**ISO/IEC JTC 1/SC 2/WG 2**
**PROPOSAL SUMMARY FORM TO ACCOMPANY SUBMISSIONS**
**FOR ADDITIONS TO THE REPERTOIRE OF ISO/IEC 10646**[1]
**Please fill all the sections A, B and C below.**
**Please read Principles and Procedures Document (P & P) from** http://www.dkuug.dk/JTC1/SC2/WG2/docs/principles.html **for guidelines and details before filling this form.**
**Please ensure you are using the latest Form from** http://www.dkuug.dk/JTC1/SC2/WG2/docs/summaryform.html**.**
**See also** http://www.dkuug.dk/JTC1/SC2/WG2/docs/roadmaps.html **for latest *Roadmaps*.**

## A. Administrative

1. **Title:** Proposal to Encode Devanagari Letter Candra A in the UCS
2. Requester's name: INCITS/L2 (US); Unicode Consortium
3. Requester type (Member body/Liaison/Individual contribution): member body, liaison contribution
4. Submission date: 2007-04-20
5. Requester's reference (if applicable): L2/07-027R

6. Choose one of the following:
This is a complete proposal: Yes
or, More information will be provided later: No

## B. Technical – General

1. Choose one of the following:
   a. This proposal is for a new script (set of characters): No
      Proposed name of script: N/A
   b. The proposal is for addition of character(s) to an existing block: Yes
      Name of the existing block: Devanagari
2. Number of characters in proposal: 1
3. Proposed category (select one from below - see section 2.2 of P&P document):
   A-Contemporary X B.1-Specialized (small collection) _____ B.2-Specialized (large collection) _____
   C-Major extinct _____ D-Attested extinct _____ E-Minor extinct _____
   F-Archaic Hieroglyphic or Ideographic _____ G-Obscure or questionable usage symbols _____
4. Proposed Level of Implementation (1, 2 or 3) (see Annex K in P&P document): 1
   Is a rationale provided for the choice? yes
      If Yes, reference: not a combining character
5. Is a repertoire including character names provided? Yes
   a. If YES, are the names in accordance with the "character naming guidelines" in Annex L of P&P document? Yes
   b. Are the character shapes attached in a legible form suitable for review? Yes
6. Who will provide the appropriate computerized font (ordered preference: True Type, or PostScript format) for publishing the standard? Everson Typography (TrueType)
   If available now, identify source(s) for the font (include address, e-mail, ftp-site, etc.) and indicate the tools used:
7. References:
   a. Are references (to other character sets, dictionaries, descriptive texts etc.) provided? Yes
   b. Are published examples of use (such as samples from newspapers, magazines, or other sources) of proposed characters attached? Yes
8. Special encoding issues:
   Does the proposal address other aspects of character data processing (if applicable) such as input, presentation, sorting, searching, indexing, transliteration etc. (if yes please enclose information)?
   Yes: suggested character properties included
9. Additional Information:

Submitters are invited to provide any additional information about Properties of the proposed Character(s) or Script that will assist in correct understanding of and correct linguistic processing of the proposed character(s) or script. Examples of such properties are: Casing information, Numeric information, Currency information, Display behaviour information such as line breaks, widths etc., Combining behaviour, Spacing behaviour, Directional behaviour, Default Collation behaviour, relevance in Mark Up contexts, Compatibility equivalence and other Unicode normalization related information. See the Unicode standard at http://www.unicode.org for such information on other scripts. Also see http://www.unicode.org/Public/UNIDATA/UCD.html and associated Unicode Technical Reports for information needed for consideration by the Unicode Technical Committee for inclusion in the Unicode Standard.

[1] Form number: N2652-F (Original 1994-10-14; Revised 1995-01, 1995-04, 1996-04, 1996-08, 1999-03, 2001-05, 2001-09, 2003-11)

