L2/24-148
JTC1/SC2/WG2 N5266

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 to Encode a Set of 128 User-Defined Variation Selectors**
Source:	**Andrew West and Michael Everson**
Status:	**Individual Contribution**
Action:	**For consideration by JTC1/SC2/WG2 and UTC**
Date:	**2024-05-30**

1. Introduction

A perennial problem encountered by academic users of historic scripts is the desire to be able to represent epigraphic or calligraphic variants of characters at the text level. Although it is possible to use the OpenType *Character Variants* feature (cv01–cv99) to define up to one hundred variant glyph forms for a character in a font, these variants are only accessible when using advanced typesetting software such as Microsoft Word or web browsers (by means of CSS), and it is not possible to preserve such variants in plain text.

An obvious solution to the problem is to use variation sequences. The Ideographic Variation Database already allows for the use of VS17–VS256 to register up to 240 ideographic variation sequences per character for the Han script and other ideographic scripts such as Tangut. This allows users of ideographic scripts the possibility to define variation sequences for epigraphic, calligraphic, or typographic variants that they encounter in the texts they study. However, this option is not available to academics studying non-ideographic scripts.

The Unicode Standard defines a set of Standardized Variation Sequences using 16 variation selectors (VS1–VS16), which in theory could be extended to allow the definition of epigraphic, calligraphic, or typographic glyph variants for historic scripts. However, the Unicode Technical Committee (UTC) has persistently refused requests for standardized variation sequences for simple glyph variants, and has indicated that the UTC does not wish to be in the business of defining and maintaining lists of glyph variants for historic scripts.[1] This is probably a reasonable position to take, and in fact standardized variation sequences would not be the best solution for end-users as the bar for getting a standardized variation sequence accepted by the UTC is very high, and the limit of 16 variation selectors is certainly too small for some scripts.

[1] Examples of rejected proposals for standardized variation sequences for glyph variants include: Karl Pentzlin: *Proposal to add Variation Sequences for Latin and Cyrillic letters* (L2/10-280, L2/11-059); Aleksandr Andreev *et al.*: *Proposal to Use Standardized Variation Sequences to Encode Church Slavonic Glyph Variants in Unicode* (L2/13-153); Mark Davis: *Playing Card Variation Selectors* (L2/14-223); Andrew West & Eiso Chan: *Proposal to define Standardized Variation Sequences for two Chinese ideographs* (L2/16-109); Andrew West: *Proposal to define 21 variation sequences for Ogham letters* (L2/16-110); Andrew West: *Preliminary proposal to define 357 variation sequences for Tangut ideographs* (L2/16-111); Michael Everson: *Proposal to add standardized variation sequences for chess notation* (L2/17-077); Eduardo Marín Silva: *Proposal to add 6 standardized variation sequences for counting rods* (L2/17-085); and Eiso Chan & Selena Wei: *Proposal to define Standardized Variation Sequences for BOPOMOFO LETTER I* (L2/18-020).

2. Proposed Solution

An alternative solution proposed in this document is to allow users to define their own variation sequences using a set of 128 User-defined Variation Selectors at E0200..E027F (abbreviated as UVS1–UVS128) in the Supplementary Special-purpose Plane. As with VS1–VS256, the set of user-defined variation selectors would have the default ignorable property (which is already pre-assigned to the proposed range of code points).

This solution would allow font developers, publishers, or user groups to define variation sequences without recourse to the UTC, and without any bureaucratic registration procedure.[2] There would be no guarantee that any given User-defined Variation Sequence (UVS) would be unique, as the same UVS could be defined for multiple different variants of the same character by multiple different sources, but this should be an acceptable and manageable limitation.[3] The use of a UVS or set of UVSes would be akin to a Private Use Area (PUA) agreement between users, with the understanding that for a particular font a particular UVS would represent a particular character variant. But it has the considerable advantages over PUA that the default fallback glyph will be the expected base character, and that text processing operations can treat the UVS and its base character the same.

Accepting the proposed set of user-defined variation selectors would give academic users the ability to use Unicode as a practical solution to their text encoding needs, and at the same time it would relieve the UTC of the burden of evaluating requests for character variants.

