

Universal Multiple-Octet Coded Character Set UCS

ISO/IEC JTC1/SC2/WG2 IRG N [1183](#)

Date: 2005-12-26

Source:	Japan
Title:	Guidelines on IDS Decomposition.
Status :	
Actions required	Review by IRG Editors for discussion at IRG meeting No. 25.
Distribution:	IRG Members and Ideographic Experts
Medium :	Electronic

1. Backgrounds

This document is the revised edition of the IRG N1153, "Guidelines on IDS Decomposition." The main difference with the IRG N1153 is that this document loosens the restrictions on the IDS decomposition. The technical contents remain the same.

The authors believe that the use of IDS greatly helps the standardization works of CJK UNIFIED IDEOGRAPHS family of characters, especially during the review process. With IDS, we can find *similar ideographs* much more easily than ever, helped by a small program.

2. Principles

The principles behind the guideline are summarized as follows:

2.1. Minimal division.

We should not divide too much. If we need further division, a program can easily generate such deep division forms, because we only use existing (already standardized) ideographs with their own IDS division. However, it is merely the recommendation, and this principle does not enforce the user to minimize the decomposition.

2.2. Concentration on visual shapes.

We should not stick to the ideographs meaning, origin, or the traditional classification/separation of components. Remember that our purpose of use of IDS is only to review the proposed ideographs. If we rely on, for example, the knowledge about the radical, IDS division by a person who doesn't know the correct radical may make a *wrong* IDS division.

By ignoring the detailed knowledge on ideograph's meaning, origin, etc., there are more chance that the IDS assigned by a person is same to those by another, regardless of the difference of knowledge on that particular ideograph.

2.3. Giving up early.

Some ideographs have a unique shape and/or structure and it is not easy to find an IDS for them. That's OK. Let them leave alone. We don't need a complete collection.

Again, we are just reviewing. We are not compiling a dictionary. As long as a number of such exceptional cases are relatively small, they have no repercussion with the entire review process.

2.4. Restricted use of *surrounding* and *overlapping* IDCs.

The use of surrounding or overlapping IDCs is sometimes ambiguous and may fail to detect the duplicate character algorithmically. This principle is to remove this difficulty.

2.5. Generousness on minor differences

Don't try to represent details of the shapes of an ideograph. Ignore minor differences. We have a set of unification rules and if the difference is important (for the unification rules), we can consider so through the eye-to-eye review after the IDS based matching. On the other hand, if the IDS is constructed under a draconian policy, two shapes to be unified may have a totally different IDS and we may fail to find them duplicate.

3. Definitions

IDC (Ideographic description character): One of 12 UCS characters whose code points are in range 2FF0 to 2FFB. See Annex F.3 of ISO/IEC 10646 for details.

CDC (Character description component): A UCS character that is included either in CJK UNIFIED IDEOGRAPHS, in CJK UNIFIED IDEOGRAPHS EXTENSION A, in CJK UNIFIED IDEOGRAPHS EXTENSION B, in KANGXI RADICALS, in CJK RADICALS SUPPLEMENT, or in CJK COMPATIBILITY IDEOGRAPHS. In other words, CDC is a DC that consists of just one UCS character.

SDC (Sequence description component): An IDS that is used as a DC in other IDSs. In other words, SDC is a DC that consists of a sequence of an IDC and following DCs.

DC: either CDC or SDC.

4. The recommended procedure for Constructing IDS

Following procedures are only for the *recommendation*, to keep the textual representation of the IDS simple and consistent as possible.

[1] See if the ideograph has a structure that two same components *pinch* another component. If so, take the division. i.e.,

[1-1] If the ideograph can be divided into three parts using 2FF2 (𠄒), where the *left-most* and *right-most* components *are the same CDC*, divide so. (The middle may be CDC or SDC in this case.)

Example:

嫩 → 𠄒女男女 (rather than 𠄒媯女)

弼 → 𠄒弓百弓 (rather than 𠄒弼弓)

[1-2] Otherwise, if an ideograph can be divided into three parts using 2FF3 (𠄓), where the *top* and *bottom* components *are the same CDC*, divide so. (The middle DC may be CDC or SDC in this case.)

Example:

器 → 𠄓𠄒犬𠄒 (rather than 𠄓哭𠄒)

[2] If the [1] above doesn't apply, see if the given ideograph is divided into two parts, and both parts are coded ideographs (CDCs). i.e.,

[2-1] If an ideograph can be divided into two parts using 2FF0 (𠄔), where the both left and right components are (not necessarily same) CDCs, divide so.

Examples:

雖 → 𠄔虽隹 (not 𠄔唯虫)

[2-2] Otherwise, if an ideograph can be divided into two parts using 2FF1 (𠄕), where the both top and bottom components are (not necessarily same) CDCs, divide so.

Examples:

笈 → 𠄕竹及

[2-3] Otherwise, if an ideograph can be divided into two parts using 2FF4 (𠄖), where the both outer and inner components are (not necessarily same) CDCs, divide so.

Examples:

袁 → 冫 口 袁

[2-4] Otherwise, if an ideograph can be divided into two parts using 2FF5(冫), where the both outer and inner components are (not necessarily same) CDCs, divide so.

Examples:

間 → 冂 門 日

[2-5] Otherwise, if an ideograph can be divided into two parts using 2FF6(冂), where the both outer and inner components are (not necessarily same) CDCs, divide so.

Examples:

𠂇 → 冂 凵 𠂇

[2-6] Otherwise, if an ideograph can be divided into two parts using 2FF7(冂), where the both outer and inner components are (not necessarily same) CDCs, divide so.

Examples:

匣 → 冂 匚 甲

[2-7] Otherwise, if an ideograph can be divided into two parts using 2FF8(冂), where the both outer and inner components are (not necessarily same) CDCs, divide so.

Examples:

厘 → 冂 厂 里

[2-8] Otherwise, if an ideograph can be divided into two parts using 2FF9(冂), where the both outer and inner components are (not necessarily same) CDCs, divide so.

Examples:

勾 → 冂 勺 厶

[2-9] Otherwise, if an ideograph can be divided into two parts using 2FFA(冂), where the both outer and inner components are (not necessarily same) CDCs, divide so.

Examples:

赶 → 冂 走 干

[2-10] Otherwise, if an ideograph can be divided into two parts using 2FFB(𠄎), where the both outer and inner components are (not necessarily same) CDCs, divide so.

Examples:

幽 → 𠄎山纟

Note the explicitly given priority of IDCs. If an ideograph can be divided into two parts either horizontally or vertically, we always divide it hirozontally (even if the division contradicts the ideographs origin!)

Examples:

众 → 𠄎从从 (rather than 𠄎欠欠)

[3] If the [1] and [2] above still don't apply, see if the given ideograph is divided into three parts, and all parts are coded ideographs (CDCs), take it. i.e.,

[3-1] If the ideograph can be divided into three parts using 2FF2, where all left, middle, and right components are CDCs, divide so.

Examples:

徹 → 𠄎彳育攴

[3-2] Otherwise, if an ideograph can be divided into three parts using 2FF3, where the both top and bottom components are CDCs, divide so.

Examples:

享 → 𠄎亠口子

[4] If the [1], [2], and [3] don't apply, we try to divide the ideograph using *two* IDCs at the same time.

Examples:

幹 → 𠄎卓𠄎人干

穎 → 𠄎𠄎匕禾頁

薛 → 𠄎⁺⁺𠄎自辛

憩 → 𠄎𠄎舌自心

圀 → 𠄎口日𠄎方
 岡 → 𠄎冂日𠄎山
 囪 → 𠄎凵日人二
 匡 → 𠄎匚日山王
 厚 → 𠄎厂日日子
 貳 → 𠄎弋日二貝
 邃 → 𠄎辵日穴豕

Note that *it is not recommended* to use SDC as the first DC of the IDC, except if IDC is either 2FF0 or 2FF1.

[5] If the [1], [2], [3] and [4] don't apply, we now try IDS with *three* IDCs. Again, it is not recommended to use SDC as the first DC of the IDC, except if IDC is either 2FF0 or 2FF1.

[6] If the ideograph is still not divided into an IDS, give up.

Examples.

Examples:

勝 → 𠄎月券 (not 𠄎朕力)
 桂 → 𠄎木圭 (prefer to 𠄎木日土土)
 土 → 𠄎土、 (not 𠄎土、)
 土 → 𠄎土、 (not 𠄎土、)
 傘 → 日人𠄎十𠄎
 傾 → 𠄎亻頃 (prefer to 𠄎化頁)
 膳 → 𠄎月飡 (not 𠄎朕言)
 京 → 日一人口小
 雝 → 𠄎隹隹隹
 繇 → 𠄎糸言糸 (prefer to 𠄎糸諄)

巖 → 日 丿 日 厂 敢 (not 日 日 丿 厂 敢)
彦 → 日 文 日 厂 彡 (not 日 日 文 厂 彡)
県 → 日 日 目 小 (not 日 日 目 小 目)
虎 → 日 卜 日 厂 斤 (not 日 日 卜 厂 斤)

4. Sample file.

The sample IDS data (ids.txt) attached with this document covers most of BMP and SIP characters. Referencing this data might be useful for constructing the IDS.

If you can't find the appropriate DC of the target character, think of any other character you know which shares the common DC part with the target character. Try search such character in this sample file and see how such character is decomposed.