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 IDEOGRAPH 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 IDEOGRAPHICS, in CJK UNIFIED IDEOGRAPHICS EXTENSION A, in CJK UNIFIED IDEOGRAPHICS EXTENSION B, in KANGXI RADICALS, in CJK RADICALS SUPPLEMENT, or in CJK COMPATIBILITY IDEOGRAPHICS. 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 (⿲), 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:

嫐 → ⽱⼥男女 (rather than ⽴娚⼥)
弼 → ⽱弓百弓 (rather than ⽴SGlobal
ds弓)

[1-2] Otherwise, if an ideograph can be divided into three parts using 2FF3 (⿳), 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 → ⽱吅犬吅 (rather than ⽴строенч)

[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:

雖然 → ⽱虽隹 (not ⽴唯虫)

[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 → ⽱竹及

[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 → □□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 → □□㐅

[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 → □勺忄

[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 → 二山绞

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

Examples:  
众 → 二从从 (rather than 仌欠欠)

[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 → 𠇋彳育攵

[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 → 仌亠口子

[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 → 仌卓仌々干  
穎 → 仌仌匕禾頁  
薛 → 仌々仌自辛  
憩 → 仌仌舌自心
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 → 月劵 (not 朕力)
桂 → 木圭 (prefer to 木土土)
土 → 土、 (not 土、)
土 → 土、 (not 土、)
傘 → 十
傾 → 亻頃 (prefer to 化頁)
膳 → 月膳 (not 朕言)
京 → 一口小
瓞 → 佳佳佳
繽 → 網言綱 (prefer to 網詠)

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.