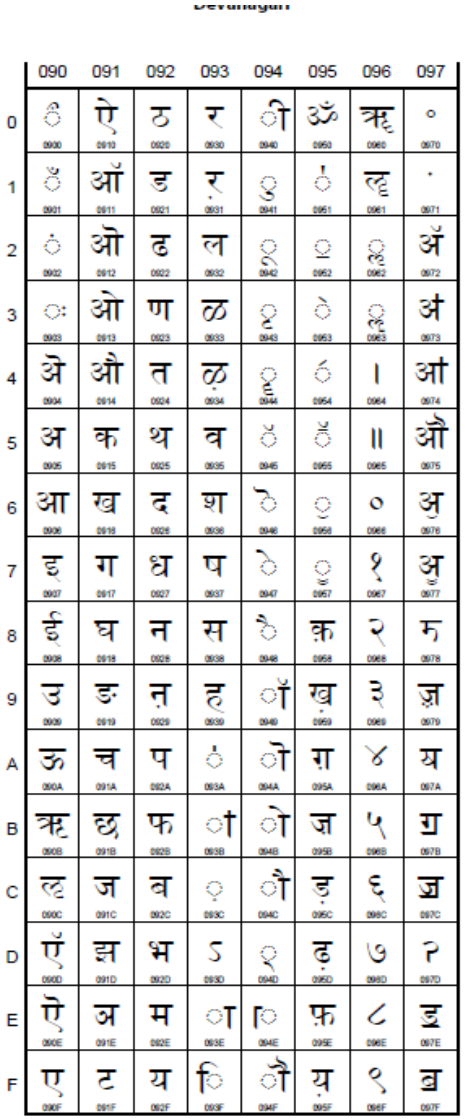
In what follows using the Devanagari code-block, basic features of Unicode required for an understanding of IDN will be presented:

1. Unicode's main source is ISO 10646 which maintains the world's scripts. However Unicode is also Community Based and receives proposals from both members as well as individuals. Each proposal has to be substantiated with concrete evidence for introducing a given character. The proposals are put up for comments and eventually accepted

2. Unicode is script based. Each script is stored in "Blocks". Each code-block bears the name of the Script to which it refers. Unicode has a total of 262 code blocks of which 160 are in the active plane termed as BMP or **Basic Multilingual Plane** and the remainder refer to obsolete scripts which are no longer in use are stored in the **Plane 1**, the **Supplementary Multilingual Plane (SMP)**, which contains historic scripts such as _Linear B, Egyptian hieroglyphs, and cuneiform scripts.

Unicode Features

3. Thus Devanagari script is stored in a code-block (often termed also as code-page or code-chart).



	090	091	092	093	094	095	096	097
0	ऀ 0900	ऐ 0910	ठ 0920	र 0930	ी 0940	ॐ 0950	ॠ 0960	॰ 0970
1	ँ 0901	ऑ 0911	ड 0921	ऱ 0931	ु 0941	॑ 0951	ॡ 0961	ॱ 0971
2	ं 0902	ओ 0912	ढ 0922	ल 0932	ू 0942	॒ 0952	ॢ 0962	ॲ 0972
3	ः 0903	ओ 0913	ण 0923	ळ 0933	ृ 0943	॓ 0953	ॣ 0963	ॳ 0973
4	अ 0904	औ 0914	त 0924	ऴ 0934	ॄ 0944	॔ 0954	। 0964	ॴ 0974
5	आ 0905	क 0915	थ 0925	व 0935	ॅ 0945	ॕ 0955	॥ 0965	ॵ 0975
6	इ 0906	ख 0916	द 0926	श 0936	े 0946	ॖ 0956	० 0966	ॶ 0976
7	ई 0907	ग 0917	ध 0927	ष 0937	ै 0947	ॗ 0957	१ 0967	ॷ 0977
8	उ 0908	घ 0918	न 0928	स 0938	ॉ 0948	क़ 0958	२ 0968	ॸ 0978
9	ऊ 0909	ङ 0919	ऩ 0929	ह 0939	ॊ 0949	ख़ 0959	३ 0969	ॹ 0979
A	ऋ 090A	च 091A	प 092A	ऺ 093A	ो 094A	ग़ 095A	४ 096A	ॺ 097A
B	ऌ 090B	छ 091B	फ 092B	ऻ 093B	ौ 094B	ज़ 095B	५ 096B	ॻ 097B
C	ऍ 090C	ज 091C	ब 092C	़ 093C	् 094C	ड़ 095C	६ 096C	ॼ 097C
D	ऎ 090D	झ 091D	भ 092D	ऽ 093D	॑ 094D	ढ़ 095D	७ 096D	ॽ 097D
E	ए 090E	ञ 091E	म 092E	ा 093E	ॎ 094E	फ़ 095E	८ 096E	ॾ 097E
F	ऐ 090F	ट 091F	य 092F	ि 093F	ॏ 094F	य़ 095F	९ 096F	ॿ 097F

