ISO/IEC JTC1/SC2/WG2 N3643
L2/09-171
2009-04-30

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 redefine the scope of Ideographic Description Sequences and to encode four additional Ideographic Description Characters**
Source: **Andrew West**
Status: **Individual Contribution**
Action: **For consideration by JTC1/SC2/WG2 and UTC**
Date: **2009-04-30**

1. Introduction

This is a proposal to redefine the scope of Ideographic Description Sequences (IDS) to cover non-Han scripts (see Section 3), and also to encode four new Ideographic Description Characters (IDC) in order to represent IDS sequences in non-Han scripts (see Section 2).

1.1 Current Definition of IDS Sequences

The Unicode Standard 5.1 section 12.2 defines Ideographic Description Sequences as follows:

> ***Ideographic Description Sequences.*** Ideographic Description Sequences are defined by the following grammar. The list of characters associated with the *Unified_CJK_Ideograph* and *CJK_Radical* properties can be found in the Unicode Character Database. See *Appendix A, Notational Conventions*, for the notational conventions used here.
> *IDS* := *Unified_CJK_Ideograph* | *CJK_Radical* | *IDS_BinaryOperator IDS IDS*
> | *IDS_TrinaryOperator IDS IDS IDS*
> *IDS_BinaryOperator* := U+2FF0 | U+2FF1 | U+2FF4 | U+2FF5 | U+2FF6 | U+2FF7 | U+2FF8 | U+2FF9 | U+2FFA | U+2FFB
> *IDS_TrinaryOperator*:= U+2FF2 | U+2FF3

This definition is also echoed in ISO/IEC 10646:2003 Annex F.3.1:

> **F.3.1 Syntax of an ideographic description sequence**
> An IDS consists of an IDC followed by a fixed number of Description Components (DC). A DC may be any one of the following :
> - a coded ideograph
> - a coded radical
> - another IDS
> NOTE 1 – The above description implies that any IDS may be nested within another IDS.

The above definitions mean that Ideographic Description Characters can only be used with CJK Unified Ideographs and CJK radicals, that is to say Ideographic Description Sequences can currently only be legitimately used to describe Han ideographs.

1.2 Potential Scope for IDS Sequences

There are a number of East Asian scripts that are in the process of being encoded or that are expected to be proposed for encoding soon that comprise a large repertoire of characters that are similar in structure to Han ideographs, and that are amenable to description using an Ideographic Description Sequence mechanism:

- **Tangut** [N3577] (6,211 characters)
- **Nüshu** [N3598] (389 characters)
- **Jurchen** [N3628] (1,376 characters)
- **Khitan Ideographs** (several hundred characters)
- **Old Han** (thousands of characters)
- **Old Yi** [N3288] (up to 88,613 characters)

These scripts all have large character repertoires, and it would be useful to be able to describe the individual characters of these scripts by means of IDS sequences for the same reasons that it is useful to describe CJK ideographs using IDS sequences. For example, IRG now requires Han character submissions to include IDS sequences for all proposed characters so that it can be checked whether they are duplicates or not. It may also be beneficial if proposals for the above scripts also included IDS sequences, both to help check for duplicates and also to help with the ordering of the proposed characters. Indeed two recent non-Han script proposals have actually used IDS sequences for analysis of the proposed character repertoire:

N3288 *Preliminary Proposal to encode Classical Yi characters* (China NB, 2007-02-15)

N3288 section 8 “Ideograph Structure Description” shows a table of sixteen description characters that were used by the authors of the proposal to help order the proposed set of characters (see Figs. 1 and 2). Although the proposal does not include any actual IDS sequences for Classical Yi characters, it is clear from the statement on page 10 (“The ordering of the proposed new set ought to be in accordance with the order of Classical Yi character structure description, and the ordering of Classical Yi strokes”) that these description characters must have been used to create IDS sequences for the proposed character set. When the Classical Yi proposal is revised and resubmitted such Yi IDS sequences will be invaluable for helping evaluate the proposed character repertoire and ordering.

