Universal Multiple-Octet Coded Character Set
International Organization for Standardization
Organisation Internationale de Normalisation
Международная организация по стандартизации

Doc Type: **Working Group Document**
Title: **Proposal for encoding the Ol Cemet' script in the BMP of the UCS**
Source: **Michael Everson, EGT (IE)**
Status: **Expert Contribution**
Action: **For consideration by JTC1/SC2/WG2 and UTC**
Date: **1999-01-29**

This is a revision of an exploratory proposal by James Kass, and contains the proposal summary form.

A. Administrative

1. Title
Proposal for encoding the Ol Cemet' script in the BMP of the UCS.
2. Requester's name
Michael Everson, EGT (WG2 member for Ireland).
3. Requester type
Expert contribution.
4. Submission date
1999-01-29.
5. Requester's reference
6a. Completion
This is a complete proposal.
6b. More information to be provided?
No.

B. Technical -- General

1a. New script? Name?
Yes. Ol Cemet'.
1b. Addition of characters to existing block? Name?
No.
2. Number of characters
43
3. Proposed category
Category A.
4. Proposed level of implementation and rationale
Level 2 as it uses diacritics in the Brahmic style.
5a. Character names included in proposal?
Yes.
5b. Character names in accordance with guidelines?
Yes.
5c. Character shapes reviewable?
Yes (see below).

6a. Who will provide computerized font?
James Kass via Michael Everson.
6b. Font currently available?
Yes.
6c. Font format?
TrueType.
7a. Are references (to other character sets, dictionaries, descriptive texts, etc.) provided?
Yes, see bibliography below.
7b. Are published examples (such as samples from newspapers, magazines, or other sources) of use of proposed characters attached?
No.
8. Does the proposal address other aspects of character data processing?
Yes, see Unicode properties below.

C. Technical -- Justification

1. Contact with the user community?
No. We need to contact Norman Zide and get his opinion of this proposal.
2. Information on the user community?
Speakers of the Santali language, whose population is 5,800,000, with 25%–50% literacy, according to the SIL *Ethnologue*.
3a. The context of use for the proposed characters?
To write the Santali language. Latin, Devanagari, Bengali, and Oriya scripts have also been used to write Santali.
3b. Reference
See bibliography.
4a. Proposed characters in current use?
Yes.
4b. Where?
In primary and adult education (general use)
5a. Characters should be encoded entirely in BMP?
Yes.
5b. Rationale
Contemporary use.
6. Should characters be kept in a continuous range?
Yes.
7a. Can the characters be considered a presentation form of an existing character or character sequence?
No.
7b. Where?
7c. Reference
8a. Can any of the characters be considered to be similar (in appearance or function) to an existing character?
No.
8b. Where?
8c. Reference
9a. Combining characters or use of composite sequences included?
Yes.
9b. List of composite sequences and their corresponding glyph images provided?

10. Characters with any special properties such as control function, etc. included?
No. They freely combine as in Brahmic scripts.

E. Proposal

The Ol Cemet' script, also called Ol or Ol Ciki, was invented by Pandit Raghunath Murmu in the first half of the 20th century CE. Ol Cemet' is alphabetic, sharing none of the syllabic properties of the other Indic scripts. Members of several linguistic groups in India apparently felt that a unique script was necessary for their cultural identities. As a result, more than a dozen scripts were devised. Most of these scripts are forgotten now, but the Ol Cemet' script has received some official recognition and Raghunath has been honoured by the Orissan government.

Languages using the Ol Cemet' script: Santali (a Munda language of India). According to Ethnologue, Santali's various dialects are spoken by 5.8 million people with 25% to 50% literacy, mostly in India with a few in Nepal and Bangladesh. The Ol Cemet' script is used for the southern dialect of Santali as spoken in the Orissan Mayurbhañj district. While this dialect has only six vowels, the Santal Parganas dialect has eight or nine vowels. The extra Santal Parganas vowels are reportedly made by combining existing vowels with diacritics. There is room in the table for 2 or 3 diacritics.

Ol Cemet' has recently been promoted by some Santal organizations, with uncertain success, for use in certain other Munda languages in the Chota Nagpur area as well as the Dravidian Kuḍux language.