Unicode Features

4. Each block is made up of **code-points**. Each code-point refers to a shape i.e. a glyph. The number of code points in a Unicode block is a multiple of 16. Unicode blocks range in size from the minimum of 16 to a maximum of 65,536 code points. A block may contain unassigned code points, which are reserved. In case a code-block gets full, an extended code-block is created. Thus Devanagari also admits an extended Devanagari block.

5. The totality of code-points within a code-block is termed as the **repertoire** of that particular script. All the code-points pertaining to the character of Devanagari in the Devanagari code block constitute its repertoire.

Unicode Features

6. The code-points within a code-block are pertinent to that script and that script alone. No mixing of scripts is permitted. If a given code-point i.e. a character glyph is shared by many scripts, Unicode provides instructions to that effect. Thus the danda and double danda of Devanagari are used by other scripts as in the case of the code-block Bengali and Assamese which covers 0900 -097F

Reserved

For viram punctuation, use the generic Indic 0964 and 0965. Note that these punctuation marks are referred to as dahri and double dahri in Bangla.

09E4 <reserved>
→ 0964 । devanagari danda

09E5 <reserved>
→ 0965 ॥ devanagari double danda

Unicode Features

7. The code-points cover not only the characters but all and every element including diacritics and even weights and measures as in the case of Tamil.

0900 **Devanagari** **094F**

Various signs

0900 ◌ँ DEVANAGARI SIGN INVERTED CANDRABINDU
= vaidika adhomukha candrabindu
0901 ◌ँ DEVANAGARI SIGN CANDRABINDU
= anunasika
→ 0310 ◌̐ combining candrabindu
0902 ◌ं DEVANAGARI SIGN ANUSVARA
= bindu
0903 ◌ः DEVANAGARI SIGN VISARGA

Independent vowels

0904 ऄ DEVANAGARI LETTER SHORT A
0905 अ DEVANAGARI LETTER A
0906 आ DEVANAGARI LETTER AA
0907 इ DEVANAGARI LETTER I
0908 ई DEVANAGARI LETTER II
0909 उ DEVANAGARI LETTER U
090A ऊ DEVANAGARI LETTER UU
090B ऋ DEVANAGARI LETTER VOCALIC R
090C ऌ DEVANAGARI LETTER VOCALIC L
090D ऍ DEVANAGARI LETTER CANDRA E
090E ऎ DEVANAGARI LETTER SHORT E
• Kashmiri, Bihari languages
• also used for transcribing Dravidian short e
090F ए DEVANAGARI LETTER E
0910 ऐ DEVANAGARI LETTER AI
0911 ऑ DEVANAGARI LETTER CANDRA O
0912 ऒ DEVANAGARI LETTER SHORT O
• Kashmiri, Bihari languages
• also used for transcribing Dravidian short o
0913 ओ DEVANAGARI LETTER O
0914 औ DEVANAGARI LETTER AU