8. Ideograph Structure Description

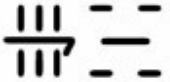
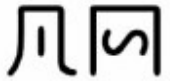
Number	DESCRIPTION	DC	IDS Acronym & Syntax	DC	IDS Presentation e.g.
1.	INDEPENDENT	1	IDC-IDP D1		
2.	LEFT TO RIGHT	2	IDC-LTR D_1 D_2		
3.	ABOVE TO BELOW	2	IDC-ATB D_1 D_2		
4.	LEFT TO MIDDLE AND RIGHT	3	IDC-LMR D_1 D_2 D_3		
5.	ABOVE TO MIDDLE AND BELOW	3	IDC-AMB D_1 D_2 D_3		
6.	FULL SURROUND	2	IDC-FSD D_1 D_2		
7.	SURROUND FROM ABOVE	2	IDC-SAV D_1 D_2		
8.	SURROUND FROM BELOW	2	IDC-SBL D_1 D_2		
9.	SURROUND FROM LEFT	2	IDC-SLT D_1 D_2		
10.	SURROUND FROM RIGHT	2	IDC-SLT D_1 D_2		

Figure 1 : N3288 page 20

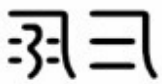
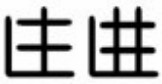
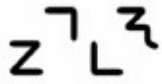
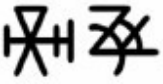
Number	DESCRIPTION	DC	IDS Acronym & Syntax	DC	IDS Presentation e.g.
11.	SURROUND FROM UPPER LEFT	2	IDC-SUL D_1 D_2		
12.	SURROUND FROM UPPER RIGHT	2	IDC-SUR D_1 D_2		
13.	SURROUND FROM LOWER LEFT	2	IDC-SLL D_1 D_2		
14.	DIAGONAL FROM LEFT TO RIGHT	2	IDC-OVL D_1 D_2		
15.	DIAGONAL FROM RIGHT TO LEFT	2	IDC-DRL D_1 D_2		
16.	OVERLAID	n	IDC-OVL Dn		

Figure 2 : N3288 page 21

Twelve of the sixteen description characters shown in N3288 are identical to existing IDC characters, but four characters ("Independent", "Surround from Right", "Diagonal from Left to Right" and "Diagonal from Right to Left") are not currently encoded:

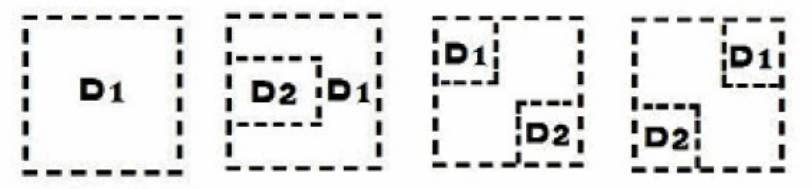


Figure 3 : Four Unencoded IDC Characters in N3288

N3577R *Proposal for a revised Tangut character set for encoding in the SMP of the UCS* (Everson *et al.*, 2009-04-08)

N3577R Appendix A provides 6,211 IDS sequences for the proposed set of Tangut characters. These IDS sequences were used by the authors of the proposal to sort the character repertoire and check for duplicates. They are also intended to enable reviewers of the proposal to easily locate any particular character by searching for its component elements.

This document uses the "independent" ideographic description character attested in N3288 in order to indicate a character for which no structural decomposition is made, as it was necessary for the automated processing of the Tangut data for all characters to be assigned an IDS sequence that starts with a description character.

Character		Radical		Residual Strokes		Total Strokes		Ideographic Description Sequence
178B6		75		8	EAGDABEC	12	CCBBEAGDABEC	
178B7		75		9	ABBBACCQB	13	CCBBABBBACCQB	
178B8		75		9	CCQCCQAMC	13	CCBBCCQCCQAMC	
178B9		75		17	DCFABBBKDBOE AAAMC	21	CCBBDCFABBBKDBOE AAAMC	
178BA		76		5	ABEAA	9	CCBEABEAA	
178BB		77		0		4	CCCQ	
178BC		77		2	HH	6	CCCQHH	
178BD		77		3	AAB	7	CCCQAAB	
178BE		77		3	AAI	7	CCCQAAI	

Figure 4 : N3577R Appendix A page 95

(*note the use of an independent IDC for U+178BB*)

2. Proposed New Characters

We believe that it is useful and appropriate to encode four additional ideographic description characters, as shown below. One further possible IDC character would be IDEOGRAPHIC DESCRIPTION CHARACTER SURROUND FROM LOWER RIGHT (⿸), but we are not proposing to encode this character as there is no indication that it is needed for any of the scripts under consideration.