**C. Technical - Justification**

1. Has this proposal for addition of character(s) been submitted before? No
   If YES explain
2. Has contact been made to members of the user community (for example: National Body, user groups of the script or characters, other experts, etc.)? Yes
   If YES, with whom? Government of India, users in India
   If YES, available relevant documents:
3. Information on the user community for the proposed characters (for example: size, demographics, information technology use, or publishing use) is included? Yes
   Reference: User community includes tens of millions of speakers of Marathi
4. The context of use for the proposed characters (type of use; common or rare) common use
   Reference: The character is used in general literature
5. Are the proposed characters in current use by the user community? yes
   If YES, where? Reference: User community includes tens of millions of speakers of Marathi
6. After giving due considerations to the principles in the P&P document must the proposed characters be entirely in the BMP? Yes
   If YES, is a rationale provided? Yes
   If YES, reference: addition to existing script block
7. Should the proposed characters be kept together in a contiguous range (rather than being scattered)? N/A
8. Can any of the proposed characters be considered a presentation form of an existing character or character sequence? No
   If YES, is a rationale for its inclusion provided? N/A
   If YES, reference: N/A
9. Can any of the proposed characters be encoded using a composed character sequence of either existing characters or other proposed characters? Potentially
   If YES, is a rationale for its inclusion provided? Yes
   If YES, reference: Discussed below.
10. Can any of the proposed character(s) be considered to be similar (in appearance or function) to an existing character? No
    If YES, is a rationale for its inclusion provided? N/A
    If YES, reference: N/A
11. Does the proposal include use of combining characters and/or use of composite sequences? No
    If YES, is a rationale for such use provided? N/A
    If YES, reference: N/A
    Is a list of composite sequences and their corresponding glyph images (graphic symbols) provided? N/A
    If YES, reference: N/A
12. Does the proposal contain characters with any special properties such as control function or similar semantics? No
    If YES, describe in detail (include attachment if necessary) N/A
13. Does the proposal contain any Ideographic compatibility character(s)? No
    If YES, is the equivalent corresponding unified ideographic character(s) identified? N/A
    If YES, reference: N/A

## D Proposed Characters

One character is proposed: general category and case mapping properties are as shown:

| Code position | Glyph | Name | Character properties |
|---|---|---|---|
| 0972 | ॲ | DEVANAGARI LETTER CANDRA A | Unicode character properties should be the same as those of similar characters, such as U+0905 DEVANAGARI LETTER A. |

## E Other Information

The Government of India brought to the attention of Unicode Consortium the need to represent a text element used in Marathi documents. This text element is used in writing load words from languages such as English to represent the vowel sound /æ/ (e.g. the “a” in “apple”).

Others have mentioned a letter they deemed to be missing from Unicode. For instance, Gautam Sengupta [1] mentions that Marathi uses “a CANDRA mounted on A”, and that Unicode does not represent this missing “DEVANAGARI LETTER CANDRA A”.

The following images illustrate the text element in question:

FIRST PART

MARATHI ALPHABET

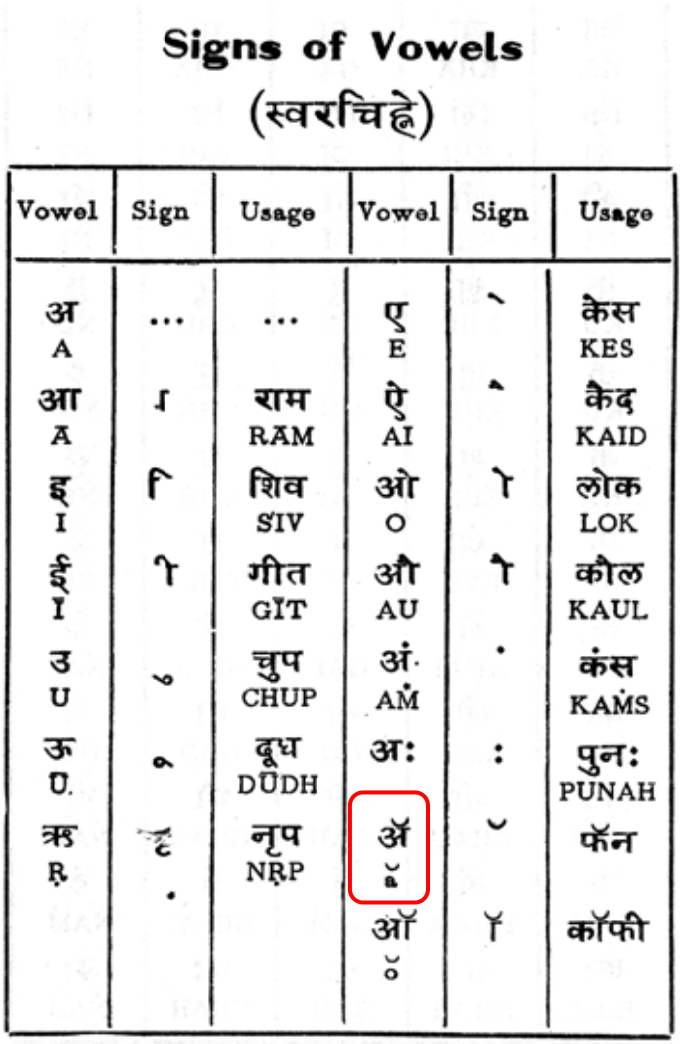
मराठी अक्षर माला

**Vowels**

(स्वर)