Consonants

0915 क DEVANAGARI LETTER KA
0916 ख DEVANAGARI LETTER KHA
0917 ग DEVANAGARI LETTER GA
0918 घ DEVANAGARI LETTER GHA
0919 ङ DEVANAGARI LETTER NGA
091A च DEVANAGARI LETTER CA
091B छ DEVANAGARI LETTER CHA
091C ज DEVANAGARI LETTER JA
091D झ DEVANAGARI LETTER JHA
091E ञ DEVANAGARI LETTER NYA
091F ट DEVANAGARI LETTER TTA
0920 ठ DEVANAGARI LETTER TTHA
0921 ड DEVANAGARI LETTER DDA
0922 ढ DEVANAGARI LETTER DDHA
0923 ण DEVANAGARI LETTER NNA
0924 त DEVANAGARI LETTER TA
0925 थ DEVANAGARI LETTER THA
0926 द DEVANAGARI LETTER DA
0927 ध DEVANAGARI LETTER DHA
0928 न DEVANAGARI LETTER NA
0929 ऩ DEVANAGARI LETTER NNNA
• for transcribing Dravidian alveolar n
≡ 0928 न 093C ◌़
092A प DEVANAGARI LETTER PA
092B फ DEVANAGARI LETTER PHA
092C ब DEVANAGARI LETTER BA
092D भ DEVANAGARI LETTER BHA
092E म DEVANAGARI LETTER MA
092F य DEVANAGARI LETTER YA
0930 र DEVANAGARI LETTER RA
0931 ऱ DEVANAGARI LETTER RRA
• for transcribing Dravidian alveolar r
• half form is represented as "Eyelash RA"
≡ 0930 र 093C ◌़
0932 ल DEVANAGARI LETTER LA
0933 ळ DEVANAGARI LETTER LLA
0934 ऴ DEVANAGARI LETTER LLLA
• for transcribing Dravidian l
≡ 0933 ळ 093C ◌़
0935 व DEVANAGARI LETTER VA
0936 श DEVANAGARI LETTER SHA
0937 ष DEVANAGARI LETTER SSA
0938 स DEVANAGARI LETTER SA
0939 ह DEVANAGARI LETTER HA

Dependent vowel signs

These dependent vowel signs are used in Kashmiri and in the Bihari languages (Bhojpuri, Magadhi, and Maithili).

093A ◌ॊ DEVANAGARI VOWEL SIGN OE
093B ◌ो DEVANAGARI VOWEL SIGN OOE

Various signs

093C ◌़ DEVANAGARI SIGN NUKTA
• for extending the alphabet to new letters
093D ऽ DEVANAGARI SIGN AVAGRAHA

Dependent vowel signs

093E ◌ा DEVANAGARI VOWEL SIGN AA
093F ◌ि DEVANAGARI VOWEL SIGN I
• stands to the left of the consonant
0940 ◌ी DEVANAGARI VOWEL SIGN II
0941 ◌ु DEVANAGARI VOWEL SIGN U
0942 ◌ू DEVANAGARI VOWEL SIGN UU
0943 ◌ृ DEVANAGARI VOWEL SIGN VOCALIC R
0944 ◌ॄ DEVANAGARI VOWEL SIGN VOCALIC RR
0945 ◌ॅ DEVANAGARI VOWEL SIGN CANDRA E
= candra
0946 ◌ॆ DEVANAGARI VOWEL SIGN SHORT E
• Kashmiri, Bihari languages
• also used for transcribing Dravidian short e
0947 ◌े DEVANAGARI VOWEL SIGN E
0948 ◌ै DEVANAGARI VOWEL SIGN AI
0949 ◌ॉ DEVANAGARI VOWEL SIGN CANDRA O
094A ◌ॊ DEVANAGARI VOWEL SIGN SHORT O
• Kashmiri, Bihari languages
• also used for transcribing Dravidian short o
094B ◌ो DEVANAGARI VOWEL SIGN O
094C ◌ौ DEVANAGARI VOWEL SIGN AU

Virama

094D ◌् DEVANAGARI SIGN VIRAMA
= halant (the preferred Hindi name)
• suppresses inherent vowel

Dependent vowel signs

094E ◌ॎ DEVANAGARI VOWEL SIGN PRISHTHAMATRA E
• character has historic use only
• combines with E to form AI, with AA to form O, and with O to form AU
094F ◌ॏ DEVANAGARI VOWEL SIGN AW
• Kashmiri, Bihari languages

Tamil numerics

0BF0 ௰ TAMIL NUMBER TEN
0BF1 ௱ TAMIL NUMBER ONE HUNDRED
0BF2 ௲ TAMIL NUMBER ONE THOUSAND

Tamil calendrical symbols

0BF3 ௳ TAMIL DAY SIGN
= naal, naali/padi
• denotes a measure of grain that equals 2 uri or 4 ulakku
= pillaiyaar suli
• denotes auspiciousness
0BF4 ௴ TAMIL MONTH SIGN
= maatham
0BF5 ௵ TAMIL YEAR SIGN
= varudam

Tamil clerical symbols

0BF6 ௶ TAMIL DEBIT SIGN
= patru
0BF7 ௷ TAMIL CREDIT SIGN
= eduppu
• denotes incoming cash which is set aside for unknown expenses
• sometimes used as the credit sign; the traditional credit sign is distinct
0BF8 ௸ TAMIL AS ABOVE SIGN
= merpadi

Currency symbol

0BF9 ௹ TAMIL RUPEE SIGN
= rupai

Tamil symbol

0BFA ௺ TAMIL NUMBER SIGN
= enn, niluvai
• denotes balance
• sometimes used as the number sign; the traditional number sign is distinct

Unicode Features

8. Each code-point or glyph has a name termed as **Unicode Name** which and also a **unique number** which refers to its position in the code-block. This is normally shown as U+0000 .

Thus क has the following value:

क U+0915 DEVANAGARI LETTER KA

9. In addition each code-point has **properties** which are not defined in the code-chart/code-block. These properties are crucial for Internationalised Domain Names since they define whether a given character will be accepted or not in resolving an IDN typed in a browser. This means that not all characters can be used to create an address. Properties such as PValid or Protocol valid are assigned to each code point [RFC 5892]

Typing डबांग०.भारत

will result in an error since the abbreviation marker is not permissible and will not be resolved.

Unicode Features: PRINCIPLES

Unicode is governed by 3 major principles. These are

1. The Principle of Economy : Only one and one glyph afferent to a particular script will be maintained and stored as a code-point. Normally this is the most frequent one. In other words allographs are not permitted. Thus the older forms of Devanagari /a/ /jh/ are not retained. However a font developer if he so desires can replace the /jh/ by the older shape the older shape in his font and Unicode will render it since both will refer to the same code-point.

अ = आ

झ = झ

Unicode Features: Principles

2. The principle of Simplicity: This principle states that only atomic glyphs will be retained. These are characters which are not decomposable. This important in the case of complex scripts of the Neo-Brahmi family since these use the Halanta or the Vowel killer (with the possible exception of Tamil) to render composite characters known as ligatures. Unicode only permits akhanda forms and composing complex conjuncts is rendered possible by the adjunction of akhanda code-points

क्ष = क+्+ष

कात्स्न्र्य= क+ा+त+्+स+्न+य

Unicode Features

3. The principle of Comprehensiveness:
This principle simply states that a given repertoire of a script represented in Unicode shall be as comprehensive as possible, and this without compromising the other two principles.
Thus, as per the principle of Economy, allographs of a given character shall not be included.
Thus ल ऌ variants of a single shape shall not be admitted. Only one form shall be permitted as per the first principle.

Unicode Features: Normalization

An exception to the above principle arises when a given character within a language can be rendered in two different ways. Unicode defines this as removing alternate representations of equivalent sequences from textual data, to convert the data into a unique form In the Unicode Standard, normalization refers specifically to processing to ensure that strings have unique representations.

रिज़र्व

ज + ◌़ = 091C+093C

ज़ = 095B

रिज़र्व = र + ि + ज + ़ + र + ् + व

Also = र + ि + ज़ + र + ् + व

This ensures in the case of IDN's that phishing and spoofing are reduced as far as possible, since the browser itself "normalizes" the two forms.

Unicode Features: Normalization

Groundwork for IDN. Points covered:

1. Unicode
2. ISO 10646
3. BMP: Basic Multilingual Plane/SMP: Supplementary Multilingual Plane
4. Code Charts/Code Blocks
5. Code point
6. Character Set/Character Repertoire/Allographs
7. Reading a Unicode Code-Block
8. Character Properties
9. Unicode Normalization
10. Governing Principles: Simplicity; Comprehensiveness, Economy

Further reading:

Pvalid code-points: https://tools.ietf.org/html/rfc5892

•Glossary of Unicode terms: http://www.unicode.org/glossary/

•Properties of South Asian Scripts:
http://www.unicode.org/versions/Unicode7.0.0/ch12.pdf

सी डैक
CDAC
Gist

धन्यवाद !