L2/08-201

Towards an Encoding for the Khojki Script in ISO/IEC 10646

Anshuman Pandey
University of Michigan
Ann Arbor, Michigan, U.S.A.
pandey@umich.edu

May 5, 2008

Contents

List of Tables

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: **Towards an Encoding for the Khojki Script in ISO/IEC 10646**
2. Requester's name: **Anshuman Pandey (pandey@umich.edu)**
3. Requester type (Member Body/Liaison/Individual contribution): **Individual contribution**
4. Submission date: **May 5, 2008**
5. Requester's reference (if applicable): **N/A**
6. Choose one of the following:
   (a) This is a complete proposal: **No**
   (b) or, More information will be provided later: **Yes**

B. Technical - General

1. Choose one of the following:
   (a) This proposal is for a new script (set of characters): **Yes**
      i. Proposed name of script: **Khojki**
   (b) The proposal is for addition of character(s) to an existing block: **No**
      i. Name of the existing block: **N/A**
2. Number of characters in proposal: **56**
3. Proposed category: **B.1 - Specialized (small collections of characters)**
4. 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**
5. Who will provide the appropriate computerized font (ordered preference: True Type, or PostScript format) for publishing the standard?: **Anshuman Pandey**; **True Type format**
   (a) If available now, identify source(s) for the font and indicate the tools used: **The characters of the digitized Khojki font are based on normalized forms of written Khojki characters. The font was drawn by Anshuman Pandey with Metafont and converted to True Type with FontForge.**
6. 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**
7. Special encoding issues:
   (a) 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; see proposal for additional details.**.
8. 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. **Character properties and numeric information are included.**

[1] Form number: N3102-F (Original 1994-10-14; Revised 1995-01, 1995-04, 1996-04, 1996-08, 1999-03, 2001-05, 2001-09, 2003-11, 2005-01, 2005-09, 2005-10, 2007-03)

C. Technical - Justification

1. Has this proposal for addition of character(s) been submitted before?: **No**
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**
   (a) If Yes, with whom?: **N/A**
      i. If Yes, available relevant documents: **N/A**
3. Information on the user community for the proposed characters (for example: size, demographics, information technology use, or publishing use) is included? **Yes**
   (a) Reference: **(Add details)**
4. The context of use for the proposed characters (type of use; common or rare): **Common**
   (a) Reference: **The script was used primarily by the Ismaili religious community in South Asia.**
5. Are the proposed characters in current use by the user community?: **Yes. Members of the Ismaili community and specialists studying Ismaili literature require the script.**
   (a) If Yes, where? Reference: **In India, Pakistan, the United Kingdom, and the United States.**
6. After giving due considerations to the principles in the P&P document must the proposed characters be entirely in the BMP?: **No**
   (a) If Yes, is a rationale provided?: **N/A**
      i. If Yes, reference: **N/A**
7. Should the proposed characters be kept together in a contiguous range (rather than being scattered)? **Yes**
8. Can any of the proposed characters be considered a presentation form of an existing character or character sequence? **No**
   (a) If Yes, is a rationale for its inclusion provided?: **N/A**
      i. 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? **No**
   (a) If Yes, is a rationale provided?: **N/A**
      i. If Yes, reference: **N/A**
10. Can any of the proposed character(s) be considered to be similar (in appearance or function) to an existing character? **Yes**
   (a) If Yes, is a rationale for its inclusion provided? **Yes**
      i. If Yes, reference: **See text of proposal**
11. Does the proposal include use of combining characters and/or use of composite sequences? **Yes**
   (a) If Yes, is a rationale for such use provided? **Yes**
      i. If Yes, reference: **See text of proposal**
   (b) Is a list of composite sequences and their corresponding glyph images (graphic symbols) provided? **N/A**
      i. If Yes, reference: **N/A**
12. Does the proposal contain characters with any special properties such as control function or similar semantics? **Yes**
   (a) If Yes, describe in detail (include attachment if necessary): **Virama**
13. Does the proposal contain any Ideographic compatibility character(s)? **No**
   (a) If Yes, is the equivalent corresponding unified ideographic character(s) identified? **N/A**
      i. If Yes, reference: **N/A**

1 Introduction

This is a working document for a project to encode the Khojki script in the Supplementary Multilingual Plane (Plane 1) of the Universal Character Set (ISO/IEC 10646).

The intent of this document is to initiate discussion regarding the character repertoire of the Khojki script, forms of the characters, and technical features of the script. Background information on the script, description of its orthography, and specimens are forthcoming.

The font used here was developed by the author. It is a basic font that strives only to be illustrative. A formal typeface will be developed to accompany revisions of this proposal.

2 Request for Allocation

Khojki is presently not allocated in the Roadmap to the Supplementary Multilingual Plane (SMP). Khojki will require a maximum of five rows.

It is requested that the Unicode Technical Committee (UTC) allocate space for Khojki in the SMP.

3 Justification for Encoding

The Khojki script is an ecclesastical script that was used by the Khoja community of South Asia to record religious literature. It is associated specifically with the Shia Nizari Ismaili religious community.

Khojki is a Brahmi-based script derived from the Landa class of scripts. It was used primarily in Sind, in modern Pakistan.

Collection of Khojki manuscripts are held by Harvard University and the Institute of Ismaili Studies (London). Khojki documents are actively studied by specialists of devotional literature and the texts contained in the documents are used by members of the Ismaili community.

An encoding for Khojki in the Universal Character Set will provide a means for the Ismaili community and specialists to preserve literature in the script and to further the study of both Ismaili literature and the Khojki script.

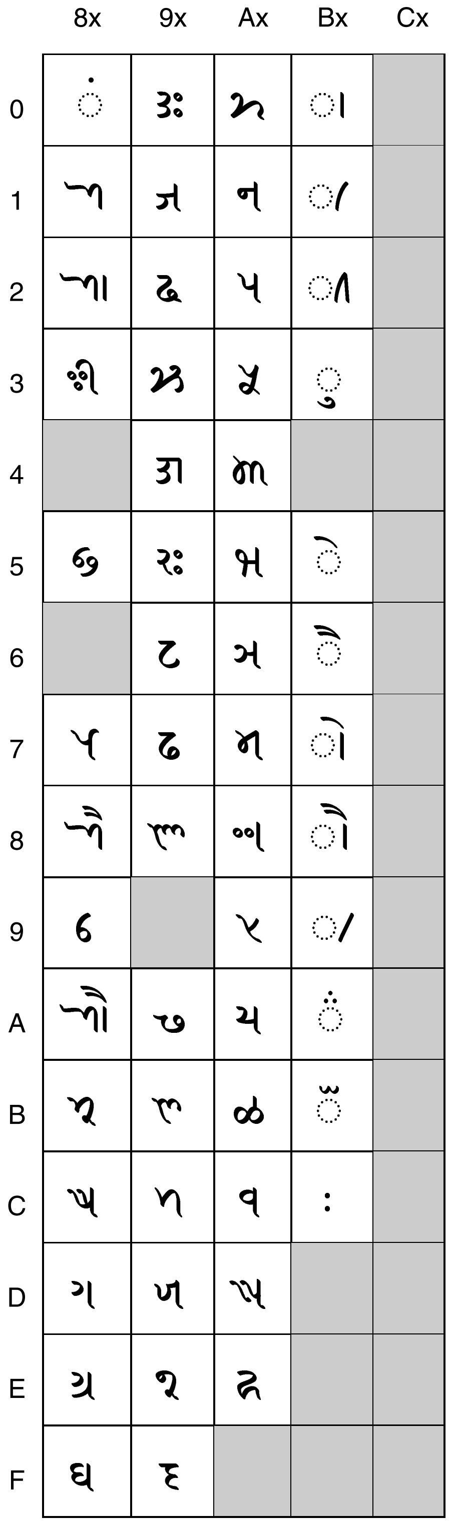
8x
9x
Ax
Bx
Cx
0
1
2
3
4
5
6
7
8
9
A
B
C
D
E
F

Table 1: Preliminary glyph chart for the Khojki script.

4 Characters Proposed

As of yet, 56 characters have been identified as constituting the core set of Khojki letters and signs.

Consonants There are 35 consonant letters:

𑈈	KHOJKI LETTER KA	𑈕	KHOJKI LETTER TTHA	𑈡	KHOJKI LETTER BA
𑈉	KHOJKI LETTER KHA	𑈖	KHOJKI LETTER DDA	𑈢	KHOJKI LETTER BBA
𑈊	KHOJKI LETTER GA	𑈗	KHOJKI LETTER DDHA	𑈣	KHOJKI LETTER BHA
𑈋	KHOJKI LETTER GGA	𑈘	KHOJKI LETTER NNA	𑈤	KHOJKI LETTER MA
𑈌	KHOJKI LETTER GHA	𑈙	KHOJKI LETTER TA	𑈥	KHOJKI LETTER YA
𑈍	KHOJKI LETTER NGA	𑈚	KHOJKI LETTER THA	𑈦	KHOJKI LETTER RA
𑈎	KHOJKI LETTER CA	𑈛	KHOJKI LETTER DA	𑈧	KHOJKI LETTER LA
𑈏	KHOJKI LETTER CHA	𑈜	KHOJKI LETTER DDDA	𑈫	KHOJKI LETTER LLA
𑈐	KHOJKI LETTER JA	𑈝	KHOJKI LETTER DHA	𑈨	KHOJKI LETTER VA
𑈒	KHOJKI LETTER JHA	𑈞	KHOJKI LETTER NA	𑈩	KHOJKI LETTER SA
𑈓	KHOJKI LETTER NYA	𑈟	KHOJKI LETTER PA	𑈪	KHOJKI LETTER HA
𑈔	KHOJKI LETTER TTA	𑈠	KHOJKI LETTER PHA		

Vowels There are 8 independent vowels:

𑈀	KHOJKI LETTER A	𑈃	KHOJKI LETTER U	𑈆	KHOJKI LETTER O
𑈁	KHOJKI LETTER AA	𑈄	KHOJKI LETTER E	𑈇	KHOJKI LETTER AU
𑈂	KHOJKI LETTER I	𑈅	KHOJKI LETTER AI		

Vowel Signs There are 8 dependent vowel signs:

◌𑈬	KHOJKI VOWEL SIGN AA	◌𑈯	KHOJKI VOWEL SIGN U	◌𑈲	KHOJKI VOWEL SIGN O
◌𑈭	KHOJKI VOWEL SIGN I	◌𑈰	KHOJKI VOWEL SIGN E	◌𑈳	KHOJKI VOWEL SIGN AU
◌𑈮	KHOJKI VOWEL SIGN II	◌𑈱	KHOJKI VOWEL SIGN AI		

Various Signs There are 4 various signs:

◌𑈴	KHOJKI SIGN ANUSVARA	◌𑈶	KHOJKI NUKTA
◌𑈵	KHOJKI SIGN VIRAMA	◌𑈷	KHOJKI SHADDA

Punctuation There is one punctuation sign:

𑈺	KHOJKI WORD SEPARATOR

5 Technical Features

5.1 Name

The name of the script in the UCS shall be Khojki.

5.2 Classification

Khojki is classified as a “Category B.1” (Specialized - small collection of characters) as per the criteria specified in ISO/IEC JTC 1/SC 2/WG 2 N3002.[1] The script consists of a small set of characters used by a small ecclesiastical community, whose literature is written in the script. Khojki is not used for ordinary communication. Although it is a minor script with a limited user base, there exists a large body of literature written and printed in Khojki.

5.3 Allocation

Khojki should be encoded in the Supplementary Multilingual Plane (SMP) (Plane 1) of the UCS. It has not yet been allocated in the SMP Roadmap. A tentative allocation of five rows is suggested for Khojki. It is likely that the script will only require four rows, but there may be punctuation and other signs that have not yet been identified. The glyph chart in Table 1 shows the characters proposed for encoding. Character properties are given in section 5.5.1

5.4 Encoding Model

The Khojki script is an abugida of the Brahmic type. It is written from left to right. The formation of syllables in Khojki follows the pattern common to north Indic scripts. The encoding model for Khojki may be based on the model implemented for Gurmukhi.

Consonant letters bear the inherent vowel *a* (KHOJKI LETTER A) when unaccompanied by a vowel sign. The inherent vowel is suppressed by the *virāma* (KHOJKI SIGN VIRAMA) to produce the bare consonant. The inherent vowel is changed by applying a vowel sign to the consonant. All dependent vowel signs are written above, below, or to the left of the consonant (including the dependent ◌𑈭 KHOJKI VOWEL SIGN I).

A sequence of consonants (in which all consonants except for the final are marked by *virāma*) is written as a consonant conjunct, which may occur as (a) a true ligature; (b) half-forms of all consonants except the final consonant, which assumes its full form; and (c) a sequence of full-form consonants marked with an explicit *virāma* except for the final consonant.

5.5 Character Properties

Vowels All independent vowels have the following properties:

General Category: Lo (Letter, Other)
Combining Class: 0 (Spacing, split, enclosing, reordrant, and Tibetan subjoined)
Bidirectional Class: L (Left-to-Right)

[1] International Organization for Standardization, 2005: 4.

Vowel Signs The dependent vowel signs are divided into two classes based upon their spacing attributes. The first class consists of the spacing marks KHOJKI VOWEL SIGN AA, KHOJKI VOWEL SIGN I, KHOJKI VOWEL SIGN II, KHOJKI VOWEL SIGN O, and KHOJKI VOWEL SIGN AU, which have the following properties:

General Category: Mc (Mark, Spacing Combining)
Combining Class: 0 (Spacing, split, enclosing, reordrant, and Tibetan subjoined)
Bidirectional Class: L (Left-to-Right)

The second class consists of the non-spacing marks KHOJKI VOWEL SIGN U, KHOJKI VOWEL SIGN UU, KHOJKI VOWEL SIGN E, and KHOJKI VOWEL SIGN AI, which have the following properties:

General Category: Mn (Mark, Nonspacing)
Combining Class: 0 (Spacing, split, enclosing, reordrant, and Tibetan subjoined)
Bidirectional Class: NSM (Non-Spacing Mark)

Consonants All consonants have the following properties:

General Category: Lo (Letter, Other)
Combining Class: 0 (Spacing, split, enclosing, reordrant, and Tibetan subjoined)
Bidirectional Class: L (Left-to-Right)

Various Signs The KHOJKI SIGN ANUSVARA is a non-spacing mark that belongs to the general category "Mn," is of combining class "0," and possesses the bidirectional class value "NSM."

The KHOJKI SIGN VIRAMA is a spacing mark that belongs to the general category "Mc," has a combining class value of "9" (Viramas), and has the bidirectional class value "L".

The KHOJKI SIGN NUKTA is a non-spacing spacing mark that belongs to the general category "Mn," is of combining class "7" (Nuktas), and possesses the bidirectional class value "NSM".

The KHOJKI SIGN SHADDA is a non-spacing spacing mark that belongs to the general category "Mn," is of combining class "0," and possesses the bidirectional class value "NSM".

