Report on the study of IDN with respect to south Indian Languages

IDNA (Internalizing Domain Names in Applications) allows the non-ASCII characters to be represented using ASCII characters. This is done by converting the internationalized domain names to punycode (which is a unique and reversible transformation of Unicode string to ASCII). During the conversion of internationalized domain names to punycodes it passes through a process called **NAMEPREP**. **NAMEPREP** specifies a framework of processing rules such as **Mapping(***For each character in the input, check if it has a mapping and, if so, replace it with its mapping*), **Normalization(***Possibly normalize the result of step 1 using Unicode Normalization*), **Prohibited Output(***Check for any characters that are not allowed in the output if any are found, return an error*), **Bidirectional characters(***Possibly check for right-to-left characters, and if any are found, make sure that the whole string satisfies the requirements for bidirectional strings*) for Unicode text.

A profile of **STRINGPREP** MUST include : - *The intended applicability of the profile - The character repertoire that is the input and output to stringprep - The mapping tables used - Any additional mapping tables specific to the profile - The Unicode normalization used, if any - The tables of characters that are prohibited as output - The bidirectional string testing used, if any - Any additional characters that are prohibited as output specific to the profile*

This profile of **STRINGPREP** can exclude characters that should not normally appear in text that is used. It can prevent such characters by changing the characters to be excluded to other characters, by removing those characters, or by causing an error if the characters would appear in the output. A profile of **STRINGPREP** converts a single string of input characters to a string of output characters, or returns an error if the output string would contain a prohibited character. Before the text can be emitted, it MUST be checked for prohibited codepoints.

200C; ZERO WIDTH NON-JOINER
200D; ZERO WIDTH JOINER comes under prohibited codepoints (ref: RFC 3454-STRINGPREP) (It is referred in RFC-3491 (Nameprep) also, but more details are given in RFC 3454-STRINGPREP)

Other Control characters in prohibited output are:

0000-001F; [CONTROL CHARACTERS]
007F; DELETE
0080-009F; [CONTROL CHARACTERS]
06DD; ARABIC END OF AYAH
070F; SYRIAC ABBREVIATION MARK
180E; MONGOLIAN VOWEL SEPARATOR
2028; LINE SEPARATOR
2029; PARAGRAPH SEPARATOR
2060; WORD JOINER
2061; FUNCTION APPLICATION

2062; INVISIBLE TIMES
2063; INVISIBLE SEPARATOR
206A-206F; [CONTROL CHARACTERS]
FEFF; ZERO WIDTH NO-BREAK SPACE
FFF9-FFFC; [CONTROL CHARACTERS]
1D173-1D17A; [MUSICAL CONTROL CHARACTERS]

In Tamil language script ZWJ and ZWNJ are not used. They are used in Telugu, only when writing foreign text which contains syllable breaks (virama)for e.g. if we write the phrase "wrong number" as "raang^nembar" (the caret represents a virama followed by a ZWNJ) . (If ZWJ and ZWNJ are avoided in Malayalam then if there is a URL name അവന് or അവൻ then both will point to the same URL as both ന് and ൻ have got the same punycode since ZWJ and ZWNJ is avoided)

As the part of experiment we checked using the website named **www.nameisp.com/punyasp** to convert a domain name to its corresponding punycode and the inference is given below.

1 ല് without inserting Unicode control character produces xn--4wc8c

3 ല് with inserting ZWJ produces xn--4wc8c

5 ല് with inserting ZWNJ produces xn--4wc8c

References:

RFC-3490 (IDNA)
RFC-3491 (Nameprep)
RFC-3492 (Punycode)
RFC-3454 (Stringprep)

www.w3.org/International/articles

www.faqs.org/rfcs/rfc3490.html

www.rfc-editor.org/rfc/rfc3491.txt

www.rfc-editor.org/rfc/rfc3492.txt

www.ietf.org/rfc/rfc3454.txt

Source: Govt of India
May 2006