| | |
|---|---|
| अ | A |
| आ | Ā |
| इ | I |
| ई | Ī |
| उ | U |
| ऊ | Ū |
| ऋ | Ṛ |
| ए | E |
| ऐ | AI |
| ओ | O |
| औ | AU |
| अं | AṂ |
| अः | AH |
| ॲ | Ă — used for writing foreign sounds |
| ऑ | Ŏ — used for writing foreign sounds |

Figure 1. CANDRA A in a list of vowel letters of the Marathi alphabet ([2], p. 13)



## Signs of Vowels
(स्वरचिह्ने)

| Vowel | Sign | Usage | Vowel | Sign | Usage |
|---|---|---|---|---|---|
| अ A | ... | ... | ए E | ॅ | केस KES |
| आ Ā | ा | राम RĀM | ऐ AI | ै | केद KAID |
| इ I | ि | शिव ŚIV | ओ O | ो | लोक LOK |
| ई Ī | ी | गीत GĪT | औ AU | ौ | कौल KAUL |
| उ U | ु | चुप CHUP | अं AṀ | ं | कंस KAṀS |
| ऊ Ū | ू | दूध DŪDH | अः AḤ | ः | पुनः PUNAH |
| ऋ Ṛ | ृ | नृप NṚP | ॲ ॅ | ॅ | फॅन |
| | | | ऑ ॉ | ॉ | कॉफी |

Figure 2. CANDRA A in a table of vowel representations ([2], p. 19)

There is little question that an encoded representation for CANDRA A is needed. The question is what that encoded representation should be: a new atomic character, or a sequence of existing characters.

The practice for encoding of independent vowels that are formed from a basic vowel letter (such as the letter a) plus some diacritic sign has been to encode these as separate, atomic characters. This includes at least the following cases:

- U+0904 DEVANAGARI LETTER SHORT A
- U+0907 DEVANAGARI LETTER I
- U+090D DEVANAGARI LETTER CANDRA E
- U+090E DEVANAGARI LETTER SHORT E
- U+0910 DEVANAGARI LETTER AI
- U+0911 DEVANAGARI LETTER CANDRA O

- U+0912 DEVANAGARI LETTER SHORT O
- U+0913 DEVANAGARI LETTER O
- U+0914 DEVANAGARI LETTER AU

All of these are potentially decomposable to sequences of other characters (e.g. <0905, 0946> could be seen as a decomposition of 0904). However, none of these characters has a decomposition mapping. To deal with confusability, The Unicode Standard (TUS), version 5.0 added text stating that such sequences should not be used (see TUS 5.0, Table 9-1 and surrounding text on p. 299).

Precedent suggests that CANDRA A can be treated the same way and encoded as a new atomic character.

The possible counterargument is that there are existing implementations that follow the guidance in TUS 5.0, not supporting sequences that would be confusable with existing characters in TUS 5.0, but that *do* support other sequences that are not in the scope of that guidance since the composite text element is not encoded as an atomic character in TUS 5.0. Such implementations could support the sequence <0905, 0945>, which would have the visual appearance of CANDRA A. Moreover, implementations that support this are known to exist (this sequence will display this text element in Microsoft Windows Vista).

In spite of this counterargument, it is our opinion that the encoding principle for extended Devanagari vowel letters, which add a diacritic mark to a basic letter such as LETTER A to form an additional vowel letter, has been established since the earliest versions of the UCS: that these are encoded as atomic characters.

On that basis, then, it is proposed that DEVANAGARI LETTER CANDRA A be encoded in the UCS. Given the potential for implementations to arise that support an alternate encoded representation, it is considered urgent to have this character encoded as soon as possible. Because the Devanagari block is being revised by amendment 3, we request and recommend that this character be added in that amendment.

## F References

[1] Sengupta, Gautam. 2006. "Multilingualism on the Internet: an Indian Perspective." Presentation at the Joint UNESCO and ITU Global Symposium on Promoting the Multilingual Internet, Geneva, 9–11 May, 2006.

[2] Sanjay. 2002. *Learn Marathi in 30 days.* (National integration language series.) Chennai: Balaji Publications.