Code Point	Character Name	Glyph
2FFC	IDEOGRAPHIC DESCRIPTION CHARACTER SURROUND FROM RIGHT	⿹
2FFD	IDEOGRAPHIC DESCRIPTION CHARACTER DIAGONAL FROM LEFT TO RIGHT	⿻
2FFE	IDEOGRAPHIC DESCRIPTION CHARACTER DIAGONAL FROM RIGHT TO LEFT	⿻
2FFF	IDEOGRAPHIC DESCRIPTION CHARACTER INDEPENDENT	⿼

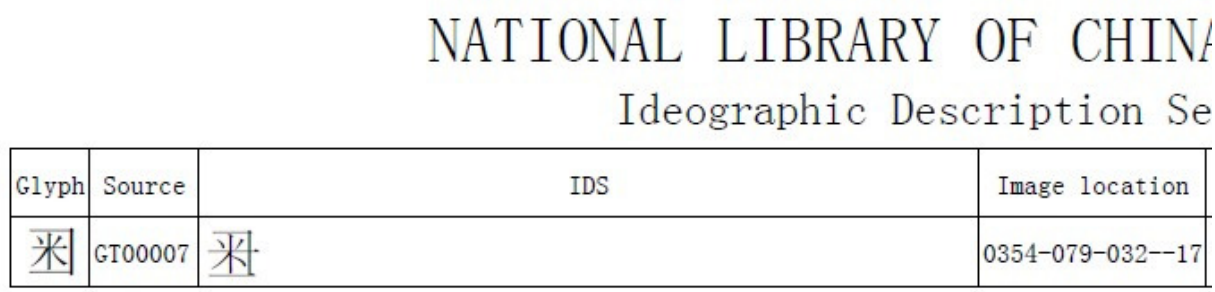
Character Properties

```
2FFC;IDEOGRAPHIC DESCRIPTION CHARACTER SURROUND FROM RIGHT;So;0;ON;;;;;N;;;;;
2FFD;IDEOGRAPHIC DESCRIPTION CHARACTER DIAGONAL FROM LEFT TO RIGHT;So;0;ON;;;;;N;;;;;
2FFE;IDEOGRAPHIC DESCRIPTION CHARACTER DIAGONAL FROM RIGHT TO LEFT;So;0;ON;;;;;N;;;;;
2FFF;IDEOGRAPHIC DESCRIPTION CHARACTER INDEPENDENT;So;0;ON;;;;;N;;;;;
```

2.1 IDC Surround from Right

This is the opposite of U+2FF7 IDEOGRAPHIC DESCRIPTION CHARACTER SURROUND FROM LEFT. In addition to its use in describing Yi characters, it would also be useful for describing some Old Han characters, for example the Oracle Bone Script form of the characters representing the Shang royal titles Baoyi 報乙, Baobing 報丙 and Baoding 報丁 are written as the characters *yi*, *bing* and *ding* within a right-surrounding enclosure::

And there is at least one CJK ideograph proposed for encoding which is also analysable as a component surrounded from the right:


NATIONAL LIBRARY OF CHINA

Ideographic Description Se

Glyph	Source	IDS	Image location
米	GT00007	米	0354-079-032--17

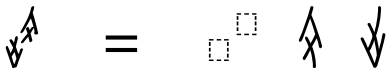
Figure 5 : IRG N1467 page 2

2.2 IDC Diagonal From Left to Right

This is a binary operator for an IDS sequence that comprises two elements that are diagonally opposite each other, the first at the top left and the second at the bottom right. In addition for use in describing Yi characters this character would also be useful for describing Nüshu characters. For example the proposed Nüshu character U+1B0DC 𛃜 could be analysed as the proposed character U+1B033 𛀳 in the top left corner and the proposed character U+1B057 𛁗 in the bottom left corner:

2.3 IDC Diagonal From Right to Left

This is a binary operator for an IDS sequence that comprises two elements that are diagonally opposite each other, the first at the top right and the second at the bottom left. In addition for use in describing Yi characters this character would be very useful for describing Nüshu characters as many Nüshu characters are diagonally skewed (see N3598 *Proposal for encoding Nüshu in the SMP of the UCS*). For example the proposed Nüshu character U+1B12E 𛄮 could be analysed as the proposed character U+1B02D 𛀭 in the top right corner and the proposed character U+1B033 𛀳 in the bottom left corner:



2.4 IDC Independent

This character is a unary operator for an IDS sequence that comprises a single element, i.e. the character itself or an equivalent character (for example a radical that has the same form as the character in question). It is used to indicate that a character is itself a basic element that is not decomposable into a sequence of any other elements. Although the current definition of an IDS sequence means that a single unified ideograph or radical by itself is also an IDS sequence, it is often useful if every character in a particular set of characters is represented by an IDS sequence that starts with an IDC character, for example in order to facilitate automated processing. The independent IDC allows for the easy visual identification and automated processing of unary IDS sequences.

