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 for guidelines and details before filling this form.)
See http://www.dkuug.dk/JTC1/SC2/WG2/docs/summaryform.html for latest *Form*.
See http://www.dkuug.dk/JTC1/SC2/WG2/docs/principles.html for latest *Principles and Procedures* document.
See http://www.dkuug.dk/JTC1/SC2/WG2/docs/roadmaps.html for latest roadmaps.

A. Administrative

1. **Title:** **Proposal to Encode Kharoṣṭhī in Plane 1 of ISO/IEC 10646**

2. Requester's name: Andrew Glass, Stefan Baums, Richard Salomon

3. Requester type (Member body/Liaison/Individual contribution): Individual contribution

4. Submission date: 19 September 2002

5. Requester's reference (if applicable): _____

6. (Choose one of the following:)

This is a complete proposal: Yes

or, More information will be provided later: _____

B. Technical - General

1. (Choose one of the following:)

   a. This proposal is for a new script (set of characters): Yes

   Proposed name of script: Kharoṣṭhī / KHAROSTHI

   b. The proposal is for addition of character(s) to an existing block: _____

   Name of the existing block: _____

2. Number of characters in proposal: 66

3. Proposed category (see section II, Character Categories): C

4. Proposed Level of Implementation (1, 2 or 3) (see clause 14, ISO/IEC 10646-1: 2000): Level 3

   Is a rationale provided for the choice? Yes

   If Yes, reference: Combining marks used.

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 ISO/IEC 10646-1: 2000? 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? Andrew Glass (True Type)

   If available now, identify source(s) for the font (include address, e-mail, ftp-site, etc.) and indicate the tools
   used: Not yet available.

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. It covers Kharoṣṭhī bidirectional behavior and gives normative rules required for rendering the script.

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/UnicodeCharacterDatabase.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: N2352-F (Original 1994-10-14; Revised 1995-01, 1995-04, 1996-04, 1996-08, 1999-03, 2001-05, 2001-09)

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? Richard Salomon, Andrew Glass	
If YES, available relevant documents: Kharoṣṭhī Manuscript Paleography	
3. Information on the user community for the proposed characters (for example: size, demographics, information technology use, or publishing use) is included?	Scholars
Reference: _____	
4. The context of use for the proposed characters (type of use; common or rare)	Scholarly; Rare
Reference: _____	
5. Are the proposed characters in current use by the user community?	Yes
If YES, where? Reference: Scholars worldwide	
6. After giving due considerations to the principles in *Principles and Procedures document* (a WG 2 standing document) must the proposed characters be entirely in the BMP?	No
If YES, is a rationale provided?	_____
If YES, reference: _____	
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
If YES, is a rationale for its inclusion provided?	_____
If YES, reference: _____	
9. Can any of the proposed characters be encoded using a composed character sequence of either existing characters or other proposed characters?	No
If YES, is a rationale for its inclusion provided?	_____
If YES, reference:	_____
10. Can any of the proposed character(s) be considered to be similar (in appearance or function) to an existing character?	Yes
If YES, is a rationale for its inclusion provided?	Yes
If YES, reference:	See below
11. Does the proposal include use of combining characters and/or use of composite sequences (see clauses 4.12 and 4.14 in ISO/IEC 10646-1: 2000)?	Yes
If YES, is a rationale for such use provided?	Yes
If YES, reference: See below; and Kharoṣṭhī Manuscript Paleography	
Is a list of composite sequences and their corresponding glyph images (graphic symbols) provided?	Yes
If YES, reference: See below; and Kharoṣṭhī Manuscript Paleography	
12. Does the proposal contain characters with any special properties such as control function or similar semantics?	Yes
If YES, describe in detail (include attachment if necessary)	Virāma (10A3F)
13. Does the proposal contain any Ideographic compatibility character(s)?	No
If YES, is the equivalent corresponding unified ideographic character(s) identified? _____	
If YES, reference: _____	

Submitter's Responsibilities

The national body or liaison organization (or any other organization or an individual) proposing new character(s) or a new script shall provide:

1. Proposed category for the script or character(s), character name(s), and description of usage.
2. Justification for the category and name(s).
3. A representative glyph(s) image on paper:
   If the proposed glyph image is similar to a glyph image of a previously encoded ISO/IEC 10646 character, then additional justification for encoding the new character shall be provided.
   Note: Any proposal that suggests that one or more of such variant forms is actually a <u>distinct</u> character requiring separate encoding, should provide detailed, printed evidence that there is actual, contrastive use of the variant form(s). It is insufficient for a proposal to claim a requirement to encode <u>as characters</u> in the Standard, glyphic forms which happen to occur in another character encoding that did not follow the Character-Glyph Model that guides the choice of appropriate characters for encoding in ISO/IEC 10646.
   Note: WG 2 has resolved in Resolution M38.12 not to add any more Arabic presentation forms to the standard and suggests users to employ appropriate input methods, rendering and font technologies to meet the user requirements.
4. Mappings to accepted sources, for example, other standards, dictionaries, accessible published materials
5. Computerized/camera-ready font:
   Prior to the preparation of the final text of the next amendment or version of the standard a suitable computerized font (camera-ready font) will be needed. Camera-ready copy is mandatory for final text of any pDAMs before the next revision. Ordered preference of the fonts is True Type or PostScript format. The minimum design resolution for the font is 96 by 96 dots matrix, for presentation at or near 22 points in print size.
6. List of all the parties consulted.
7. Equivalent glyph images:
   If the submission intends using composite sequences of proposed or existing combining and non-combining characters, a list consisting of each composite sequence and its corresponding glyph image shall be provided to better understand the intended use.
8. Compatibility equivalents:
   If the submission includes compatibility ideographic characters, identify the equivalent unified CJK Ideograph character(s).
9. Any additional information that will assist in correct understanding of the different characteristics and linguistic processing of the proposed character(s) or script.

Proposal for Kharoṣṭhī script

This is a proposed assignment for Kharoṣṭhī characters. The Kharoṣṭhī script was used to write Gāndhārī and Sanskrit as well as various mixed dialects termed 'Gāndhārī Hybrid Sanskrit' (see Salomon 2001). The characters in this proposal are derived from sources in the Kharoṣṭhī script from across the whole range of known manuscripts and inscriptions. The intention is to provide a standard method for writing Kharoṣṭhī, and also a common means for the electronic storage of manuscript data. The Unicode Consortium has not previously published a proposal for Kharoṣṭhī.

Brief History of the Kharoṣṭhī script

The Kharoṣṭhī script is one of the two ancient writing systems of India in the historical period. Unlike the pan-Indian Brāhmī script, Kharoṣṭhī was confined to the northwest of India centered on the region of Gandhāra (modern northern Pakistan and eastern Afghanistan, see map). The exact details of its origin remain obscure despite the attention of several generations of scholars, but it is almost certainly related to Aramaic, stemming from the time of the Achaemenid conquest and occupation of that region in 559–336 BCE (Salomon 1998: 51–4). The Kharoṣṭhī script first appears in a fully developed form in the Aśokan inscriptions at Shāhbāzgaṛhī and Mānsehrā which have been dated to around 250 BCE (Hultzsch 1925: xxxv). The script continued to be used in Gandhāra and neighboring regions, sometimes alongside Brāhmī, until around the third century CE, when it disappeared from its homeland (Salomon 1996: 375). The Kharoṣṭhī script was also used for official documents and epigraphs in the Central Asian cities of Khotan and Niya in the third and fourth centuries CE, and appears to have survived in Kucha and neighboring areas along the Northern Silk Road as late as the seventh century.

Map: Geographical extent of the Kharoṣṭhī script.

The Kharoṣṭhī script was initially deciphered around the middle of the nineteenth century by

James Prinsep and others who worked from the short biscript inscriptions (Greek and Kharoṣṭhī) on the coins of the Indo-Greek and Indo-Scythian kings. The decipherment has been refined over the last 150 years as more material has come to light. We now have several examples of Sanskrit, or Sanskritized Gāndhārī, written in the Kharoṣṭhī script. The current proposal makes provision for encoding the level of Sanskrit found in the known documents (see Salomon 2001).

The Writing System

The Kharoṣṭhī script is a member of the Indic script family and conforms to the alphasyllabic or abugida script type. However, unlike the other scripts of this group, it is written from right to left. Kharoṣṭhī letters do not have positional variants as in Arabic and Hebrew.

Unicode Bidirectional Algorithm. Kharoṣṭhī can be implemented using the rules of the Unicode Bidirectional Algorithm as they apply to Arabic and Hebrew, with the exception that in Kharoṣṭhī both letters and numerals are written from right to left.

Convention. In this proposal we follow the Unicode naming conventions for the other Indic scripts (see http://www.unicode.org/charts/PDF/U0900.pdf), with slight adaptations based on current scholarly conventions for naming Kharoṣṭhī letters (see Glass 2000: 33–113).

Diacritic Marks/Vowels. All vowels other than *a* are written with diacritic marks in Kharoṣṭhī. In addition, there are four vowel modifiers and three consonant modifiers which are written with combining diacritics. Some letters may take more than one such diacritical mark. In these cases the preferred encoding sequence is: Letter (L) + [Consonant Modifier (CM)] + [Vowel (V)] + [Vowel Modifier (VM)]. For example the Sanskrit word *parārdhyaiḥ* might be rendered in Kharoṣṭhī script as **parārȷaiḥ* (written from right to left):
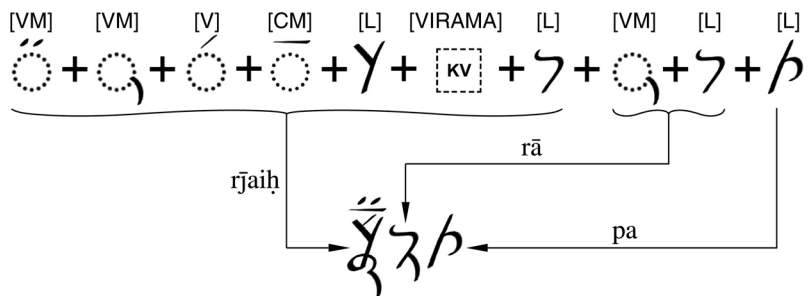
Numeral Signs. Kharoṣṭhī employs a set of numeral signs unique to the script. They have been included in this proposal. The numerals, like the letters, are written from right to left. Numbers in Kharoṣṭhī are based on an additive system. There is no zero, nor separate signs for the numbers 5–9. The number 1996, for example, would appear as: 1000 4 4 1 100 20 20 20 20 10 4 2 (see Glass 2000: 139–43).

Punctuation. Nine different punctuation marks are used in Kharoṣṭhī manuscripts and inscriptions. They have been included in this proposal (see Glass 2000: 144–7).

Minimum Rendering Requirements. Rendering requirements for Kharoṣṭhī are similar to those for Devanāgarī. The remainder of this section specifies a minimum set of rules that provide legible Kharoṣṭhī diacritic and ligature substitution behavior.

Combining Classes. The various combining diacritics attach to the full characters in different ways. A number of classes have been determined on the basis of their standard positions.

VOWEL SIGNS:

Combining *-i:*

Horizontal: example *a + -i → i*

members of this class: *a, na, ha.*

Diagonal: example *ka + -i → ki*

members of this class: *ka, ḱa, kha, ga, gha, ca, cha, ja, ña, ṭa, ṭha, ṭ́ha, ḍa, ḍha, ṇa, ta, da, dha, ba, bha, ya, ra, va, ṣa, sa, za.*

Vertical: example *tha + -i → thi*

members of this class: *tha, pa, pha, ma, la, śa.*

Combining *-u:*

Attached: example *a + -u → u*

members of this class: *a, ka, ḱa, kha, ga, gha, ca, cha, ja, ña, ṭha, ṭ́ha, ḍa, ḍha, ṇa, ta, tha, da, dha, na, pa, pha, ba, bha, ya, ra, la, va, śa, ṣa, sa, za.*

Independent: example *ha + -u → hu*

members of this class: *ṭa, ha.*

Ligatured: example *ma + -u → mu*

members of this class: *ma.*

Combining *-ṛ:*

Attached: example *a + -ṛ → ṛ*

members of this class: *a, ka, ḱa, kha, ga, gha, ca, cha, ja, ta, da, dha, na, pa, pha, ba, bha, va, śa, sa.*

Independent: example *ma + -ṛ → mṛ*

$$∪ + \text{⬡} ⇒ \text{ʊ̬}$$

members of this class: *ma, ha.*

Combining *-e:*

Horizontal: example *a + -e → e*

$$9 + \dot{○} ⇒ 9̶$$

members of this class: *a, na, ha.*

Diagonal: example *ka+ -e → ke*

$$\text{ℬ} + \dot{○} ⇒ \text{ℬ́}$$

members of this class: *ka, ḱa, kha, ga, gha, ca, cha, ja, ña,*
ṭa, ṭha, ṭ́ha, ḍa, ḍha, ṇa, ta, dha, ba, bha, ya, ra, va, ṣa, sa, za.

Vertical: example *tha + -e → the*

$$+ + \dot{○} ⇒ 4$$

members of this class: *tha, pa, pha, la, śa.*

Ligatured: example *da + -e → de*

$$\text{S} + \dot{○} ⇒ \text{C}$$

members of this class: *da, ma.*

Combining *-o:*

Diagonal: example *a + -o → o*

$$9 + \text{⬡} ⇒ 9$$

members of this class: *a, ka, ḱa, kha, ga, gha, ca, cha, ja,*
ña, ṭa, ṭha, ṭ́ha, ḍa, ḍha, ṇa, ta, tha, da, dha, na, ba, bha, ma, ra, la, va, ṣa,
sa, za, ha.

Vertical: example *pa + -o → po*

$$\text{ɦ} + \text{⬡} ⇒ \text{ɦ}$$

members of this class: *pa, pha, ya, śa.*

VOWEL MODIFIERS:

Combining VOWEL LENGTH MARK:

This sign may be used with *-a, -i, -u, -ṛ,* to indicate the equivalent long vowel *-ā,*
-ī, -ū, -ṝ. In combination with *-e* and *-o* it indicates the diphthongs *-ai* and *-au.*

Example *ma + ¯ → mā*

$$∪ + \text{◌̦} ⇒ \text{ʊ̦}$$

combines with: *-a, -i, -ṛ, -u, -e, -o.*

Combining DOUBLE RING BELOW:

This sign appears in some of the Central Asian documents. Its precise

phonetic value has not yet been established.

Example *sa* + ˳ → *sạ*

$$ꝑ + ○_∞ \Rightarrow ꝑ_∞$$

combines with: *-a, -u.*

Combining ANUSVARA:

This sign indicates nasalization of the vowel or a nasal segment following the vowel.

Example *a* + *-ṃ* → *aṃ*

$$୨ + ◌ \Rightarrow ୨$$

combines with: *-a, -i, -u, -ṛ, -e, -o.*

Combining VISARGA:

This sign is generally used to indicate unvoiced syllable-final [h]. A secondary usage is as a vowel length marker.

Example *ka* + *-ḥ* → *kaḥ*

$$ꝛ + ◌̈ \Rightarrow ꝛ̈$$

combines with: *-a, -i, -u, -ṛ, -e, -o.*

CONSONANT MODIFIERS:

Combining BAR ABOVE:

This sign is used to indicate various modified pronunciations depending on the consonants involved, such as nasalization or aspiration.

Example *ja* + ¯ → *j̄a*

$$ү + ◌̄ \Rightarrow ȳ$$

combines with: *kṣa ga, ca, ja, na, ma, śa, ṣa, sa, ha.*

Combining CAUDA:

This sign is used to indicate various modified pronunciations of the consonants involved, particularly fricativization.

Example *ga* + ´ → *ǵa*

$$ φ + ◌̨ \Rightarrow φ̨$$

combines with: *ga, ja, ḍa, ta, da, pa, ya, va, śa, sa.*

Combining DOT BELOW:

The precise value of this sign has not yet been determined.

Example *ma* + ˳ → *ṃa*

$$ υ + ○̣ \Rightarrow υ̣$$

combines with: *ma, ha.*

COMBINING VIRAMA:

This is a control character. When not followed by a consonant it causes the preceding consonant to be written as subscript to the left of the letter before it. If followed by

another consonant, it will trigger a combined form consisting of two or more consonants. The resulting form may also be subject to combinations with the above combining diacritics.

Examples:

Pure VIRAMA:

dha + i + k + [VIRAMA] → dhik

$$ 3 + \text{(i)} + \text{ঌ} + \boxed{KV} \Rightarrow \text{ß} $$

Ligatures:

ka + [VIRAMA] + ṣa → kṣa

$$ \text{ঌ} + \boxed{KV} + \mathcal{P} \Rightarrow Y $$

ma + [VIRAMA] + ra → mra

$$ \cup + \boxed{KV} + 7 \Rightarrow \mathcal{S} $$

va + [VIRAMA] + ha → vha

$$ 7 + \boxed{KV} + 2 \Rightarrow \mathcal{H} $$

sa + [VIRAMA] + ta → sta

$$ \mathcal{P} + \boxed{KV} + 5 \Rightarrow 7 $$

members of this class: *kṣ*V, *ts*V, *mr*V, *vh*V, *st*V.

Consonants with special combining forms:

sa + [VIRAMA] + ya → sya

$$ \mathcal{P} + \boxed{KV} + \wedge \Rightarrow \mathcal{P} $$

ṛa + [VIRAMA] + ta → rta

$$ 7 + \boxed{KV} + 5 \Rightarrow \mathcal{d} $$

ta + [VIRAMA] + ṛa → tra

$$ 5 + \boxed{KV} + 7 \Rightarrow \mathcal{ɾ} $$

ḷa + [VIRAMA] + pa → lpa

$$ \mathcal{7} + \boxed{KV} + \mathcal{P} \Rightarrow \mathcal{P} $$

pa + [VIRAMA] + ḷa → pla

$$ \mathcal{P} + \boxed{KV} + \mathcal{7} \Rightarrow \mathcal{P} $$

ka + [VIRAMA] + ḷa → kla

$$ \text{ঌ} + \boxed{KV} + \mathcal{7} \Rightarrow \text{ঌ} $$

ta + [VIRAMA] + *ya* → *tva*

$$\text{ち} + \boxed{\text{KV}} + \text{フ} \Rightarrow \text{ら}$$

members of this class: C*y*V, *r*CV, C*r*V, *l*CV, C*l*V, C*v*V.

Consonants with full combined forms:

ka + [VIRAMA] + *ta* → *kta*

$$\text{ス} + \boxed{\text{KV}} + \text{ち} \Rightarrow \text{ス}$$

kha + [VIRAMA] + *ka* + [VIRAMA] + *ṣa* → *khkṣa*

$$\text{ら} + \boxed{\text{KV}} + \text{ス} + \boxed{\text{KV}} + \text{プ} \Rightarrow \text{ら}$$

members of this class: *k, kh, g, ġ, c, j, ñ, ṭ, ṭh, ḍ, ḍh, ṇ, t, th, d, dh, n, p, b, bh, m, y* (in *ry*V), *l* (in *lm*V), *v* (in *vr*V), *ś, ṣ, s, z, h.*

Kharoṣṭhī

Range: 10A00 to 10A5F

These charts contain only proposed assignments and should not be considered valid until such time as the Unicode Consortium formally accepts them.
Andrew Glass created the fonts used in these charts.

Code chart

The code chart characters are normalized forms based on manuscripts of the first century CE.

	10A0	10A1	10A2	10A3	10A4	10A5
0	9 10A00	ꨐ 10A10	ⴕ 10A20	Ⴘ 10A30	⥉ 10A40	· 10A50
1	10A01	ς 10A11	ς 10A21	2 10A31	ⴑ 10A41	ο 10A51
2	10A02	φ 10A12	ϡ 10A22	ⴘ 10A32	ш 10A42	○ 10A52
3	10A03	ⴓ 10A13	ς 10A23	ⴗ 10A33	⤬ 10A43	∈ 10A53
4			ⴔ 10A24		ⴈ 10A44	⊛ 10A54
5	10A05	ⴝ 10A15	ⴔ 10A25		ⴟ 10A45	⦂ 10A55
6	10A06	ⴖ 10A16	ⴖ 10A26		ⴈ 10A46		10A56
7		ⴧ 10A17	Ⴕ 10A27		ⴁ 10A47	‖ 10A57
8			∪ 10A28	⃘ 10A38		≈ 10A58
9		ⴙ 10A19	∧ 10A29	⦿ 10A39		
A		ⴊ 10A1A	ⴚ 10A2A	⦿ 10A3A		
B		ⴁ 10A1B	⴫ 10A2B			
C	10A0C	ⴕ 10A1C	ⴤ 10A2C			
D	10A0D	T 10A1D	ⴖ 10A2D			
E	10A0E	ⴖ 10A1E	ⴖ 10A2E			
F	10A0F	ⴢ 10A1F	ⴖ 10A2F	KV 10A3F		

Name chart

The name chart characters are normalized forms based on manuscripts of the first century CE
Additional information about individual characters in this block can be found in Appendix 1.

Glyph	Unicode code point	Name	Transcription
？	10A00	KHAROSTHI LETTER A	a
	10A01	KHAROSTHI VOWEL SIGN I	i
	10A02	KHAROSTHI VOWEL SIGN U	u
	10A03	KHAROSTHI VOWEL SIGN VOCALIC R	ṛ
	10A04	(This position shall not be used)	
	10A05	KHAROSTHI VOWEL SIGN E	e
	10A06	KHAROSTHI VOWEL SIGN O	o
	10A07	(This position shall not be used)	
	10A08	(This position shall not be used)	
	10A09	(This position shall not be used)	
	10A0A	(This position shall not be used)	
	10A0B	(This position shall not be used)	
	10A0C	KHAROSTHI VOWEL LENGTH MARK	̄
	10A0D	KHAROSTHI SIGN DOUBLE RING BELOW	̰
	10A0E	KHAROSTHI SIGN ANUSVARA	ṃ
	10A0F	KHAROSTHI SIGN VISARGA	ḥ
	10A10	KHAROSTHI LETTER KA	ka
	10A11	KHAROSTHI LETTER KHA	kha
	10A12	KHAROSTHI LETTER GA	ga
	10A13	KHAROSTHI LETTER GHA	gha
	10A14	(This position shall not be used)	

Glyph	Unicode code point	Name	Transcription
ろ	10A15	KHAROSTHI LETTER CA	ca
ゐ	10A16	KHAROSTHI LETTER CHA	cha
Y	10A17	KHAROSTHI LETTER JA	ja
	10A18	(This position shall not be used)	
ᚦ	10A19	KHAROSTHI LETTER NYA	ña
Z	10A1A	KHAROSTHI LETTER TTA	ṭa
⅂	10A1B	KHAROSTHI LETTER TTHA	ṭha
Ч	10A1C	KHAROSTHI LETTER DDA	ḍa
T	10A1D	KHAROSTHI LETTER DDHA	ḍha
∫	10A1E	KHAROSTHI LETTER NNA	ṇa
ⴝ	10A1F	KHAROSTHI LETTER TA	ta
ⴲ	10A20	KHAROSTHI LETTER THA	tha
S	10A21	KHAROSTHI LETTER DA	da
ろ	10A22	KHAROSTHI LETTER DHA	dha
ⵢ	10A23	KHAROSTHI LETTER NA	na
♭	10A24	KHAROSTHI LETTER PA	pa
♭	10A25	KHAROSTHI LETTER PHA	pha
Ꭹ	10A26	KHAROSTHI LETTER BA	ba
�	10A27	KHAROSTHI LETTER BHA	bha
Ʊ	10A28	KHAROSTHI LETTER MA	ma

Glyph	Unicode code point	Name	Transcription
Λ	10A29	KHAROSTHI LETTER YA	ya
�	10A2A	KHAROSTHI LETTER RA	ra
〈	10A2B	KHAROSTHI LETTER LA	la
〉	10A2C	KHAROSTHI LETTER VA	va
Π	10A2D	KHAROSTHI LETTER SHA	śa
𝒫	10A2E	KHAROSTHI LETTER SSA	ṣa
ʔ	10A2F	KHAROSTHI LETTER SA	sa
Ч	10A30	KHAROSTHI LETTER ZA	za
2	10A31	KHAROSTHI LETTER HA	ha
℀	10A32	KHAROSTHI LETTER KKA	ḱa
᧡	10A33	KHAROSTHI LETTER TTTHA	ṭha
	10A34	(This position shall not be used)	
	10A35	(This position shall not be used)	
	10A36	(This position shall not be used)	
	10A37	(This position shall not be used)	
◌̄	10A38	KHAROSTHI SIGN BAR ABOVE	¯
◌	10A39	KHAROSTHI SIGN CAUDA	´ or ˍ
◌	10A3A	KHAROSTHI SIGN DOT BELOW	.
	10A3B	(This position shall not be used)	
	10A3C	(This position shall not be used)	
	10A3D	(This position shall not be used)	
	10A3E	(This position shall not be used)	
KV	10A3F	KHAROSTHI VIRAMA = halant · suppresses inherent vowel	see VIRAMA

Glyph	Unicode code point	Name	Transcription
)	10A40	KHAROSTHI DIGIT ONE	1
Ƿ	10A41	KHAROSTHI DIGIT TWO	2
ɯ	10A42	KHAROSTHI DIGIT THREE	3
X	10A43	KHAROSTHI DIGIT FOUR	4
?	10A44	KHAROSTHI NUMBER TEN	10
3	10A45	KHAROSTHI NUMBER TWENTY	20
𝟣	10A46	KHAROSTHI NUMBER ONE HUNDRED	100
૪	10A47	KHAROSTHI NUMBER ONE THOUSAND	1000
	10A48	(This position shall not be used)	
	10A49	(This position shall not be used)	
	10A4A	(This position shall not be used)	
	10A4B	(This position shall not be used)	
	10A4C	(This position shall not be used)	
	10A4D	(This position shall not be used)	
	10A4E	(This position shall not be used)	
	10A4F	(This position shall not be used)	
˙	10A50	KHAROSTHI PUNCTUATION DOT	
o	10A51	KHAROSTHI PUNCTUATION SMALL CIRCLE	∘
◯	10A52	KHAROSTHI PUNCTUATION CIRCLE	○
∈	10A53	KHAROSTHI PUNCTUATION CRESCENT BAR	∈
⊕	10A54	KHAROSTHI PUNCTUATION MANGALAM	⊕
❁	10A55	KHAROSTHI PUNCTUATION LOTUS	❂
∣	10A56	KHAROSTHI PUNCTUATION DANDA	∣

Glyph	Unicode code point	Name	Transcription
‖	10A57	KHAROSTHI PUNCTUATION DOUBLE DANDA	‖
≈	10A58	KHAROSTHI PUNCTUATION LINES	~
	10A59	(This position shall not be used)	
	10A5A	(This position shall not be used)	
	10A5B	(This position shall not be used)	
	10A5C	(This position shall not be used)	
	10A5D	(This position shall not be used)	
	10A5E	(This position shall not be used)	
	10A5F	(This position shall not be used)	

Text Samples

Figure 1: Aśokan inscription at Shahbazgaṛhi, ca. 250 BCE (Hultzsch 1925).
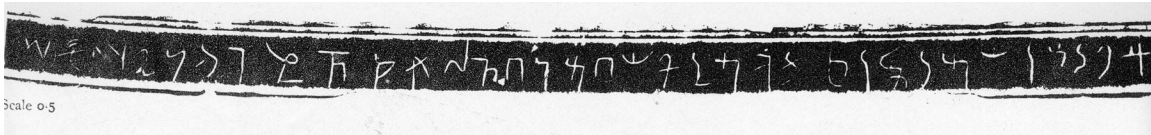
Figure 2: Relic vase inscription of Theodoros, ca. 50 BCE (Konow 1929: Plate 1).

Figure 3: Coin of King Azes with legend in Greek and Kharoṣṭhī, ca. 50 BCE. (The Royal Collection of Coins and Medals, National Museum, Denmark. Photographs by Stefan Baums and Helle Horsnæs. Inventory Number B.P. 917.)
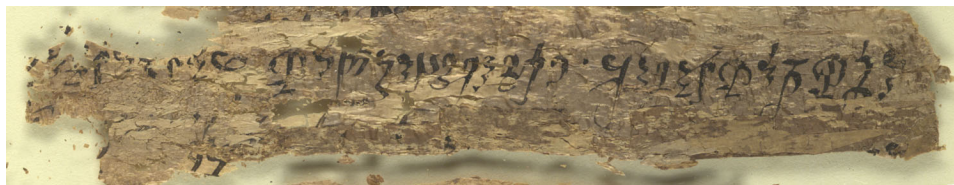
Figure 4: Detail from British Library Kharoṣṭhī Fragment 5B, ca. 50 CE (Salomon 2000: Plate 2).
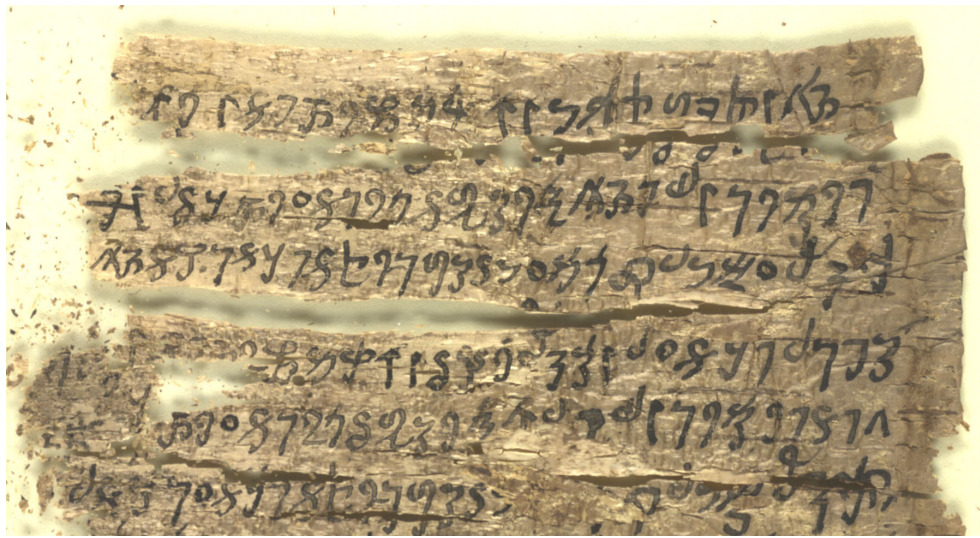
Figure 5: Detail from British Library Kharoṣṭhī Fragment 14, ca. 50 CE (Allon 2001: Plate 7).
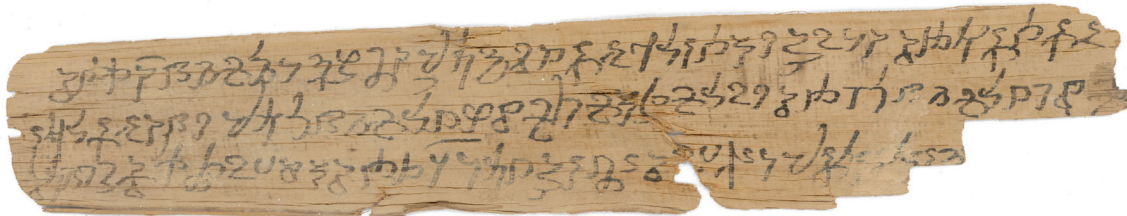
Figure 6: Fragment 44 from the Schøyen Collection, ca. 150 CE (Braarvig 2000: Plate 10.2).

budhabayaṇo (𐨦𐨂𐨣𐨿𐨩) in line 29, the scribe originally omitted the *ya* and wrote *budhabaṇa* (𐨦𐨂𐨣) before realizing his mistake and making the right leg of the *ya* from the tip of the *ṇa*, that is, *ya* 𐨩 (29.6). The *khu* 𐨐 in 38.16 was first written as a plain *kha,* and the *u*-vowel loop was added subsequently. Other examples where the scribe modified an incorrect character include *śa* 𐨠 (10.12), where the scribe corrected what he probably originally wrote as *va.* Similarly, the *ji* in 51.18 seems to have been corrected from an original *ci*. In *tva* 𐨪 (44.11), the scribe wrote a normal *ta* 𐨟 before separately adding a postconsonantal *v*. Likewise in *pra* 𐨤

Figure 7: Sample text including Kharoṣṭhī characters from a recent publication (Allon 2001: 66).

Figure 8: Typeset version of the text shown in fig. 4.

Appendix 1: Usage of Characters

- •10A00. This is the independent form of the vowel *a,* and the vowel carrier for the other independent vowels.

- 10A01 – 10A06. These are the combining vowel signs. In principle only one may be applied to each syllable. However there are some examples of akṣaras taking two vowel diacritics in Central Asian Kharoṣṭhī.

- 10A0C – 10A0D. These are vowel modifiers in the narrow sense (as opposed to 10A0E and 10A0F). They have only been found in manuscripts and inscriptions from the first century CE onwards. They are transparent for sorting purposes, see [Appendix 2](#).

- 10A0E. This is the Kharoṣṭhī *anusvāra,* indicating either vowel nasalization or a nasal consonant segment. The sort order of this glyph is thus context dependent, see [Appendix 2](#).

- 10A0F. This is the Kharoṣṭhī *visarga.* It is found only in Sanskritized forms of the language. It indicates either a variant articulation of the vowel or a [h] segment following the vowel. In the former usage, but not the latter, it is transparent for sorting purposes, see [Appendix 2](#). It cannot co-occur in the same akṣara with *anusvāra.*

- 10A10 – 10A31. These are the basic consonant signs. All unmarked consonants include the inherent vowel *a*. Other vowels are indicated by one of the combining vowel diacritics. Consequently, these consonant signs can combine with vowel diacritics and both consonant and vowel modifiers, see [Diacritic Marks/Vowels](#) above.

- 10A32 – 10A33. These are special modified forms of two of the basic consonant signs that are not obtainable by combination of those basic signs with one of the consonant modifiers. The modified forms *k̄a* and *ṭha* are consistently distinguished from *ka* and *ṭha* in the writing system.

- 10A38 – 10A3A. These are the consonant modifiers. Usually only one consonant modifier can be applied to a single consonant. The resulting combined form may also combine with the vowel diacritics and/or one of the vowel modifiers and/or *anusvāra* or *visarga,* see [Diacritic Marks/Vowels](#) above. They are transparent for sorting purposes, see [Appendix 2](#).

- 10A3F. This is the Kharoṣṭhī *virāma.* It is used to indicate the suppression of the inherent vowel. It not a mark or sign in itself, but a control character that causes the consonant which it follows to appear as a subscript to the preceding akṣara. When followed immediately by another consonant it triggers a conjunct form representing both consonants, see [Combining with VIRAMA](#) above. It can only follow a consonant, or a consonant modifier. It cannot follow a space, a vowel, a vowel modifier, a numeral sign, a punctuation sign, or another VIRAMA.

- 10A40 – 10A47. These are the Kharoṣṭhī numerals. They are written from right to left like the letters. The Kharoṣṭhī number system is additive/multiplicative, there is no zero, and no decimal point.

- 10A50 – 10A57. These are the Kharoṣṭhī punctuation signs. Nine punctuation signs have been identified from across the range of Kharoṣṭhī sources. Some of these punctuation signs could be considered to similar (in appearance or function) to existing characters. However, we feel that independent code points should be assigned to the Kharoṣṭhī punctuation signs so that Kharoṣṭhī documents posted on the Internet may be searchable for those who do not have specialized Kharoṣṭhī fonts installed. For example, such

documents should be searchable using a future version of Arial Unicode or any other single, fallback Unicode font.

Appendix 2: Sort Order

There is an ancient abecedary connected with the Kharoṣṭhī script called Arapacana, named after its first five akṣaras. There is, however, no evidence that words were ever sorted in this order. A further complication is that there is no record in Kharoṣṭhī of the complete Arapacana sequence, while Sanskrit records of it are not in total agreement about the inventory and order of letters. Therefore, we do not propose using the Arapacana as the basis for sorting.

In modern scholarly practice, Gāndhārī is sorted in much the same order as Sanskrit. Vowel length, however, even when marked is ignored in Kharoṣṭhī. In the following table, when two signs are given in a single row, they should be treated as equivalent in the sorting algorithm, the first sign having priority in tie-resolving situations, for example, *ka, ḱa, ki.*

Character	Unicode code point	Transcription
?	10A00	a
?	10A01	i
?	10A02	u
?	10A03	ṛ
?	10A05	e
?	10A06	o
?	10A0E	ṃ (preceding ∅, y–h) see note below
?	10A0F	ḥ
KV	10A3F	see VIRAMA
?, ?	10A10, 10A32	k, ḱ
?	10A11	kh
?	10A12	g
?	10A13	gh

Character	Unicode code point	Transcription
◌	10A0E	ṃ (preceding k–gh) see note below
3	10A15	c
⅀	10A16	ch
Y	10A17	j
ß , ◌	10A19, 10A0E	ñ, ṃ (preceding c–ñ) see note below
Z	10A1A	ṭ
Ⅎ , ५	10A1B, 10A33	ṭh, t́h
Y	10A1C	ḍ
T	10A1D	ḍh
⌡ , ◌	10A1E, 10A0E	ṇ, ṃ (preceding ṭ-ṇ) see note below
ꓩ	10A1F	t
⨍	10A20	th
S	10A21	d
ʒ	10A22	dh
ꟍ , ◌	10A23, 10A0E	n, ṃ (preceding t-n) see note below
⅕	10A24	p
⅔	10A25	ph
Ч	10A26	b
Ⅎ	10A27	bh

Character	Unicode code point	Transcription
ᴜ , ᴄ	10A28, 10A0E	m, ṃ (preceding p-m) see note below
∧	10A29	y
ᒐ	10A2A	r
ᒐ	10A2B	l
ᒐ	10A2C	v
ᴨ	10A2D	ś
ᴩ	10A2E	ṣ
ᒐ	10A2F	s
ᴚ	10A30	z
ᒎ	10A31	h
ᛁ	10A40	1
ᴘ	10A41	2
ᴍ	10A42	3
ᴄ	10A43	4
ᒎ	10A44	10
ᴣ	10A45	20
ᒐ	10A46	100
ᴙ	10A47	1000
'	10A50	·

Character	Unicode code point	Transcription
ᴑ	10A51	◦
◯	10A52	○
∈	10A53	∈
⊕	10A54	⊕
⁂	10A55	◉
|	10A56	|
‖	10A57	‖
≈	10A58	～

The following characters, omitted in the above table, should be transparent to the sorting algorithm:

Character	Unicode code point	Transcription
◌ꜗ	10A0C	-
◌	10A0D	∞
◌̄	10A38	-
◌	10A39	´ or ˍ
◌	10A3A	.

The sort value of ANUSVARA (10A0E) is context dependent:
- When followed by a space, the letters *y–h* (10A29 – 10A31), a number (10A40 – 10A47), a punctuation mark (10A50 – 10A57), or any non-Kharoṣṭhī character, it is considered to be a 'true' *anusvāra* and follows *o* (10A07) in the sort order.
- When followed by the letters *k–gh,* or *ḱ* (10A10 – 10A13, or 10A32), it is considered to be a velar nasal and follows *gh* (10A13) in the sort order.
- When followed by the letters *c–ñ,* (10A15 – 10A19), it is functionally equivalent to *ñ* (10A19), and follows *j* (10A17) in the sort order.
- When followed by the letters *ṭ–ṇ,* or *ṭh* (10A1A – 10A1E, or 10A33), it is functionally

equivalent to *ṇ* (10A1E), and follows *ḍh* (10A1D) in the sort order.

- When followed by the letters *t–n,* (10A1F – 10A23), it is functionally equivalent to *n* (10A23), and follows *dh* (10A22) in the sort order.
- When followed by a vowel or the letters *p–m,* (10A00 or 10A24 – 10A28), it is functionally equivalent to *m* (10A28), and follows *bh* (10A27) in the sort order.

The sort values of the Kharoṣṭhī digits will not produce a correct sorting of Kharoṣṭhī numerals, because of the multiplicative element in the Kharoṣṭhī numeral system. If possible, Kharoṣṭhī numbers should be sorted according to their numeric values.

Appendix 3: Word Breaks, Line Breaks and Hyphenation

Most Kharoṣṭhī manuscripts are written as continuous text with no indication of word boundaries. Only a few examples are known where spaces have been used to separate words or verse quarters. Most scribes have tried to finish words before starting a new line. There are no examples of anything akin to hyphenation in Kharoṣṭhī manuscripts. In cases where a word would not completely fit into a line, its continuation simply appears at the beginning of the next line. Modern scholarly practice will in most cases make use of spaces and hyphenation. When necessary, hyphenation should be applied on the model of Sanskrit.

References

Allon, Mark. 2001. *Three Ekottarikāgama-Type Sūtras: British Library Kharoṣṭhī Fragments 12 and 14*. Gandhāran Buddhist Texts 2. Seattle: University of Washington Press.

Boyer, A. M., E. J. Rapson, and E. Senart. 1920–9. *Kharoṣṭhī Inscriptions Discovered by Sir Aurel Stein in Chinese Turkestan.* 3 pts. (pt. 3 by Rapson and P. S. Noble). Oxford: Clarendon Press.

Braarvig, Jens, ed. 2000. *Manuscripts in the Schøyen Collection I: Buddhist Manuscripts,* vol. 1. Oslo: Hermes Publishing.

Glass, Andrew. 2000. "A Preliminary Study of Kharoṣṭhī Manuscript Paleography." Master's thesis, Department of Asian Languages and Literature, University of Washington.

Hultzsch, E. 1925. *The Inscriptions of Aśoka*. Second edition. Corpus Inscriptionum Indicarum 1. Oxford: Clarendon Press.

Konow, Sten, ed. 1929. *Kharoshṭhī Inscriptions with the Exception of Those of Aśoka*. Corpus Inscriptionum Indicarum 2.1. Calcutta: Government of India. Plate 1.

Salomon, Richard. 1996. "Brahmi and Kharoshthi" in Daniels and Bright, eds. *The World's Writing Systems.* New York: Oxford University Press.

──────. 1998. *Indian Epigraphy: A Guide to the Study of Inscriptions in Sanskrit, Prakrit, and Other Indo-Aryan Languages.* New York: Oxford University Press.

──────. 2000. *A Gāndhārī Version of the Rhinoceros Sūtra: British Library Kharoṣṭhī Fragment 5B.* Gandhāran Buddhist Texts 2. Seattle: University of Washington Press.

──────. 2001. "'Gāndhārī Hybrid Sanskrit': New Sources for the Study of the Sanskritization of Buddhist Literature." *Indo-Iranian Journal* 44: 241–252.

Comments or Discussion

Please send any responses to this proposal to Andrew Glass (email: asg@u.washington.edu). Please also CC to Richard Salomon (email: rsalomon@u.washington.edu) and Stefan Baums (email: baums@u.washington.edu).