ISO/IEC JTC1/SC2/WG2 N1714R
1998-03-18

Universal Multiple-Octet Coded Character Set
International Organization for Standardization
Международная организация по стандартизации

ISO/IEC JTC1/SC2/WG2
Universal Multiple-Octet Coded Character Set (UCS) – ISO/IEC 10646
Secretariat: ANSI

Doc Type: **Expert Contribution**
Title: **Bengali encoding in ISO/IEC 10646 and BSD 1520**
Source: **Michael Everson, Everson Gunn Teoranta (IE)**
Project: **JTC1.02.18.01**
Status: **Discussion**
Action: **FYI**
Due Date: **—**
Distribution: **WG2**
Medium: **Paper and PDF**
No. of Pages: **6**

This contribution is in response to N1634, a contribution from Dr Golam Mowlah, Director General of Bangladesh Standards and Testing Institution (BSTI), dated 1997-06-29. In that document, BSTI requests the “incorporation” of the code table of BSD 1520:1995 “Bangla Coded Character Set” into ISO/IEC 10646. The rationale given for this incorporation is that the “Bengali coded character which is available in ISO/IEC publication 10646:1993 edition in row 9, locating 0980–09FF (hex) is not applicable for Bangladesh as that coded characters are not similar to [the Bangladeshi] national coded character”.

WG2 has made a thorough review of the code table in N1634 and has compared it to the Bengali code table in ISO/IEC 10646 and has found that BSD 1520 is mappable to 10646 as it currently stands, that no change to 10646 is required, and that with an adequate table-lookup it will be possible for data coded in BSD 1520 to be transformed into 10646 coding. This determination is in accordance with the agreed WG2 and UTC principle that character codes (or character names) *shall not be changed* once they have been allocated.

The analysis given below is imperfect because, unfortunately, N1634 does not give name identifications for BSD 1520 characters 83–FF, and so this analysis may have errors deriving from our imperfect knowledge of Bengali conjunct formation. However, it is clear from the analysis that even those conjuncts which we have not identified can be represented with the current 10646 coding conventions – it is simply a question of decomposing the conjunct forms. We were unable to identify 4 of the 125 characters in BSD 1520 from the range 83–FF (99, C6, F2, F8).

Almost all of the characters from that range are glyph representations of an underlying coding compatible with 10646 coding for this script. The character set appears (because of the characters CA, and D0–D6) to be based on an Apple Macintosh implementation for Bengali. Numerous precomposed conjuncts (e.g. 8D–92) are given, as well as a number of on-screen composable variants (such as the positional variants for following -RA, intended to combine with characters of differing widths and heights (DE–E0, ৢ ৣ ৲). If the implementations in Bangladesh are fully

WorldScript compatible, the table-lookup for conversion to 10646 will be relatively straightforward. If the implementations are glyph-based, the lookup-table mapping will be more complex, but still definable and simple enough to implement.

Four characters appearing in UCS do not appear in BSD 1520: U+09C4, U+09E1, U+09E2, U+09E3. One character, BENGALI ABAGRAHA, does not appear in either ISO/IEC 10646 or BSD 1520 (cf. the LOC Romanization tables). Was this unified with DEVANAGARI AVAGRAHA or is a unique BENGALI ABAGRAHA required? The Unicode Standard should note, informatively, that ৭ BENGALI DIGIT SEVEN is used to denote the Hindu god Ganesha.

The table which follows gives the BSD 1520 character set, with its hexadecimal notation, and with strings of 10646 characters to which they can be mapped if the implementation in Bangladesh is glyph-based. Many scripts require mapping tables to convert between 10646 and local standards, and most vendors' technology already has facilities to support such mapping. However, full information on local standards is important to the correct formation of those tables.

BSTI should also be aware that SC2/WG2 and the Unicode Consortium take its contribution in N1634 seriously, and wish to ensure them that implementors of 10646 and Unicode intend to support mapping between BSD 1520 and 10646. However, because UCS is *not* a "registry" for national standards, BSD 1520:1995 cannot be adopted to *replace* the existing code table in Row 9 of the Standard. What BSTI can do, to facilitate mapping between BST 1520 and 10646, in order to produce comprehensive lookup tables for mapping between UCS and the Bangladeshi Standard, is provide us with a full description of the implementation (Worldscript-based or glyph-based) and, if glyph based, an exhaustive list of Bangla syllables coded according to the principles of BSD 1520:1995.

If careful analysis of BSD 1520 shows that one or more characters cannot be mapped directly (or with reasonable, local, context analysis) to 10646, then those characters may be candidates to be added to the standard.

Char.	1520	UCS
SP	20	0020
!	21	0021
"	22	2033
#	23	0023
৳	24	09F3
%	25	0025
×	26	00D7
'	27	0027
(	28	0028
)	29	0029
*	2A	002A
+	2B	002B
,	2C	0026
-	2D	002D
.	2E	002E
/	2F	002F
০	30	09E6
১	31	09E7
২	32	09E8
৩	33	09E9
৪	34	09EA
৫	35	09EB
৬	36	09EC
৭	37	09ED
৮	38	09EE
৯	39	09EF
:	3A	003A
;	3B	003B
<	3C	003C
=	3D	003D
>	3E	003E
?	3F	003F
@	40	0040
অ	41	0985
আ	42	0986
ই	43	0987
ঈ	44	0988
উ	45	0989
ঊ	46	098A
ঋ	47	098B
এ	48	098F
ঐ	49	0990
ও	4A	0993
ঔ	4B	0994
ক	4C	0995
খ	4D	0996
গ	4E	0997
ঘ	4F	0998
ঙ	50	0999
চ	51	099A
ছ	52	099B
জ	53	099C
ঝ	54	099D
ঞ	55	099E
ট	56	099F
ঠ	57	09A0
ড	58	09A1

