

**Universal Multiple-Octet Coded Character Set
International Organization for Standardization**

Doc Type: ISO/IEC JTC1/SC2/WG2/IRG

Title: Proposal to encode five new Ideographic Description Characters

Authors: Ken Lunde, John Jenkins & Andrew West

Status: Member Body Contribution

Action: For consideration by the IRG and UTC

Date: 2022-08-24

To follow up on [L2/21-118R](#) (aka [IRG N2492](#)) and UTC #172 Action Item 172-A52, this document is a proposal to encode five (5) new Ideographic Description Characters (IDCs) in order to handle a modest number of edge cases when managing Ideographic Description Sequences (IDSes) and IDS databases. IDCs and IDSes are extensively documented in [Section 18.2, Ideographic Description Characters](#), of the Core Specification of the Unicode Standard.

Five New Ideographic Description Characters


Four new IDCs were most recently proposed in [L2/18-012](#) (aka [IRG N2273](#)) as shown in the first four rows of the table below (the representative glyph of the fourth one was adjusted per UTC feedback), along with a fifth one that was introduced in [L2/21-118R](#):

IDC	Type	Character Name
	Binary	IDEOGRAPHIC DESCRIPTION CHARACTER SURROUND FROM RIGHT
	Binary	IDEOGRAPHIC DESCRIPTION CHARACTER SURROUND FROM LOWER RIGHT
	Unary	IDEOGRAPHIC DESCRIPTION CHARACTER HORIZONTAL REFLECTION
	Unary	IDEOGRAPHIC DESCRIPTION CHARACTER HALF-TURN ROTATION
	Binary	IDEOGRAPHIC DESCRIPTION CHARACTER COMPONENT SUBTRACTION

The first two proposed new IDCs— IDEOGRAPHIC DESCRIPTION CHARACTER SURROUND FROM RIGHT and IDEOGRAPHIC DESCRIPTION CHARACTER SURROUND FROM LOWER RIGHT—follow the pattern of similar IDCs that involve an ideograph component partially surrounding another ideograph component. Other than the possible use cases being relatively low compared to the similar IDCs, these two proposed new IDCs are not expected to be problematic nor controversial.






The third and fourth proposed new IDCs— IDEOGRAPHIC DESCRIPTION CHARACTER HORIZONTAL REFLECTION and IDEOGRAPHIC DESCRIPTION CHARACTER HALF-TURN

ROTATION—are novel in that they would become the very first *unary* IDCs. They indicate the reflection or rotation of the ideograph component that follows.

The fifth proposed new IDC— IDEOGRAPHIC DESCRIPTION CHARACTER COMPONENT SUBTRACTION—is also novel in that it specifies an ideograph component that is removed. It is a binary IDC and is therefore followed by two components:











1. An ideograph component
2. An ideograph component, such as stroke from the [CJK Strokes](#) block, that is omitted from the first ideograph component

Below are examples of this IDC used in IDSes:

- The IDSes for U+2002A 其 and U+2002B 其 are difficult to represent with existing ideograph components, but could be easily represented as 其ノ and 其ノ, respectively.
- The IDS for U+2CEBB 豕 is also difficult to represent with existing ideograph components, but could be represented as 豕ノノ.
- The IDS for U+27C27 豕 is also difficult to represent with existing ideograph components, but could be represented as 豕ノ.




A counter-example for the first example above would be to instead encode the common ideograph component of U+5176 其, U+2002A 其, and U+2002B 其 as a new ideograph component, but that accommodates only this particular case. Encoding a new IDC is much more productive.

The following table provides examples of how each of these IDCs would be used to represent existing ideographs in IDSes:

IDC	Ideograph	IDS
	U+355A 叉	 叉ノ
	U+6C37 水	 水ノ
	U+23944 五	 正
	U+20114 予	 予
	U+2002A 其	 其ノ

In terms of existing IDS implementations that use one or more of the proposed new IDCs, the [IDS.TXT](#) IDS database currently uses U+2194 ↔ LEFT RIGHT ARROW, U+21B7 ↻ CLOCKWISE TOP SEMICIRCLE ARROW, and U+2296 ⊖ CIRCLED MINUS as placeholder IDCs for the last three IDCs that are proposed in this document.

Ambiguity & Other Concerns

The two proposed new unary IDCs resolve as no-ops if used in sequence. For example, 正 corresponds to 五, but 正 corresponds to 正 itself, which is a no-op. The same is true of

𠄎𠄎𠄎, which corresponds to 𠄎 itself. In addition, reflected or rotated components can be used as ideograph components as a way to represent their non-reflected or non-rotated counterparts, such as 𠄎𠄎 and 𠄎𠄎 to represent 正 and 𠄎, respectively.