3. Proposed Changes to the Unicode and ISO/IEC 10646 Standards

At present the Unicode Standard section 12.2 restricts IDS sequences to characters with one of the following properties: *IDS_Binary_Operator*, *IDS_Trinary_Operator*, *Unified_CJK _Ideograph* or *CJK_Radical*. We note in passing that **Unified_CJK_Ideograph** and **CJK_Radical** are not defined in the Unicode Standard, and appear to be non-standard aliases for the **Unified_Ideograph** and **Radical** properties respectively (see PropertyAliases.txt and UCD.htm).

In order to accommodate the need to use description sequences for non-Han scripts, we propose that a number of changes to Unicode character properties be made, as outlined below.

3.1 IDS_Unary_Operator Property

A new **IDS_Unary_Operator** property (abbreviation **IDSU**) should be added to the list of Binary Properties. The only character to have this property should be the new IDEOGRAPHIC DESCRIPTION CHARACTER INDEPENDENT character that we are proposing to encode.

3.2 Radical Property

The **Radical** property should be extended to all sets of *radical* characters, not just *CJK radicals*. Specifically, the Radical property should be applied to all characters in the existing Kangxi Radicals, CJK Radicals Supplement and Yi Radicals blocks, as well as to all sets of radicals and/or character components that may be encoded in the future for Jurchen, Khitan Ideographs, Old Han, Old Yi and Tangut.

3.3 IDS_Component Property

A new **IDS_Component** property (abbreviation **IDSC**) should be added to the list of Binary Properties. This property should be applied to all characters that have either the Unified_Ideograph or Radical property, as well to Yi Syllables and all Nüshu, Jurchen, Khitan Ideographs, Old Han, Old Yi and Tangut characters if and when they are encoded.

If these changes are made the definition of *Ideographic Description Sequences* in the Unicode Standard can be changed to:

> ***Ideographic Description Sequences.*** Ideographic Description Sequences are defined by the following grammar. The list of characters associated with the *IDS_Component* property can be found in the Unicode Character Database. See *Appendix A, Notational Conventions*, for the notational conventions used here.
> *IDS* := *IDS_Component* | *IDS_Unary_Operator IDS_Component* | *IDS_Binary_Operator IDS IDS* | *IDS_Trinary_Operator IDS IDS IDS*
> *IDS_Unary_Operator*:= U+2FFF
> *IDS_Binary_Operator* := U+2FF0 | U+2FF1 | U+2FF4 | U+2FF5 | U+2FF6 | U+2FF7 | U+2FF8 | U+2FF9 | U+2FFA | U+2FFB | U+2FFC | U+2FFD | U+2FFE
> *IDS_Trinary_Operator*:= U+2FF2 | U+2FF3

Note that for backwards compatibility with existing collections of IDS sequences, the syntax defined above continues to allow an IDS sequence to comprise a single IDS Component character with no preceding IDS operator.

If this proposal is accepted then Section 12.2 of the Unicode Standard would need to be rewritten and moved out of Chapter 12.

Annex I of the forthcoming ISO/IEC 10646:2009 would also need to be rewritten to make it clear that IDS sequences are not limited to CJK "ideographs" and "radicals". This could be done by adding a note to the effect that in Annex I "ideograph" refers to a coded CJK unified ideograph or a coded character in the Yi Syllables block (and Tangut, Jurchen, Nushu etc. blocks when these scripts are encoded); and that "radical" refers to a coded character in the Kangxi Radicals block, CJK Radicals Supplement block and Yi Radicals block (and Tangut Radicals block, Jurchen Radicals block etc. when these scripts are encoded).

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:** ***Proposal to encode four additional Ideographic Description Characters***
2. Requester's name: *Andrew West*
3. Requester type (Member body/Liaison/Individual contribution): *Individual*
4. Submission date: *2009-04-30*
5. Requester's reference (if applicable): *N/A*
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):
     - Proposed name of script:
   - b. The proposal is for addition of character(s) to an existing block: *X*
     - Name of the existing block: *Ideographic Description Characters*
2. Number of characters in proposal: *4*
3. Proposed category (select one from below - see section 2.2 of P&P document):

A-Contemporary	X	B.1-Specialized (small collection)		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? *Yes*
5. Who will provide the appropriate computerized font (ordered preference: True Type, or PostScript format) for publishing the standard? *Andrew West*
   If available now, identify source(s) for the font (include address, e-mail, ftp-site, etc.) and indicate the tools used:
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? *Yes*
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)? *Yes*

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.

[1] Form number: N3152-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)

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? *other experts*
   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) *specialized*
   Reference:
5. Are the proposed characters in current use by the user community? *Yes*
   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? *Yes*
   If YES, is a rationale provided? *No*
   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? *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? *No*
    If YES, describe in detail (include attachment if necessary)
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: