L2/00-417

Unicode Representation of Indian Scripts

Unicode is a data storage codification for various languages. All file storage will be in unicode. If the Indian community has any ambition of maintaining databases in Unicode, it is absolutely essential for us to ensure that the codification follows a sort order, which is defined as per Indian language sorting rules. In absence of this we will pay a very heavy price in processing overheads in Indian languages. Our observations for various scripts are as follows:

Devanagri:

(1) Consonants with nukta are isolated and are not in sort order.
(2) There are quite a few customers who need 'Ksha', 'Shra' and 'Gnya' as independent consonants and not as conjuncts.
(3) 'Dirgha Ri' and 'Dirgha Li' are completely out of sort order.
(4) If Indian Language numerals are located almost at the end of the code it is not clear how to use them and they are clearly out of sort order. This comment will apply to all Indian Scripts.
(5) Some of the vedic 'swara chinha' are not proper and are not adequate.

Gujarati:

(1) 'Dirgha Ri' are completely out of sort order.
(2) 'Abbreviation' sign is missing.
(3) There are quite a few customers who need 'Ksha', 'Shra' and 'Gnya' as independent consonants and not as conjuncts.

Punjabi:

(1) Consonants with nukta are not in proper sort order.

Bengali:

(1) Comments 1, 2, and 3 mentioned in Devanagri applies.

Assamese:

(1) Assamese may be treated as a separate language since there are two consonants that are different. The current locations in Bengali script will not satisfy the sort order.

Oriya:

(1) Comments 1, 2, and 3 mentioned in Devanagri applies.

Tamil:

(1) The 6 Grantha characters may be located at the end of all consonants.
(2) There is no anuswar in Tamil.
(3) In Tamil sort order, the half characters(consonant + pulli) comes before the full consonants.
(4) Tamil character set is not complete. A reasonably complete set may be obtained from Tam99 standard.

(5) ‘Om’ is missing.
(6) Tamil experts may decide whether to have old Tamil characters.

Kannada:

(1) Nukta equivalent character is missing.
(2) ‘Dirgha Ri’ and ‘Dirgha Li’ are completely out of sort order.
(3) Vedic ‘swara chinaha’ are missing.
(4) Symbols like ‘Om’ and ‘Avagraha’ are missing.

Telugu:

(1) ‘Dirgha Ri’ and ‘Dirgha Li’ are completely out of sort order.
(2) Vedic ‘swara chinaha’ are missing.
(3) Few Vedic characters are missing.

Malayalam:

(1) ‘Dirgha Ri’ and ‘Dirgha Li’ are completely out of sort order.

Presented by Dr. M. N. Cooper
Modular InfoTech
PUNE.
Ph 020-4227994
Email modular@giaspn01.vsnl.net.in