There is also inherit ambiguity in the proposed new IDC, 𠄎 IDEOGRAPHIC DESCRIPTION CHARACTER COMPONENT SUBTRACTION, about which some experts may have concerns for introducing a new dimension of adverse effects on automatic matching algorithms. For example, there are three instances of the 丿 stroke in the ideograph U+27C7 𠄎, and it is ambiguous as to which instance is removed. The way in which IDCs are currently used, which requires a non-zero amount of human intervention for interpretation, strongly suggests that this will not be issues in practical usage. Besides, an existing IDC, U+2FFB 𠄎 IDEOGRAPHIC DESCRIPTION CHARACTER OVERLAID, is already ambiguous in that human intervention is required to determine the nature of the overlaid ideograph components.

In other words, one or more of the new proposed IDCs, in particular 𠄎 IDEOGRAPHIC DESCRIPTION CHARACTER HORIZONTAL REFLECTION, 𠄎 IDEOGRAPHIC DESCRIPTION CHARACTER HALF-TURN ROTATION, and 𠄎 IDEOGRAPHIC DESCRIPTION CHARACTER COMPONENT SUBTRACTION, are likely to be considered problematic by some experts, but like other characters in the Unicode Standard, they can be ignored by those who find them to be problematic. For example, if one or more of these new IDCs pose problems for the IRG (*Ideographic Research Group*), such as when performing IDS matching against IRG submission data, the IRG could simply mandate in its P&P (*Principles & Procedures*) that particular IDCs cannot be used in IDSes for IRG submissions. IDS database maintainers do not necessarily have such constraints.

Proposed Code Points, Character Names & Properties

The [Ideographic Description Characters](#) block, which is the most appropriate block for encoding these five new IDCs, has exactly four available code points: **U+2FFC through U+2FFF**. We recommend encoding the first four of these new IDCs using these particular code points. It was suggested during the UTC #172 meeting that **U+31EF**, which is at the very end of the [CJK Strokes](#) block, be recommended as the code point for the fifth IDS.

Therefore, the following are the proposed code points, character names, and property values for the five proposed new IDCs:

```
2FFC;IDEOGRAPHIC DESCRIPTION CHARACTER SURROUND FROM RIGHT;So;0;ON;;;;;N;;;;;
2FFD;IDEOGRAPHIC DESCRIPTION CHARACTER SURROUND FROM LOWER RIGHT;So;0;ON;;;;;N;;;;;
2FFE;IDEOGRAPHIC DESCRIPTION CHARACTER HORIZONTAL REFLECTION;So;0;ON;;;;;N;;;;;
2FFF;IDEOGRAPHIC DESCRIPTION CHARACTER HALF-TURN ROTATION;So;0;ON;;;;;N;;;;;
31EF;IDEOGRAPHIC DESCRIPTION CHARACTER COMPONENT SUBTRACTION;So;0;ON;;;;;N;;;;;
```

New Character Property

The two new unary IDCs will require that a new character property, **IDS_Unary_Operator**, be defined. This new property needs to be reflected in the “CJK” section of Table 7, *Property Index by Scope of Use*, in [Section 5.1, Property Index](#), of UAX #44 as a link to a new entry in the “PropList.txt” section of Table 9, *Property Table*, in [Section 5.3, Property Definitions](#), of the same UAX with Property Type, Property Status, and Property Description fields being identical to those of *IDS_Binary_Operator* and *IDS_Tertiary_Operator*:

Property Type: **B**
Property Status: **N**
Property Description: **Used in Ideographic Description Sequences.**

The following are the proposed changes to the IDC-related lines in the UCD’s *PropList.txt* file, showing changes and new lines in **red**:

```
2FFE..2FFF ; IDS_Unary_Operator # So [2] IDEOGRAPHIC DESCRIPTION CHARACTER
HORIZONTAL REFLECTION..IDEOGRAPHIC DESCRIPTION CHARACTER HALF-TURN ROTATION

2FF0..2FF1 ; IDS_Binary_Operator # So [2] IDEOGRAPHIC DESCRIPTION CHARACTER LEFT
TO RIGHT..IDEOGRAPHIC DESCRIPTION CHARACTER ABOVE TO BELOW
2FF4..2FFD ; IDS_Binary_Operator # So [10] IDEOGRAPHIC DESCRIPTION CHARACTER FULL
SURROUND..IDEOGRAPHIC DESCRIPTION CHARACTER SURROUND FROM LOWER RIGHT
31EF ; IDS_Binary_Operator # So IDEOGRAPHIC DESCRIPTION CHARACTER
COMPONENT SUBTRACTION

2FF2..2FF3 ; IDS_Tertiary_Operator # So [2] IDEOGRAPHIC DESCRIPTION CHARACTER LEFT
TO MIDDLE AND RIGHT..IDEOGRAPHIC DESCRIPTION CHARACTER ABOVE TO MIDDLE AND BELOW
```

The proposed short name for the *IDS_Unary_Operator* property is **IDSU**, and following is the proposed change to the IDC-related lines in the “Binary Properties” section of the UCD’s *PropertyAliases.txt* file, showing new lines in **red**:

```
IDSU ; IDS_Unary_Operator
IDSB ; IDS_Binary_Operator
IDST ; IDS_Tertiary_Operator
```

IDS Grammar

The grammar in Section 18.2, *Ideographic Description Characters*, of the Core Specification should be updated to accommodate unary IDCs and the three new binary IDCs, as follows (additions are shown in **red**):

```
IDS := Ideographic | Radical | CJK_Stroke | Private Use | U+FF1F
| IDS_UnaryOperator IDS
| IDS_BinaryOperator IDS IDS
| IDS_TertiaryOperator IDS IDS IDS
CJK_Stroke := U+31C0 | ... | U+31E3
IDS_UnaryOperator := U+2FFE | U+2FFF
IDS_BinaryOperator := U+2FF0 | U+2FF1 | U+2FF4 | ... | U+2FFD | U+31EF
IDS_TertiaryOperator:= U+2FF2 | U+2FF3
```

TrueType Font

A TrueType font with an open source (OFL) license that provides representative glyphs for all 17 IDCs—12 existing plus five proposed—that map from code points in the Ideographic Description Characters (U+2FF0 through U+2FFF) and CJK Strokes (U+31EF) blocks is attached to this PDF.

That is all.

**ISO/IEC JTC 1/SC 2/WG 2
PROPOSAL SUMMARY FORM TO ACCOMPANY SUBMISSIONS
FOR ADDITIONS TO THE REPERTOIRE OF ISO/IEC 10646¹**

Please fill all the sections A, B and C below.

Please read Principles and Procedures Document (P & P) from <http://std.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://std.dkuug.dk/JTC1/SC2/WG2/docs/summaryform.html>.

See also <http://std.dkuug.dk/JTC1/SC2/WG2/docs/roadmaps.html> for latest **Roadmaps**.

A. Administrative

1. Title:
2. Requester's name:
3. Requester type (Member body/Liaison/Individual contribution):
4. Submission date:
5. Requester's reference (if applicable):
6. Choose one of the following:
- This is a complete proposal:
- (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:
- Name of the existing block:
2. Number of characters in proposal:
3. Proposed category (select one from below - see section 2.2 of P&P document):
- A-Contemporary 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?
- a. If YES, are the names in accordance with the "character naming guidelines" in Annex L of P&P document?
- b. Are the character shapes attached in a legible form suitable for review?
5. Fonts related:
- a. Who will provide the appropriate computerized font to the Project Editor of 10646 for publishing the standard?
- b. Identify the party granting a license for use of the font by the editors (include address, e-mail, ftp-site, etc.):
6. References:
- a. Are references (to other character sets, dictionaries, descriptive texts etc.) provided?
- b. Are published examples of use (such as samples from newspapers, magazines, or other sources) of proposed characters attached?
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)?

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.

¹ Form number: N4502-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? If YES explain	Yes See proposal
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.)? If YES, with whom? If YES, available relevant documents:	Yes See proposal See proposal
3. Information on the user community for the proposed characters (for example: size, demographics, information technology use, or publishing use) is included? Reference:	N/A
4. The context of use for the proposed characters (type of use; common or rare) Reference:	Common See proposal
5. Are the proposed characters in current use by the user community? If YES, where? Reference:	N/A
6. After giving due considerations to the principles in the P&P document must the proposed characters be entirely in the BMP? If YES, is a rationale provided? If YES, reference:	Yes Yes See proposal
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? If YES, is a rationale for its inclusion provided? If YES, reference:	No
9. Can any of the proposed characters be encoded using a composed character sequence of either existing characters or other proposed characters? If YES, is a rationale for its inclusion provided? If YES, reference:	No
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? If YES, is a rationale for its inclusion provided? If YES, reference:	No
11. Does the proposal include use of combining characters and/or use of composite sequences? 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:	No
12. Does the proposal contain characters with any special properties such as control function or similar semantics? If YES, describe in detail (include attachment if necessary)	No
13. Does the proposal contain any Ideographic compatibility characters? If YES, are the equivalent corresponding unified ideographic characters identified? If YES, reference:	No