3. Use Cases

This proposal has been motivated by two recent documents by academic users of the Ogham and Runic scripts. Firstly, a blog post by Adrian Doyle, *The Future of Digital Ogam: Potential Updates to the Unicode Ogham Block to Facilitate Modern Usage* (24 April 2024), which bemoans the fact that there is no way to accurately represent the six-stroke graphical variant of U+168F ᚏ using Unicode. Doyle's suggested solution of encoding a set of combining stroke characters for each of the four *aicmí* (i.e. combining versions of ᚁ ᚂ ᚃ ᚄ) is not viable in our opinion, as it would create the potential for multiple spellings, and would be a spoofing risk. However, a user-defined variation sequence could be an acceptable solution.

Secondly, an article from runologists Elisabeth Maria Magin and Marcus Smith, *(R)Unicode: Encoding and Sustainability Issues in Runology* (L2/24-129; DOI:10.5617/dhnbpub.10657), suggests the use of variation sequences as their preferred solution for the issue of runic variants. However, the authors recognize that the UTC has not been supportive of similar proposals for variation sequences in the past, and they conclude that "variation sequences for runes could ... be defined for runological use and data interchange within the domain, outside of but in complement to the Unicode standard". This implies that runologists could define their own non-conformant variation sequences if the UTC rejects a future proposal for standardized variation sequences. This would not be desirable, and I believe that user-defined variation sequences would be a far better solution for both runologists and the UTC.

[2] An example of a user group would be the Medieval Unicode Font Initiative (MUFI), which was responsible for successful encoding proposals for a large number of Latin letters, marks, punctuation, and symbols. However, the latest version of MUFI (v. 4.0) still defines some 738 PUA characters which are unlikely to be candidates for encoding. Converting these PUA assignments to user-defined variation sequences would be advantageous.

[3] It is expected that conflicting assignments of variation sequences would not be a problem in most script domains, where experts would collaborate to define an agreed set of user-defined variation sequences.

4. Unicode Properties

Block name: User-defined Variation Selectors

Block range: E0200..E027F

Character names: USER-DEFINED VARIATION SELECTOR-1 through USER-DEFINED VARIATION SELECTOR-128 (abbreviated UVS1 through UVS128)

General Category: Mn

Canonical Combining Class: 0

Bidi Class: NSM

Bidi Mirrored: No

Line Break: CM

Other properties: Grapheme Extend; Default Ignorable Code Point; ID Continue; XID Continue; Variation Selector

5. Proposal Summary Form

SO/IEC JTC 1/SC 2/WG 2
PROPOSAL SUMMARY FORM TO ACCOMPANY SUBMISSIONS
FOR ADDITIONS TO THE REPERTOIRE OF ISO/IEC 10646[4]
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 a Set of 128 User-Defined Variation Selectors***
2. Requester's name: *Andrew West and Michael Everson*
3. Requester type (Member body/Liaison/Individual contribution): *Individual contribution*
4. Submission date: *2024-05-30*
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: *User-Defined Variation Selectors*
   b. The proposal is for addition of character(s) to an existing block: *NO*
      Name of the existing block:
2. Number of characters in proposal: *128*
3. Proposed category (select one from below - see section 2.2 of P&P document):
   A-Contemporary ( ) B.1-Specialized (small collection) (X) 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. 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? *NO*
5. Fonts related:
   a. Who will provide the appropriate computerized font to the Project Editor of 10646 for publishing the standard?
   Michael Everson
   b. Identify the party granting a license for use of the font by the editors (include address, e-mail, ftp-site, etc.):
   Michael Everson
6. References:
   a. Are references (to other character sets, dictionaries, descriptive texts etc.) provided? *NO*
   b. Are published examples of use (such as samples from newspapers, magazines, or other sources) of proposed characters attached? *NO*
7. 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)? *NO*

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 Unicode Character Database (http://www.unicode.org/reports/tr44/) and associated Unicode Technical Reports for information needed for consideration by the Unicode Technical Committee for inclusion in the Unicode Standard.

TP[4]PT Form number: N4102-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, 2008-05, 2009-11, 2011-03, 2012-01)

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.)? *NO*
   If YES, with whom?
   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? *NO*
   Reference:
4. The context of use for the proposed characters (type of use; common or rare) *N/A*
   Reference:
5. Are the proposed characters in current use by the user community? *N/A*
   If YES, where? Reference:
6. After giving due considerations to the principles in the P&P 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, or could be confused with, an existing character? *NO*
    If YES, is a rationale for its inclusion provided?
    If YES, reference:
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?
    If YES, reference:
    Is a list of composite sequences and their corresponding glyph images (graphic symbols) provided?
    If YES, reference:
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)
13. Does the proposal contain any Ideographic compatibility characters? *NO*
    If YES, are the equivalent corresponding unified ideographic characters identified?
    If YES, reference: