

Universal Multiple-Octet Coded Character Set UCS

ISO/IEC JTC1/SC2/WG2 IRG N [1154](#)

Date: 2005-11-21

Source:	Taichi Kawabata
Title:	Algorithm for Identifying the Duplicate Ideograph Characters by the IDS.
Status :	Informative
Actions required	Review by IRG Editors for discussion at IRG meeting No. 25.
Distribution:	IRG Members and Ideographic Experts
Medium :	Electronic

1. Introduction

In this paper, the author will present the algorithm to “normalize” the IDS sequence. The IDS itself is usually ambiguous because (1) the single character may be described by the IDS in the multiple ways, and (2) the single DC may be described by the several UCS code points (for example, the character “示” can be described as U+793A, U+2EAC (Radical Supplement), or U+2F70 (Kang-Xi Radical)), and (3) sometimes the different ideographs can be considered the variants when they appear in the IDS.

Procedures described here will solve these problems by *recursively flattening* the IDS and replacing recognized identical or variant component characters to single unified character to build the normalized form, for just the comparison purpose.

Remember that, the replacements are sometimes *inappropriate*, so this merely helps identifying the *possible* duplicate characters. Also, it *does not assure* that all duplicates would be found by this method. This merely helps identifying the possible duplicates.

2. Definition

Flatten IDS: *Flatten IDS* is the special form of the IDS where horizontally (or vertically) aligned IDCs are equally put together in the form of “ (= ...)”, or “ (= ...)”.

Example:

亭 (U+02180) can be described by the IDS in various forms, such as “
→  →   +”, “ →  →   +” or
“ →  →   +”. In the Flatten IDS, 亭 can be

uniformly described as “𠄎(一𠄎一𠄎一𠄎)”

3. Algorithm

- (a) Prepare the IDS data of the candidate characters, and “ids.txt” data file which covers IDS data of the most of UCS characters.
- (b) **Flatten** the IDS of the candidate characters *horizontally* or *vertically* according to the data described in “ids.txt”.

Horizontal Flattening is that, when the IDC of the IDS is either 𠄎, 𠄎, 𠄎, or 𠄎, then the DCs of the character will be equally put together in the form of “𠄎(⋯)” . If any of the DCs of such IDS are also the IDS with the IDC of 𠄎, 𠄎, 𠄎, or 𠄎, then they will also be expanded and recursively filed into such flatten IDS.

Vertical Flattening is that, when IDC of the IDS is either 𠄎, 𠄎, 𠄎, 𠄎 or 𠄎, then the the DCs of the IDS will be equally put together in the form of “𠄎(⋯)” . If any of the DCs of the IDS are also the IDS with the IDC of 𠄎, 𠄎, 𠄎, 𠄎 or 𠄎, then they will also be expanded and recursively filed into such flatten IDS.

Example:

湊 (𠄎 彳 致)	→ 𠄎(彳 致)	
	→ 𠄎(彳 至 攴)	(Horizontal Flattening)
	→ 𠄎(彳 𠄎(𠄎土) 攴)	
	→ 𠄎(彳 𠄎(一𠄎土) 攴)	(Vertical Flattening)
樂 (𠄎 丩 樂)	→ 𠄎(丩 樂)	(樂 → 𠄎𠄎 幺 白 幺 木)
	→ 𠄎(丩 𠄎(𠄎(幺 白 幺) 木)	
纛	→ 𠄎(糸 𠄎(𠄎𠄎𠄎 口 𠄎(土 几) 一 几) 糸)	
燿	→ 𠄎(火 𠄎(日 廿 一 廿 一 一))	
曩	→ 𠄎(𠄎 一 由 圭 攴)	

- (c) After each step of horizontal or vertical flattening is completed, replace each DCs of the flatten IDS with the possible variant or unifying component characters. The list of the variant characters will be prepared based on the Annex S or Normalization Data (UAX #15) (See appendix below).

- (d) If horizontal flattening is completed, try vertical flattening for each DCs. Likewise, if vertical flattening is completed, then try horizontal flattening for each DCs.
- (e) If the flattening and replacing process can not be done any more in any DCs, then it will be called “normalized IDS” for convenience.

4. Identification of the Unification Possibility

If the two characters have the same “normalized IDS”, then there is a *possibility* that these two characters might be unified. If they are similar enough, the one who proposes such character may prepare the evidence (meaning, pronunciation, etc.) that shows two characters are distinct.

5. Examples of the Flatten and Unified IDS

Highlighted are the pairs detected as being the candidate for the unification.

𠂔 → 𠂔(月 𠂔(人三))
 𠂔 → 𠂔(月 𠂔(人二ム))
 𠂔 → 𠂔(月 𠂔(八一⁺⁺))
 𠂔 → 𠂔(月 𠂔(八一⁺⁺))
 𠂔 → 𠂔(月 𠂔(八𠂔))
 𠂔 → 𠂔(月 𠂔(八人八人))
 𠂔 → 𠂔(月 𠂔(八十))
 𠂔 → 𠂔(月 𠂔(八ム心))
 𠂔 → 𠂔(月 𠂔(八ム羽))
 𠂔 → 𠂔(月 𠂔(八又))
 𠂔 → 𠂔(月 𠂔(八口八))
 𠂔 → 𠂔(月 𠂔(八口八))
 𠂔 → 𠂔(月 𠂔(八口八)𠂔(門八大))
 𠂔 → 𠂔(月 𠂔(八夫己))
 𠂔 → 𠂔(月 𠂔(八夫目))

殺 → 𠂔(𠂔(×木)𠂔(八又))
 𠂔 → 𠂔(𠂔(×木)𠂔(弋工))
 𠂔 → 𠂔(𠂔(×木)𠂔(弋工))
 網 → 𠂔(𠂔(×木)𠂔(門人))
 網 → 𠂔(𠂔(×木)𠂔(門木))
 𠂔 → 𠂔(𠂔(×木)𠂔(𠂔))
 𠂔 → 𠂔(𠂔(×木)𠂔(𠂔))
 𠂔 → 𠂔(𠂔(×木)𠂔(支))

𠂔 → 𠂔(𠂔(×木)𠂔(辛))
 𠂔 → 𠂔(𠂔(×木)𠂔(辛))
 𠂔 → 𠂔(𠂔(×木)𠂔(鬼))
 𠂔 → 𠂔(𠂔(×木)𠂔(鬼))
 𠂔 → 𠂔(𠂔(×木)𠂔(鳥))

𠂔 → 𠂔(𠂔(一八)𠂔(于))
 𠂔 → 𠂔(𠂔(一八)𠂔(人))
 𠂔 → 𠂔(𠂔(一八)𠂔(𠂔))
 𠂔 → 𠂔(𠂔(一八)𠂔(𠂔))
 𠂔 → 𠂔(𠂔(一八)𠂔(力))
 𠂔 → 𠂔(𠂔(一八)𠂔(千))
 𠂔 → 𠂔(𠂔(一八)𠂔(ム))
 𠂔 → 𠂔(𠂔(一八)𠂔(又))
 𠂔 → 𠂔(𠂔(一八)𠂔(又))
 𠂔 → 𠂔(𠂔(一八)𠂔(夕))

滴 → 𠂔(𠂔(𠂔)𠂔(一八門十口))
 滌 → 𠂔(𠂔(𠂔)𠂔(一八門巾))
 滌 → 𠂔(𠂔(𠂔)𠂔(一八門巾口))
 滌 → 𠂔(𠂔(𠂔)𠂔(一八門方))
 滌 → 𠂔(𠂔(𠂔)𠂔(一八門方))
 滌 → 𠂔(𠂔(𠂔)𠂔(一八門羊))
 滌 → 𠂔(𠂔(𠂔)𠂔(一八門))

吳(U+5433)	吳(U+5434)	吳(U+5449)
告(U+543F)	告(U+544A)	
噯(U+568F)	噯(U+5694)	
圀(U+56EF)	圀(U+56FD)	
圓(U+570E)	圓(U+5713)	
圖(U+5716)	圖(U+5717)	
𠂔(U+5759)	𠂔(U+5DE0)	
埒(U+57D2)	埒(U+57D3)	
壽(U+58FD)	壽(U+5900)	
𩇛(U+5910)	𩇛(U+657B)	
本(U+5932)	本(U+672C)	
奧(U+5965)	奧(U+5967)	
姬(U+59EB)	姬(U+59EC)	
婕(U+5A55)	婕(U+5AAB)	
孳(U+5B73)	孳(U+5B76)	
寬(U+5BDB)	寬(U+5BEC)	
寧(U+5BDC)	寧(U+5BE7)	
專(U+5C02)	專(U+5C08)	
尔(U+5C13)	尔(U+5C14)	
尙(U+5C19)	尙(U+5C1A)	
尫(U+5C2A)	尫(U+5C2B)	
𪔐(U+5C36)	𪔐(U+5C37)	
帶(U+5E2F)	帶(U+5E36)	
并(U+5E76)	并(U+5E77)	
強(U+5F37)	強(U+5F3A)	
𠂇(U+5F50)	𠂇(U+5F51)	
彥(U+5F65)	彥(U+5F66)	
德(U+5FB3)	德(U+5FB7)	
徵(U+5FB4)	徵(U+5FB5)	
惠(U+6075)	惠(U+60E0)	
戶(U+6236)	戶(U+6237)	戶(U+6238)
戾(U+623B)	戾(U+623E)	
拋(U+629B)	拋(U+62CB)	
拔(U+629C)	拔(U+62D4)	
挿(U+633F)	挿(U+63D2)	挿(U+63F7)
捏(U+634F)	捏(U+63D1)	
搜(U+635C)	搜(U+641C)	
搖(U+63FA)	搖(U+6416)	搖(U+6447)
擊(U+6483)	擊(U+64CA)	
教(U+654E)	教(U+6559)	
既(U+65E2)	既(U+65E3)	
昂(U+6602)	昂(U+663B)	
曾(U+66FD)	曾(U+66FE)	
𣎵(U+67B4)	𣎵(U+67FA)	
查(U+67E5)	查(U+67FB)	
樣(U+69D8)	樣(U+6A23)	
步(U+6B65)	步(U+6B69)	
歲(U+6B72)	歲(U+6B73)	
歿(U+6B7F)	歿(U+6B81)	
𣪠(U+6BBB)	𣪠(U+6BBC)	
毀(U+6BC0)	毀(U+6BC1)	
每(U+6BCE)	每(U+6BCF)	
汚(U+6C5A)	汚(U+6C61)	
沒(U+6C92)	沒(U+6CA1)	
潛(U+6F5B)	潛(U+6FF3)	

瀨(U+7028)	瀨(U+702C)
為(U+70BA)	爲(U+7232)
𩇛(U+712D)	𩇛(U+7162)
熙(U+7155)	熙(U+7199)
狀(U+72B6)	狀(U+72C0)
瑤(U+7464)	瑤(U+7476)
產(U+7522)	產(U+7523)
瘦(U+75E9)	瘦(U+7626)
眞(U+771E)	真(U+771F)
𩇛(U+773E)	衆(U+8846)
禿(U+79BF)	禿(U+79C3)
篡(U+7BE1)	篡(U+7C12)
粵(U+7CA4)	粵(U+7CB5)
絕(U+7D55)	絕(U+7D76)
羹(U+7FAE)	羹(U+7FB9)
𩇛(U+8203)	𩇛(U+8204)
舍(U+820D)	舍(U+820E)
蓓(U+83D1)	蓓(U+8458)
薰(U+85AB)	薰(U+85B0)
虛(U+865A)	虛(U+865B)
衛(U+885B)	衛(U+885E)
袞(U+886E)	袞(U+889E)
謠(U+8B20)	謠(U+8B21)
走(U+8D70)	𨇗(U+8D71)
輜(U+8F1C)	輜(U+8F3A)
达(U+8FBE)	达(U+8FD6)
遙(U+9059)	遙(U+9065)
郎(U+90CE)	郎(U+90DE)
鄉(U+90F7)	鄉(U+9109)
陞(U+9667)	陞(U+9689)
青(U+9751)	青(U+9752)
馱(U+99B1)	馱(U+99C4)
𩇛(U+9AA9)	𩇛(U+9AAB)
高(U+9AD8)	高(U+9AD9)
髮(U+9AEA)	髮(U+9AEE)
鬪(U+9B2C)	鬪(U+9B2D)
鳳(U+9CEF)	鳳(U+9CF3)
麪(U+9EAA)	麪(U+9EAB)
黃(U+9EC3)	黃(U+9EC4)
黑(U+9ED1)	黑(U+9ED2)

鄉(U+9115)

Miscellaneous

𩇛 (U+4E2C)
 蔣, 狀-狀, 裝-裝, 醬-醬)
 开(U+5F00)
 鉞)
 免(U+4FDB)
 皐(U+7690)
 𩇛(U+6740)
 東(U+6771)

 亏 (U+4E8F)
 产 (U+4EA7)
 茲 (U+8332)

𩇛 (U+723F)
 𩇛(U+5E75)

 兔(U+5154)
 皐(U+81EF)
 𩇛(U+233C2)
 東(U+67EC)

 亏 (U+4E90)
 𩇛文厂
 茲 (U+5179)

(妝-妝, 壯-壯, 寢-寢, 將-將, 莊-莊, 蔣-
 (妍-妍, 研-研, 𩇛-𩇛, 猊-猊, 邢-邢, 鉞-
 (嫵-嫵, 晚-晚)
 (翱-翱, 皐-皐 → <白, 自>)
 (刹-刹, 𩇛-𩇛, 檄-檄)
 (諫-諫, 鍊-鍊, 鵝-鵝)

 (汚-汚)
 (產-產, 彦-彦, 顏-顏)
 (萼-萼 → <卅, 𩇛>)

二 (U+4E8C)	𠂇 (U+51AB)		(勻-勻)
几 (U+51E0)	𠂇 (U+5342)		(𠂇-𠂇)
𠂇 (U+5342)	凡 (U+51E1)		([𠂇]CJ)
叟 (U+53DF)	𠂇由又		(搜-搜, 瘦-瘦)
另 (U+53E6)	另 (U+53E7)		(𠂇-𠂇)
另 (U+53E6)	另 (U+20AB0)		(別-別)
員 (U+54E1)	員 (U+8C9F)		(圓-圓)
孝 (U+5B5D)	孝 (U+21949)		(教-教)
𠂇 (U+5BFD)	𠂇𠂇寸		(埤-埤)
小 (U+50CF)	𠂇 (U+2E8C)		(尙-尙)
𠂇 (U+758C)	𠂇卜𠂇		(嫖-嫖)
𠂇 (U+7590)	𠂇 (U+24D21)		(噓-噓)
𠂇 (U+7F95)	𠂇 (U+23D0E)		(樣-樣)
臣 (U+81E3)	臣 (U+268DD)	臣 (U+268DE)	(姬-姬, 熙-熙)
𠂇 (U+81FF)	𠂇 (U+2218D)		(挿-挿-挿)
色 (U+8272)	𠂇刀巴		(絕-絕)
𠂇 (U+226F3)	𠂇十𠂇心		(德-德)
刀 (U+5200)	𠂇 (U+2E88)		(絕-絕)
三 (U+4E09)	𠂇 (U+5F61)		(叁-叁, 參-參)
𠂇 (U+4E0F)	𠂇 (U+4E10)		(麪-麪)
业 (U+4E1A)	𠂇 (CDP-8D6B)		(噓-噓, 戲-戲, 虚-虚)
八 (U+516B)	𠂇 (U+4E37)		(兌-兌, 悅-悅, 挽-挽, 斂-斂, 稅-稅, 浼-)
浼, 脫-脫, 蛻-蛻, 說-說, 銳-銳, 閱-閱)			
九 (U+4E5D)	丸 (U+4E38)		(飢-飢)
八 (U+516B)	儿 (U+513F)	几 (U+51E0)	(秃-秃, 頽-頽, 傻, 爰, 鏤, 「船」)
兀 (U+5140)	尢 (U+5C22)	(允 U+5C23)	(尫-尫, 拋-拋, 廬-廬)
公 (U+516C)	台 (U+3563)		(兇-兇, 滾-滾, 袞-袞)
𠂇 (U+5202)	𠂇 (U+5DDC)		(倫-倫, 俞-俞)
𠂇 (U+5204)	𠂇 (U+5205)		(𠂇-𠂇)
𠂇 (U+536C)	𠂇 (U+536D)		(昂-昂)
友 (U+53CB)	友 (U+72AE)	(拔-拔, 髮-髮)	
回 (U+56DE)	𠂇 (U+56EC)		(圖-圖)
呈 (U+573C)	𠂇白工		(捏-捏, 陞-陞)
𠂇白工	𠂇白土		(毀-毀)
久 (U+5092)	𠂇 (U+590A)	𠂇 (U+6535)	(𠂇-𠂇)
大 (U+5927)	太 (U+592A)		(馱-馱)
大 (U+5927)	犬 (U+72AC)		([器]CJ, 獎-獎, 戾-戾, 达-达, 淚-淚)
天 (U+5929)	𠂇 (U+592D)		(吞-吞, [忝])
少 (U+5C11)	少 (U+23942)		(步-步 → から間接的導入)
己 (U+5DF1)	巳 (U+5DF3)	巳 (U+353E)	(卷-卷, [包]GJ 巳? U+5DF2)
千 (U+5343)	干 (U+5E72)		(刊-刊, [舌]CT)
日 (U+65E5)	𠂇 (U+81FC)		(捏-捏, 陞-陞)
木 (U+6728)	木 (U+6729)	术 (U+672F)	(刹-刹)
木 (U+6728)	𠂇 (U+672E)	𠂇 (U+233B3)	
𠂇 (U+6BB3)	𠂇𠂇又		(沒-沒, 歿-歿 → <𠂇, 几>)
母 (U+6BCB)	母 (U+6BCD)		(每-每)
熏 (U+718F)	熏 (U+242F1)		(薰-薰)
牛 (U+725B)	生 (U+20092)		(告-告)
王 (U+738B)	玉 (U+7389)		(国-国)
申 (U+7533)	𠂇 (U+26954)		
𠂇 (U+753E)	𠂇𠂇𠂇一田		(𠂇-𠂇, 輻-輻)
𠂇 (U+758B)	𠂇𠂇		(噓-噓)
米 (U+7C73)	𠂇 (U+91C6)		(奧-奧, 粵-粵)
𠂇 (U+8279)	𠂇 (U+4491)		(孳-孳)

