漢字編碼與應用

葉健欣(善那)

漢語大字典修訂講座

四川大學 文學與新聞學院

2024.5.31

從造字到拼形

1991 錄入佛典，合併造字檔

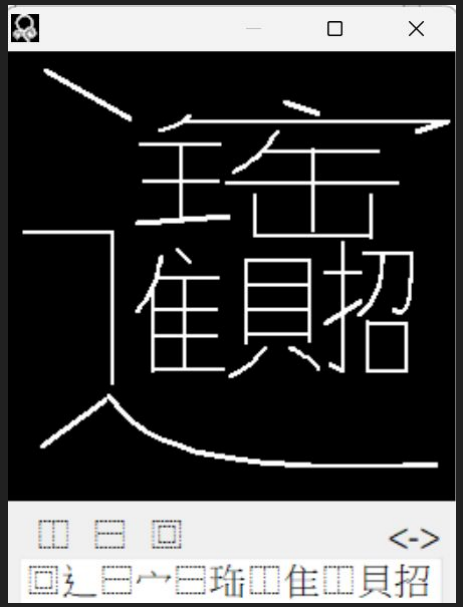
1998 漢語大字典 54000 字矢量字形。分成 12 個字形檔，稀疏編碼（頁碼+每頁32字）。掃描字表、曲線擬合(Curve Fitting)，文件較大。中央研究院漢字構形資料庫採用。

2003 完成 IDS 單線體 組字（招財進寶）

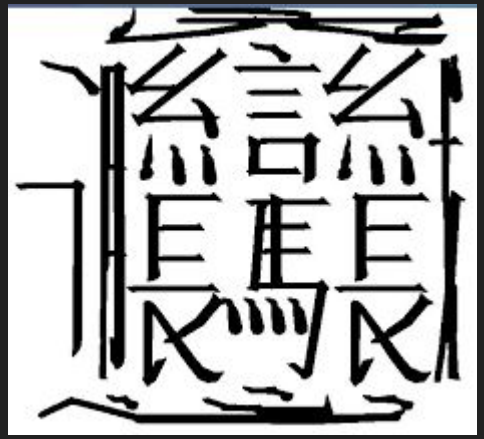
2016 完成 漢字拼形系統

各種嘗試

單線體

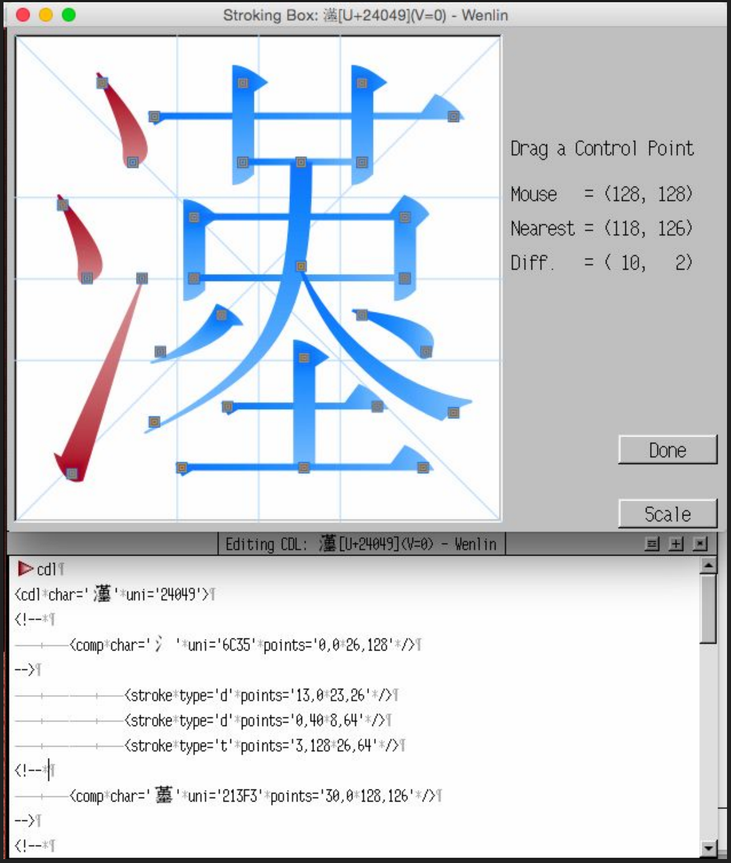


張時劍



朱邦復

文林 (基於XML，商用)

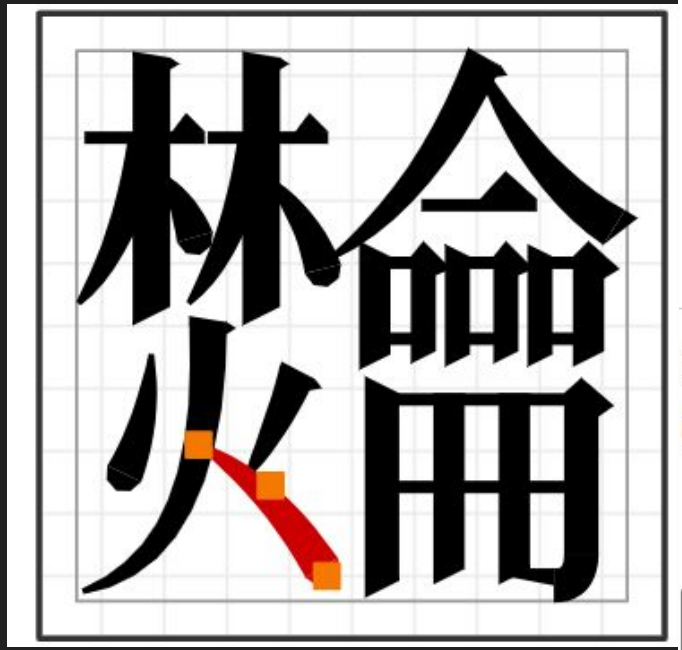


Stroking Box: 𤁉[U+24049](V=0) - Wenlin
Drag a Control Point
Mouse = (128, 128)
Nearest = (118, 126)
Diff. = (10, 2)
Done
Scale
Editing CDL: 𤁉[U+24049](V=0) - Wenlin
cdl
<cdl char='𤁉' uni='24049'>
<!--
<comp char='氵' uni='6C35' points='0,0 26,128' />
-->
<stroke type='d' points='13,0 23,26' />
<stroke type='d' points='0,40 8,64' />
<stroke type='t' points='3,128 26,64' />
<!--
<comp char='𡏳' uni='213F3' points='30,0 128,126' />
-->
<!--

CDL with this comp: 𡏳[U+213F3]
Direct (Recursion Depth=0; Total=35):
(V=0): (19/35)
(V=1): (0/35)
(V=2): (13/35)
(V=3): (2/35)
(V=4): (1/35)
Indirect (Recursion Depth=1; Total=12):
(V=0): (1/12)
(V=1): (0/12)
(V=2): (11/12)
(V=3): (0/12)
(V=4): (0/12)

GlyphWiki+Kage開源

開源，花園字體



部品 **u20c60-jv**

(96,71) → (187,113)

入替 ←　←　4 / 5　→　→入替

元に戻す　やり直す

すべてを選択　選択範囲を反転

コピー　貼り付け

切り取り　手書き開始

部品分解　表示設定...

漢字在計算機

一）編碼（身份證）

二）輸入（鍵盤、筆跡、語音→編碼）

三）顯示（編碼→字形）

缺字問題本質上是缺碼問題（黑戶）

漢字拼形 編碼思路

將 漢字(二維) 視為 西文單詞(一維) 而非字母

是自由長度字符串而非碼位

a, i 是字母也是詞。基本字也有碼位。

英文單詞可自由創造，海納百川。

顯示/打印 路徑

現有系統:字符串→ 單元切分(UTF, IVS)→調取矢量字形→位圖轉換(Rasterize)→繪製/打印

拼形系統:字符串→ 單元切分(IDS/拼形)→組字系統產生矢量→位圖轉換(Rasterize)→繪製/打印

顯示單元切分

一）Unicode UTF-16 (2 , 4 , 6 bytes)

二）IDS 表意序列，不定長度（部件+運算符）

三）漢字拼形，不定長度（部件）

Unicode 字符走既有路徑。

本模塊必須置入 操作系統或文書編輯器內。

漢字拼形 組字系統 原理

https://hanziku.github.io/hzpx/

e.g 初 礻 礻

一）系統 調取初的構字信息。即 礻 刀 以及佈局（框）。

二）礻 礻 表示將 礻 刀 改為 礻 刀。

三）沿用初的佈局，重新產生矢量。

字形矢量之原理

部件及框

99:0:0:3:0:188:200:u8864-01:0:0:0 衤

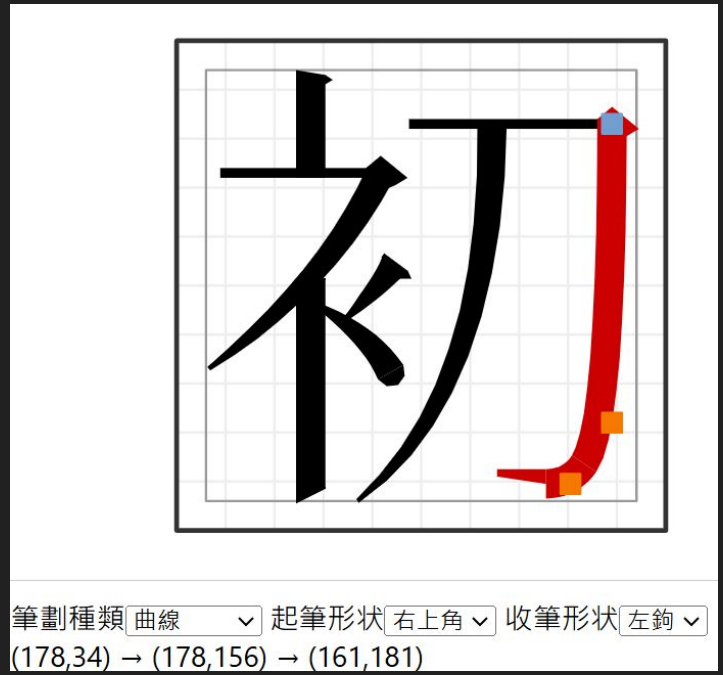
99:0:0:-5:0:205:200:u5200-02:0:0:0 刀

刀 u5200（拆分成三划）

1:0:2:95:34:178:34 直線

2:22:4:178:34:178:156:161:181 曲線帶鉤

2:32:7:129:34:129:144:74:188 曲線（撇）



機會

一）近九萬字只須 5MB (CJK , Ext A - Ext G)

二）相容Unicode 而不受制於Unicode Consortium

我們已可以自主設計操作系統，並布署到硬件。

正是奪回漢字的主導權的最佳時機！