

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

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

Title: Preliminary proposal to add a new provisional *kIDS* property (Unihan)

Authors: Ken Lunde & John H. Jenkins

Status: Member Body Contribution

Action: For consideration by the IRG and UTC

Date: 2021-08-11 (revised)

The purpose of this document, which is a revised preliminary proposal to add a new provisional Unihan database property, *kIDS*, is threefold:

1. Outline the standardization timeline
2. Identify any barriers early on in the process through constructive and meaningful feedback from both the UTC and IRG
3. Solicit help in collecting ideograph components for use in IDSeS

IDSeS is an abbreviation for *Ideographic Description Sequences*, which is extensively documented in [Section 18.2](#), *Ideographic Description Characters*, of the Core Specification of the Unicode Standard.

The standardization timeline has two targets, both of which are subject to change: Unicode Version 15.0 (2022) and Unicode Version 16.0 (2023).

Unicode Version 15.0 (2022)

The targets for Unicode Version 15.0 are to:

1. Encode up to five new Ideographic Description Characters (IDCs)
2. Encode a modest number of ideograph components as new CJK Unified Ideographs for use in IDSeS

This then sets the stage for adding the provisional *kIDS* property in the subsequent version of the Unicode Standard.

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:

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 ONE HUNDRED EIGHTY DEGREE ROTATION
	Binary	IDEOGRAPHIC DESCRIPTION CHARACTER STROKE SUBTRACTION

The fifth new IDC that is being considered whose recommended name is IDEOGRAPHIC DESCRIPTION CHARACTER STROKE SUBTRACTION is binary, and is therefore followed by two components:

1. An ideograph component
2. A CJK stroke from the [CJK Strokes](#) block that is omitted from the ideograph component, or multiple CJK strokes if an IDC is used

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 represented as 其ノ and 其ノ, respectively.
- The IDSeS for U+2CEBB 豸 is also difficult to represent with existing ideograph components, but could be represented as 豸ノ.

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.

The two new unary IDCs will require that a new character property, *IDS_Unary_Operator*, be defined, and that the grammar in Section 18.2, *Ideographic Description Characters*, of the Core Specification be updated to accommodate unary IDCs and the three new binary IDCs, such as the following (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+2FE0 | U+2FF0 | U+2FF1 | U+2FF4 | ... | U+2FFB | U+2FFC
    | U+2FFD
IDS_TertiaryOperator := U+2FF2 | U+2FF3
```

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. The formal proposal will recommend encoding four of these new IDCs using these particular code

points. There is currently an unassigned block of 16 code points immediately before the *Ideographic Description Characters* block, specifically the range U+2FE0 through U+2FEF, which could be used to encode the fifth IDC.

New Ideograph Components

A non-trivial number of CJK Unified Ideographs include components that cannot easily be represented in IDSes, and a modest number of additional ideograph components—encoded as new CJK Unified Ideographs—would serve to improve the IDSes of such ideographs.

Examples of candidate ideograph components can be found in [L2/21-134](#), *Collections of ideograph components for use in IDSes*, which is a repertoire of 119 ideograph components that was prepared and submitted as a formal response to the original version of this preliminary proposal.

All existing ideograph components from a variety of sources will be collected, studied, and cataloged, which will serve as the basis for the formal proposal. A non-trivial number of ideograph components are already encoded as CJK Unified Ideographs, meaning that there is a precedent to continue to treat ideograph components as CJK Unified Ideographs. The first step toward encoding the new ideograph components will be to add them to [UAX #45](#), *U-Source Ideographs*, meaning that each will be assigned a U-Source source reference with the usual “UTC-” prefix.

Given the aggressive timeline, we hereby solicit help from experts in collecting, studying, and cataloging ideograph components. In addition, a new CJK Unified Ideographs extension block that includes sufficient unused code points for encoding additional ideograph components in the future will be proposed. Our recommendation is to use U+3FF00 through U+3FFFD for this new block, which provides 254 code points, and to use *CJK Unified Ideographs Components* as the block name to distinguish it from the CJK Unified Ideographs extension blocks.

Unicode Version 16.0 (2023)

With up to five new IDCs and a modest number of ideograph components encoded in Unicode Version 15.0, the target for Unicode Version 16.0 is to add the provisional *kIDS* property to the UniHan database, based on [IDS.TXT](#). This IDS database, which is developed and maintained by Andrew West, has no copyright and is not encumbered by a license.

Delaying the addition of this new provisional property until Unicode Version 16.0 gives sufficient time for this IDS database to incorporate the new IDCs and new ideograph components, and also to make continued refinements to other IDSes. This should result in more robust property data.

That is all.