Madan Puraskar Pustakalaya
01-04-2008

Character set and encoding constraints for Nepali IDNs

Bal Krishna Bal, Basanta Shrestha, Dibyendra Hyoju Laxmi Khatiwada

Madan Puraskar Pustakalaya, Lalitpur, Nepal

bal@mpp.org.np ,basanta@mpp.org.np, dibyendra@mpp.org.np

Abstract

This paper gives an overview of the character set proposed by the Nepal Country Component for Internationalized Domain Names (IDN) in Nepali and the existing encoding constraints in this regard.

1. Introduction

Localization of software has taken a big hype all over the world in the recent years. This movement has been taken as a big revolution in making possible the effective transfer of information and technology by means of the usage of local language. In this context, while this revolution has broken down the dominance of the English language in the field of information technology and at the same time provided opportunity to the non-English speakers around the world to go hand in hand with the latest technology. In the past, without the knowledge of English, this was not possible. The introduction of IDNs is yet another milestone in the direction of localization which means that after the successful implementation of this technology, one can type the names of the websites in the address bar of browsers in his/her own local language and get access to the desired pages. In the context of Nepal, the availability of web content in Nepali is quite low at present and from this perspective it might seem that the issue of IDN translation is not as much significant but from a technological advancement point of view, this initiative certainly carries far reaching consequences.

2. Character set for Nepali

The character set that Nepali uses is the subset from the Devanagari character set which has been allocated space from 0900 to 097F in the Unicode table. Below in Table 1, we try to list down this subset of the character set applicable for Nepali and their respective Unicode values.

Table 1. Character set for Nepali

Character	Unicode value
ँ (Devanagari sign Candrabindu)	0901
ं (Devanagari sign anusvara or bindu)	0902
ः (Devanagari sign visarga)	0903
अ (Devanagari letter A)	0905
आ (Devanagari AA)	0906
इ (Devanagari I)	0907
ई (Devanagari II)	0908

उ(Devanagari U)	0908
ऊ (Devanagari UU)	090A
ऋ (Devanagari letter vocalic R)	090B
ए (Devanagari letter E)	090F
ऐ(Devanagari letter AI)	0910
ओ (Devanagari letter O)	0913
औ (Devanagari letter AU)	0914
क (Devanagari letter KA)	0915
ख (Devanagari letter KHA)	0916
ग (Devanagari letter GA)	0917
घ (Devanagari letter GHA)	0918
ङ (Devanagari letter NGA)	0919
च (Devanagari letter CA)	091A
छ(Devanagari letter CHA)	091B
ज (Devanagari letter JA)	091C
झ (Devanagari letter JHA)	091D
ञ (Devanagari letter NYA)	091E
ट (Devanagari letter TTA)	091F
ठ (Devanagari letter TTHA)	0920
ड (Devanagari Letter DDA)	0921
ढ (Devanagari Letter DDHA)	092

ण (Devanagari letter NNA)	0923
त (Devanagari letter TA)	0924
थ (Devanagari letter THA)	0925
द (Devanagari letter DA)	0926
ध (Devanagari letter DHA)	0927
न (Devanagari letter NA)	0928
प (Devanagari letter PA)	092A
फ (Devanagari letter PHA)	092B
ब (Devanagari letter BA)	092C
भ (Devanagari letter BHA)	092D
म (Devanagari letter MA)	092E
य (Devanagari letter YA)	092F
र (Devanagari letter RA)	0930
ल (Devanagari letter LA)	0932
व (Devanagari letter VA)	0935
श (Devanagari letter SHA)	0936
ष (Devanagari letter SSA)	0937
स (Devanagari letter SA)	0938
ह (Devanagari letter HA)	0939
ा (Devanagari vowel sign AA)	093E
ि (Devanagari vowel sign I)	093F
ी (Devanagari vowel	0940

sign II)	
ु (Devanagari vowel sign U)	0941
ू (Devanagari vowel sign UU)	0942
ृ (Devanagari vowel sign vocalic R)	0943
े (Devanagari vowel sign E)	0947
ै (Devanagari vowel sign AI)	0948
ो (Devanagari vowel sign O)	094B
ौ (Devanagari vowel sign AU)	094C
् (Devanagari vowel sign Virama)	094D
ॐ (Devanagari OM)	0950
। (Devanagari Danda)	0964
॥ (Devanagari Double Danda)	0965
० (Devanagari Digit ZERO)	0966
१ (Devanagari Digit ONE)	0967
२ (Devanagari Digit TWO)	0968
३(Devanagari Digit THREE)	0969
४(Devanagari Digit FOUR)	096A
५ (Devanagara Digit FIVE)	096B
६(Devanagari Digit SIX)	096C
७(Devanagari Digit SEVEN)	096D
८(Devanagari Digit EIGHT)	096E
९(Devanagari Digit NINE)	096F
॰ (Devanagari Abbreviation sign)	0970

In addition to the above characters, Nepali also requires two more character codes, namely ZERO WIDTH JOINER(0200D) and ZERO WIDTH NON-JOINER(0200C). These are invisible characters that play an important role in producing certain unique text in Nepali. We try to illustrate the usage of these two symbols in Nepali below:

Case I: Using ZWNJ – श्रीमान्को
(श+्+र+म+ा+न+्+ZWNJ+ क+ो)

Not using ZWNJ above would produce श्रीमान्को, which is incorrect.

Case II: Using ZWJ – च्याल
(र+ZWJ+्+ य+ा+ल)
Not using ZWJ above would produce र्याल, which is incorrect.

Madan Puraskar Pustakalaya
01-04-2008

3. Discussion

Several encoding constraints for Nepali IDNs exist today. In this section, we try to discuss briefly on these limitations.

Lacking alternatives

Although it would be revolutionary to have a complete usage of one's own character set for domain names and correspondingly entering the Uniform Resource Locator (URL) addresses, in many cases it would be very irrelevant and uncomfortable. In case of Nepali, for instance, the label separator, which is "।" used for marking the end of sentences is not relevant for replacing the period sign used for label separation for latin languages. Another example is the alternative for world wide web(www). Till date, a suitable replacement has not come up for this as well. Hence, currently the period symbol "." and "www" are lacking alternatives in the Nepali language.

Lacking use

Not all the characters listed above in the character set table for Nepali are applicable while formulating domain names and URL addresses. The Devanagari abbreviation sign "॰" for instance, is not applicable for the above mentioned purpose. In fact, abbreviations are rarely used or not used at all for naming domains and entering URL addresses for latin languages as well.

Mandatory use

We have mentioned above in the "Character set for Nepali" section about the two invisible characters that are not in the Nepali character set but are integral part of Nepali Unicode typing. Need to note that these two symbols also should be taken into consideration while devising technology for entering URL addresses and domain names in Nepali.

4. Conclusion

This document in addition to listing the valid character set for the language also has discussed over the existing constraints and the possible solutions to the problems. The work is still in the research phase and hence many things might be subjected to some changes in future.

Acknowledgement

"This work was carried out with the aid of a grant from the International Development Research Centre, Ottawa, Canada administered through the centre for Research in Urdu Language Processing (CRLUP),National University of Computing and Emerging Sciences, Lahore, Pakistan (NUCES)".

5. References

[1] "GUIDELINES FOR THE IMPLEMENTATION OF INTERNATIONALIZED DOMAIN NAMES", Version 2.2 draft 0.03, http://www.icann.org/topics/idn/idn-guidelines-26apr07.pdf, 2007.

[2] http://unicode.org/charts/PDF/U0900.pdf