寬(U+8408)	寬(U+83A7)		(寬-寬)
虽(U+867D)	𩺰(U+27227)		(強-強 → <口,厶>)
賴(U+983C)	賴(U+8CF4)		(瀨-瀨)
輓(U+8ED7)	輓(U+23AA0)		(擊-擊)
韋(U+97CB)	𠂔𠂔𠂔𠂔		(衛-衛)
頁(U+9801)	負(U+8CA0)	𠂔刀貝	(瀨-瀨)
鳥(U+9CE5)	寫(U+29F8F)		(鳳-鳳)
日(U+65E5)	日(U+66F0)		([曹]CJ)
丰(U+4E2F)	丰(U+4E30)		([契]CJ)
𠂔(U+20AD3)	𠂔𠂔𠂔		([育]CT)
丿(U+4E3F)	丨(U+4E28)	(丿 U+4E85)	
乚(U+4E29)	𠂔(U+20088)		
十(U+5341)	𠂔(U+20087)		
壳(U+58F2)	壳(U+58F3)		
卜(U+535C)	卜(U+2E8A)		
冂(U+5182)	冂(U+2E86)		([有]CJ)
𠂔(U+34C1)	𠂔(U+5197)	𠂔(U+2EB3)	(𠂔 should be 𠂔)
匕(U+5315)	匕(U+4E03)	𠂔(U+2090E)	([化]CJ)
厶(U+4EBC)	𠂔(U+4EBD)		([今]CJ)
羊(U+7F8A)	𠂔(U+2EB6)		([犇]GJ)
𠂔(U+5382)	𠂔(U+4E06)		
水(U+6C34)	𠂔(U+6C3A)		
𠂔(U+3563)	𠂔𠂔𠂔		([沿]CT, [匠]CT, [𠂔]CT, [𠂔]CT)
𠂔(U+20B26)	𠂔𠂔𠂔		([鑊]CJ)
𠂔(U+777f)	𠂔(U+25220)		([𠂔]CT)
人(U+4EBA)	𠂔(U+201A2)		
𠂔(U+7F52)	𠂔(U+2EAB)		
𠂔(U+8279)	𠂔(U+5EFE)	𠂔(U+5EFF)	
𠂔(U+8279)	𠂔(U+5344)	𠂔(U+535D)	
𠂔(U+51AB)	𠂔(U+2E80)		([冬]CT)
冂(U+5182)	𠂔(U+5196)		# they usually use different IDC.
𠂔(U+53E3)	𠂔(U+56D7)		
𠂔(U+5188)	𠂔(U+7F53)		
𠂔(U+20509)	𠂔(U+9FB9)		
𠂔(U+2099D)	𠂔(U+9FBA)		
𠂔(U+470C)	𠂔(U+9FBB)		
日(U+65E5)	𠂔(U+5183)		([冒]TJ)
𠂔(U+20A27)	𠂔𠂔𠂔		
𠂔(U+21D49)	𠂔𠂔𠂔𠂔	𠂔𠂔𠂔𠂔	([微]GTJ)
𠂔(U+9B5A)	𠂔(U+29D4B)		
𠂔(U+5140)	𠂔𠂔𠂔		
𠂔(U+26B07)	𠂔(U+26B20)		
吉(U+5409)	吉(U+20BB7)		
𠂔(U+2626A)	𠂔(U+2053F)		(𠂔 <𠂔 for BMP, 𠂔 for extB>)
𠂔(U+5365)	𠂔(U+209E7)		
𠂔(U+5349)	𠂔(U+20983)	𠂔(U+20984)	
𠂔(U+518A)	𠂔(U+20541)		
一(U+4E00)	丿(U+4E3F)		