Universal Multiple-Octet Coded Character Set UCS

ISO/IEC JTC1/SC2/WG2 IRG N 1153
Date: 2005-11-21

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

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.

IDS database for already standardized ideographs, i.e., those for all ideographs in URO, Ext-A, and Ext-B already exist, although it may contain some errors. However, such errors make no serious problem. (We anyway need eye-to-eye review). Just consider the IDS-based report the program generates a suggestion or candidates. Even if the accuracy of the IDS based review is 90%, it greatly reduces the required workloads to look up duplicates from nearly a hundred thousands of ideographs.

An ideograph often can be divided into two or more different IDSs. A program to find duplicate makes its best effort to find same or similar ideographs even if they have different IDS. (See the document IRG N1154 for the current algorithm the program uses and how smartly it can recognize different IDSs represent same ideograph.) However, some simple guideline about the way of dividing will gain the accuracy.

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. On the contrary, if we starts with the maximum division, its is not easy to algorithmically re-structuring the original shapes.

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 ideograph have a unique shape and/or structure and not easy to find an IDS. 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 ideographs. 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 procedure for Constructing IDC

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

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

Example:

嬲 → ⿲女男女 (rather than ⿰娚女)

弼 → ⿲弓百弓 (rather than ⿰弼弓)

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

Example:

器 → ⿳口犬口 (rather than ⿳哭口)

[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 (⿰), where the both left and right components are (not necessarily same) CDCs, divide so.

Examples:

雖 → ⿰虽隹 (not ⿰唯虫)

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

Examples:

笈 → ⿱竹及

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

Examples:

園 → ⿴口袁

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

Examples:

間 → ⿵門日

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

Examples:

凶 → ⿶凵乂

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

Examples:

匣 → ⿷匚甲

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

Examples:

厘 → ⿸厂里

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

Examples:

勾 → ⿹勹厶

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

Examples:

赶 → ⿺走干

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

Examples:

幽 → ⿻山幺

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:

灸 → ⿱从从 (rather than ⿰夂夂)

[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 → ⿲彳育攵

[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 → ⿱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. During this rule [4], we assume an SDC is an IDS for a component of the ideograph under consideration that if the component was an ideograph and applying the rules [1] through [3] above for it would cause the IDS.

[4-1] If an ideograph can be divided into two parts using 2FF0 (⿰), where the left component is a CDC, and the right component is an SDC, divide so.

Examples:

幹 → ⿰卓𠂉人干

[4-2] Otherwise, if an ideograph can be divided into two parts using 2FF0 (⿰), where the right component is a CDC, and the left component is an SDC, divide so.

Examples:

穎 → ⿰禾𠂉匕禾頁

[4-3] Otherwise, if an ideograph can be divided into two parts using 2FF1(⿱), where the top component is a CDC, and the bottom component is an SDC, divide so.

Examples:

薛 → ⿱艹⿰𠂉自辛

[4-4] Otherwise, if an ideograph can be divided into two parts using 2FF1(⿱), where the bottom component is a CDC, and the top component is an SDC, divide so.

Examples:

憩 → ⿱舌⿰𠂉自心

[4-5] Otherwise, if an ideograph can be divided into two parts using 2FF4(⿴), where the outer component is a CDC, and the inner component is an SDC, divide so.

Examples:

圀 → ⿴囗⿱罒方

[4-6] Otherwise, if an ideograph can be divided into two parts using 2FF5(⿵), where the outer component is a CDC, and the inner component is an SDC, divide so.

Examples:

岡 → ⿵冂⿱罒山

[4-7] Otherwise, if an ideograph can be divided into two parts using 2FF6(⿶), where the outer component is a CDC, and the inner component is an SDC, divide so.

Examples:

㠯 → ⿴凵⿱人二

[4-8] Otherwise, if an ideograph can be divided into two parts using 2FF7(⿲), where the outer component is a CDC, and the inner component is an SDC, divide so.

Examples:

匰 → ⿲匚山王

[4-9] Otherwise, if an ideograph can be divided into two parts using 2FF8(⿳), where the outer component is a CDC, and the inner component is an SDC, divide so.

Examples:

厚 → ⿳厂日子

[4-10] Otherwise, if an ideograph can be divided into two parts using 2FF9 (⿴), where the outer component is a CDC, and the inner component is an SDC, divide so.

Examples:

貳→ ⿴弋二貝

[4-11] Otherwise, if an ideograph can be divided into two parts using 2FFA (⿵), where the outer component is a CDC, and the inner component is an SDC, divide so.

Examples:

邃 → ⿵辶穴豕

Note again on the priority. Also note that, except for the cases for 2FF0 and 2FF1, *we don't allow SDC as the first DC to the IDC.*

[5] If the [1], [2], [3] and [4] don't apply, we now try IDS with three IDCs. Just repeat the [4] with consideration that the SDCs explained in [4] can now be the IDS. (Exact conditions [5-1] to [5-11] are omitted, since they are exactly the same sentences as [4-1] to [4-11].)

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

Examples.

Examples:

勝 → ⿰月券 （not ⿰朕力）
桂 → ⿰木圭 (prefer to ⿰木⿱土土)
圭 → ⿱土丶 (not ⿻土丶)
圭 → ⿱土丶 (not ⿻土丶)
傘 → ⿱人⿻十灬
傾 → ⿰亻頃（prefer to ⿰化頁）
謄 → ⿰月誊 （not ⿰朕言）
京 → ⿳亠口小
雥 → ⿳隹隹隹
纞 → ⿳糸言糸 (prefer to ⿰糸誶)
厳 → ⿱𠂉⿰厂敢 (not ⿰⿱𠂉厂敢)
彦 → ⿱文⿰厂彡 (not⿰⿱文厂彡)
県 → ⿱⿺乚目小 (not ⿺ ⿱乚小目)
虎 → ⿱卜⿰厂斤 (not ⿰⿱卜厂斤)

4. Sample file.

The sample IDS data (ids.txt) attached with this document covers most of BMP and SIP characters. They might be useful on constituting the IDS of any character, as the most efficient way to constitute the IDS is to refer to the IDS of the similar character and copy (the part of) them.

If you can't find the appropriate DC of the target character, think of any other character which shares the common DC part, then search and see how that character is constituted in the sample file.