5.5.1 Unicode Character Database Format

The properties for Khojki characters in the Unicode Character Database format are:

```
xxx80;KHOJKI SIGN ANUSVARA;Mn;0;NSM;;;;;N;;;;;
xxx81;KHOJKI LETTER A;Lo;0;L;;;;;N;;;;;
xxx82;KHOJKI LETTER AA;Lo;0;L;;;;;N;;;;;
xxx83;KHOJKI LETTER I;Lo;0;L;;;;;N;;;;;
xxx84;<reserved>
xxx85;KHOJKI LETTER U;Lo;0;L;;;;;N;;;;;
xxx86;<reserved>
xxx87;KHOJKI LETTER E;Lo;0;L;;;;;N;;;;;
xxx88;KHOJKI LETTER AI;Lo;0;L;;;;;N;;;;;
xxx89;KHOJKI LETTER O;Lo;0;L;;;;;N;;;;;
xxx8A;KHOJKI LETTER AU;Lo;0;L;;;;;N;;;;;
xxx8B;KHOJKI LETTER KA;Lo;0;L;;;;;N;;;;;
xxx8C;KHOJKI LETTER KHA;Lo;0;L;;;;;N;;;;;
xxx8D;KHOJKI LETTER GA;Lo;0;L;;;;;N;;;;;
xxx8E;KHOJKI LETTER GGA;Lo;0;L;;;;;N;;;;;
xxx8F;KHOJKI LETTER GHA;Lo;0;L;;;;;N;;;;;
```

```
xxx90;KHOJKI LETTER NGA;Lo;0;L;;;;;N;;;;;
xxx91;KHOJKI LETTER CA;Lo;0;L;;;;;N;;;;;
xxx92;KHOJKI LETTER CHA;Lo;0;L;;;;;N;;;;;
xxx93;KHOJKI LETTER JA;Lo;0;L;;;;;N;;;;;
xxx94;KHOJKI LETTER JHA;Lo;0;L;;;;;N;;;;;
xxx95;KHOJKI LETTER NYA;Lo;0;L;;;;;N;;;;;
xxx96;KHOJKI LETTER TTA;Lo;0;L;;;;;N;;;;;
xxx97;KHOJKI LETTER TTHA;Lo;0;L;;;;;N;;;;;
xxx98;KHOJKI LETTER DDA;Lo;0;L;;;;;N;;;;;
xxx9A;<reserved>
xxx9B;KHOJKI LETTER DDHA;Lo;0;L;;;;;N;;;;;
xxx9C;KHOJKI LETTER NNA;Lo;0;L;;;;;N;;;;;
xxx9D;KHOJKI LETTER TA;Lo;0;L;;;;;N;;;;;
xxx9E;KHOJKI LETTER THA;Lo;0;L;;;;;N;;;;;
xxx9F;KHOJKI LETTER DA;Lo;0;L;;;;;N;;;;;
xxx99;KHOJKI LETTER DDDA;Lo;0;L;;;;;N;;;;;
xxxA0;KHOJKI LETTER DHA;Lo;0;L;;;;;N;;;;;
xxxA1;KHOJKI LETTER NA;Lo;0;L;;;;;N;;;;;
xxxA2;KHOJKI LETTER PA;Lo;0;L;;;;;N;;;;;
xxxA3;KHOJKI LETTER PHA;Lo;0;L;;;;;N;;;;;
xxxA4;KHOJKI LETTER BA;Lo;0;L;;;;;N;;;;;
xxxA5;KHOJKI LETTER BBA;Lo;0;L;;;;;N;;;;;
xxxA6;KHOJKI LETTER BHA;Lo;0;L;;;;;N;;;;;
xxxA7;KHOJKI LETTER MA;Lo;0;L;;;;;N;;;;;
xxxA8;KHOJKI LETTER YA;Lo;0;L;;;;;N;;;;;
xxxA9;KHOJKI LETTER RA;Lo;0;L;;;;;N;;;;;
xxxAA;KHOJKI LETTER LA;Lo;0;L;;;;;N;;;;;
xxxAB;KHOJKI LETTER LLA;Lo;0;L;;;;;N;;;;;
xxxAC;KHOJKI LETTER VA;Lo;0;L;;;;;N;;;;;
xxxAD;<reserved>
xxxAE;KHOJKI LETTER SA;Lo;0;L;;;;;N;;;;;
xxxAF;KHOJKI LETTER HA;Lo;0;L;;;;;N;;;;;
xxxB0;KHOJKI VOWEL SIGN AA;Mn;0;NSM;;;;;N;;;;;
xxxB1;KHOJKI VOWEL SIGN I;Mc;0;L;;;;;N;;;;;
xxxB2;KHOJKI VOWEL SIGN II;Mc;0;L;;;;;N;;;;;
xxxB3;KHOJKI VOWEL SIGN U;Mn;0;NSM;;;;;N;;;;;
xxxB4;<reserved>
xxxB5;KHOJKI VOWEL SIGN E;Mn;0;NSM;;;;;N;;;;;
xxxB6;KHOJKI VOWEL SIGN AI;Mn;0;NSM;;;;;N;;;;;
xxxB7;KHOJKI VOWEL SIGN O;Mn;0;NSM;;;;;N;;;;;
xxxB8;KHOJKI VOWEL SIGN AU;Mn;0;NSM;;;;;N;;;;;
xxxB9;KHOJKI SIGN VIRAMA;Mc;9;NSM;;;;;N;;;;;
xxxBA;KHOJKI SIGN NUKTA;Mn;7;NSM;;;;;N;;;;;
xxxBB;KHOJKI SIGN SHADDA;Mn;0;NSM;;;;;N;;;;;
xxxBC;KHOJKI WORD SEPARATOR;Po;0;L;;;;;N;;;;;
xxxBD;<reserved>
xxxBE;<reserved>
xxxBF;<reserved>
```

6 Orthography

6.1 General Features

6.2 Vowels

Khojki does not have an independent character for ii; the i is used instead. It does, however, have the dependent vowel sign for ii.

The script lacks both the independent letter and dependent vowel sign for UU. The independent and dependent forms of U are used for both short and long forms.

KHOJKI VOWEL SIGN I Unlike the practice in other Indic scripts, the ◌𑈭 KHOJKI VOWEL SIGN I is written to the right of the consonant, not to the left.

6.3 Other Signs

KHOJKI VOWEL VIRAMA Unlike the practice is other Indic scripts, the ◌𑈵 KHOJKI SIGN VIRAMA is written to the right of the consonant, not beneath.

6.4 Gemination

Geminated consonants are marked by ◌𑈷 KHOJKI SHADDA. The sign is written above the consonant to be doubled. This practice is similar to that of Gurmukhi, where the GURMUKHI SIGN ADDAK is used to denote doubled consonants, and also to the Arabic script, where the ARABIC SHADDA is used.

6.5 Nukta

Nukta is marked using ◌𑈶 KHOJKI NUKTA. Both vowel and consonant letters can take nukta, which is written above the letter. It is primarily used to transcribe Arabic letters in Khojki.

6.6 Word Boundaries

Khojki separates words using 𑈺 KHOJKI WORD SEPARATOR, which resembles a colon and also VISARGA.

6.7 Sentence Boundaries

Danda is used.