ঃ	76	0982
ঁ	77	0983
ং	78	0981
়	79	
ঽ	7A	0964
া	7B	007B
ি	7C	0965
ী	7D	007D
ু	7E	007E
ূ		
ৃ	80	09FA
ৄ	81	098C
ে	82	09BE
ৈ	83	09BF
ো	84	09C0
ৌ	85	09C1
্	86	09C2
ৎ	87	09C3
ৗ	88	09C7
ড়	89	09C8
ঢ়	8A	09CB?
য়	8B	09CC?
ৠ	8C	09CD 0995
ৡ	8D	0995 09CD 0995
ৢ	8E	0995 09CD 099F
ৣ	8F	0995 09CD 0961
০	90	0995 09CD 09B0
১	91	09CD 0995 09CD 09B0
২	92	0995 09CD 096E

ঢ	59	09A2
ণ	5A	09A3
[	5B	005B
\	5C	005C
]	5D	005D
^	5E	005E
_	5F	005F
`	60	0060
ত	61	09A4
থ	62	09A5
দ	63	09A6
ধ	64	09A7
ন	65	09A8
প	66	09AA
ফ	67	09AB
ব	68	09AC
ভ	69	09AD
ম	6A	09AE
য	6B	09AF
র	6C	09B0
ল	6D	09B2
শ	6E	09B6
ষ	6F	09B7
স	70	09B8
হ	71	09B9
ড়	72	09DC
ঢ়	73	09DD
য়	74	09DF
ৎ	75	0961 09CD

ক্ষ	93	0995 09CD 09AE
খ	94	09CD 0096
গ	95	09CD 0997
ঘ	96	0997 09C1
ঙ	97	0997 09CD 09A6
চ	98	0997 09CD 09A6
ছ	99	
জ	9A	0999 09CD 0995
ঝ	9B	0950 09CD 094E
ঞ	9C	09CD 099A
ট	9D	099A 09CD 099E
ঠ	9E	099C 09CD 099C
ড	9F	099C 09CD 099D
ঢ	A0	099C 09CD 099E
ণ	A1	09F9
ত	A2	099E 09CD 099A
থ	A3	099E 09CD 099B
দ	A4	099E 09CD 099C
ধ	A5	099E 09CD 099D
ন	A6	0956 09C1
প	A7	09A1 09CD 09A1
ফ	A8	09A3 09CD 0956
ব	A9	09A3 09CD 09A0
ভ	AA	09A3 09C1
ম	AB	09CD 0961
য	AC	09CD 0961 09CD 0961
র	AD	0961 09CD 0961
ল	AE	0961 09CD 09B0
শ	AF	09CD 0961 09CD 09B0

ষ	B0	09CD 09B9
স	B1	09A4 09CD 09A5
হ	B2	09A6 09CD
ড়	B3	09A6 09CD
ঢ়	B4	09A6 09CD 09A6
য়	B5	09A6 09CD 09A7
ৠ	B6	09A6 09CD 09AC
ঌ	B7	09A8 09CD
ৡ	B8	09A8 09CD
ৢ	B9	09CD 09A8
ৣ	BA	09CD 09A8
০	BB	09A8 09CD 09A0
১	BC	09A8 09CD 09A1
২	BD	09A8 09CD 09A7
৩	BE	09A8 09CD 09B6
৪	BF	09B6 09CD
৫	C0	09CD 09B6
৬	C1	09B6 09CD 099F
৭	C2	09B6 09CD 09B6
৮	C3	09B6 09C1
৯	C4	09CD 09B7
ৰ	C5	09F0
ৱ	C6	
৲	C7	09CD 09AC
৳	C8	09CD 09AC
৴	C9	09CD 09AC
৵	CA	00CA
NBSP	CB	09AC 09CD 099C
৷	CC	09AC 09CD 09A6

ক্ষ	CD	09AC 09CD 09A7
ৎ	CE	09CD 09AD
ৗ	CF	09AD 09CD 09B0
৷	D0	2013
৸	D1	2014
“	D2	201C
”	D3	201D
‘	D4	2018
’	D5	2019
÷	D6	00D6
৯	D7	09CD 09AD 09CD 09B0
ঃ	D8	096A 09CD
অ	D9	09CD 096A
আ	DA	09CD 096A
ই	DB	09CD 09AF
ঈ	DC	09B0 09CD
উ	DD	
ঊ	DE	09CD 09B0
ঋ	DF	09CD 09B0
ঌ	E0	09CD 09B0
঍	E1	09B2 09CD
঎	E2	09B2 09CD
এ	E3	09CD 09B2
ঐ	E4	09CD 09B2
ও	E5	09B2 09CD 09A1
ঔ	E6	09B6 09C1
ক	E7	09B6 09CD
খ	E8	09B7 09CD
গ	E9	09B7 09CD 099E

ঘ	EA	09B7 09CD 099F
ঙ	EB	09B7 09CD 09A0
চ	EC	09B8 09CD
ছ	ED	___ 09CD
জ	EE	09B9 09C1
ঝ	EF	09B9 09C3
ঞ	F0	09B9 09CD 09A3
ট	F1	09B9 09CD 09AE
ঠ	F2	
ড	F3	09C1
ঢ	F4	09CD 09A4
ণ	F5	09C2
ত	F6	09C3
থ	F7	09C3
দ	F8	09C8?
ধ	F9	09CD
ন	FA	09F4
঩	FB	09F5
প	FC	09F6
ফ	FD	09F7
ব	FE	09F8
ভ	FF	09BC
ম		
য		
র		
঱		
ল		
঳		
঴		
঵		
শ		
ষ		
স		
হ		