Zide 1996 says: "One ingenious – "scientific" – and unique feature of Ol Cemet' that certainly increases the efficiency of writing Santali is the deglotallizing *o̤họt'* diacritic. This neatly preserves the morphophonemic relationships between the glottalized and voiced equivalents: the former occurs word-finally and at certain word-internal preconsonantal junctures, the latter prevocalically, but never morpheme-initially in these alternations. Thus, *o̤k'* is the name of a letter that represents both [k′] and [g]. Two further diacritics include a horizontal loop added at the top right of the character for the aspiration of consonants, and a raised dot for vowel nasalization." Unfortunately Zide does not indicate the glyph of the *o̤họt'* character.

Names and ordering

Characters are arranged in a 5 by 6 matrix, named in a conventional way as shown in the names list. The first characters in each row (LO, LA, LI, LU, LE, LOO) are vowels. xx2D has been named VISARGA, but it indicates consonant aspiration, and a better name should be sought. xx2E has been named ANUSVARA, and it does indicate vowel nazalization, but a better name should be sought.

Unicode Character Properties

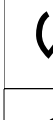
Spacing letters, category "Lo", bidi category "L" (strong left to right)
xx00 - xx1D
Numbers, decimal digits, category "Nd", bidi category "L" (strong left to right)
xx20-xx29
Non-spacing marks, category "Mc", bidi category "ON" (other neutral); combining priorities in parentheses:
xx2D (232)
Non-spacing marks, category "Mn", bidi category "ON" (other neutral); combining priorities in parentheses:
xx2E - xx2F (230) [xx2F is conjectural]

Bibliography

Zide, Norman. 1996. “Scripts for Munda languages”, in Peter T. Daniels and William Bright, eds. *The world’s writing systems*. New York; Oxford: Oxford University Press. ISBN 0-19-507993-0

Zide gives other sources in his bibliography, none of which I have seen.

TABLE XXX - Row xx: OL CEMET

	xx0	xx1	xx2
0			
1			
2			
3			
4			
5			
6			
7			
8			
9			
A			
B			
C			
D			
E			
F			

G = 00
P = 00

TABLE XXX - Row xx: OL CEMET

dec	hex	Name
	00	OL CEMET LETTER LO
	01	OL CEMET LETTER OT
	02	OL CEMET LETTER OKH
	03	OL CEMET LETTER ONG
	04	OL CEMET LETTER OL
	05	OL CEMET LETTER LA
	06	OL CEMET LETTER AK
	07	OL CEMET LETTER ACH
	08	OL CEMET LETTER AM
	09	OL CEMET LETTER AW
	0A	OL CEMET LETTER LI
	0B	OL CEMET LETTER IS
	0C	OL CEMET LETTER IH
	0D	OL CEMET LETTER INY
	0E	OL CEMET LETTER IR
	0F	OL CEMET LETTER LU
	10	OL CEMET LETTER UC
	11	OL CEMET LETTER UTH
	12	OL CEMET LETTER UNN
	13	OL CEMET LETTER UY
	14	OL CEMET LETTER LE
	15	OL CEMET LETTER EP
	16	OL CEMET LETTER EDD
	17	OL CEMET LETTER EN
	18	OL CEMET LETTER ERR
	19	OL CEMET LETTER LOO
	1A	OL CEMET LETTER OOTT
	1B	OL CEMET LETTER OOPH
	1C	OL CEMET LETTER OOWW
	1D	OL CEMET LETTER OOHH
	1E	(This position shall not be used)
	1F	(This position shall not be used)
	20	OL CEMET DIGIT ZERO
	21	OL CEMET DIGIT ONE
	22	OL CEMET DIGIT TWO
	23	OL CEMET DIGIT THREE
	24	OL CEMET DIGIT FOUR
	25	OL CEMET DIGIT FIVE
	26	OL CEMET DIGIT SIX
	27	OL CEMET DIGIT SEVEN
	28	OL CEMET DIGIT EIGHT
	29	OL CEMET DIGIT NINE
	2A	(This position shall not be used)
	2B	(This position shall not be used)
	2C	(This position shall not be used)
	2D	OL CEMET VISARGA
	2E	OL CEMET ANUSVARA
	2F	OL CEMET OHOTT

dec	hex	Name

Group 00 **Plane 00** **Row xx**