

## 汉字智慧编码与应用

主讲人：叶健欣 教授

主持人：周晓文 教授

### 主讲人简介



叶健欣（Yap Cheah Shen），马来西亚华裔，台湾仓颉输入法发明人朱邦复团队成员，在台湾“中研院”信息处理研究所工作多年，曾担任香港文化传信（朱邦复工作室）字形及Linux研发工程师、台湾“中研院”信息科学研究所访问学人、鸿海集团企业信息系统产品事业群创新产品部门专理、中华开放古籍协会研发长，在信息处理方面建树颇多，曾参与佛光大辞典系统规画及设计、印顺法师全集检索光盘、中华佛教百科全书光盘版、巴利大藏经全文检索网等研究计划与产品开发，是Treasure Association Digital Kangyur Project technical advisor. Adarsha Web/iOS Search Engine Provider 核心检索软件提供者。

### 讲座摘要

回顾台湾地区从八十年代末期以来，为一劳永逸地解决，古籍数字化所碰到的“缺字问题”，围绕在“字形的编码与生成”的种种探索。以“中研院”文献处理实验室以及Unicode IDS为代表的：“解构再重构”的编码方案，一直无法克服，部件占比及布局的难点，直到以“替换法”替代“部件组字法”，以不到5MB的数据量，产生印刷等级的字形品质。本讲座重点讲述“替换法”实现思路并有现场演示。

时间：2024年5月31日 下午15:30-17:30

地点：四川大学江安校区文科楼一区311会议室

# 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 字向量字形。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



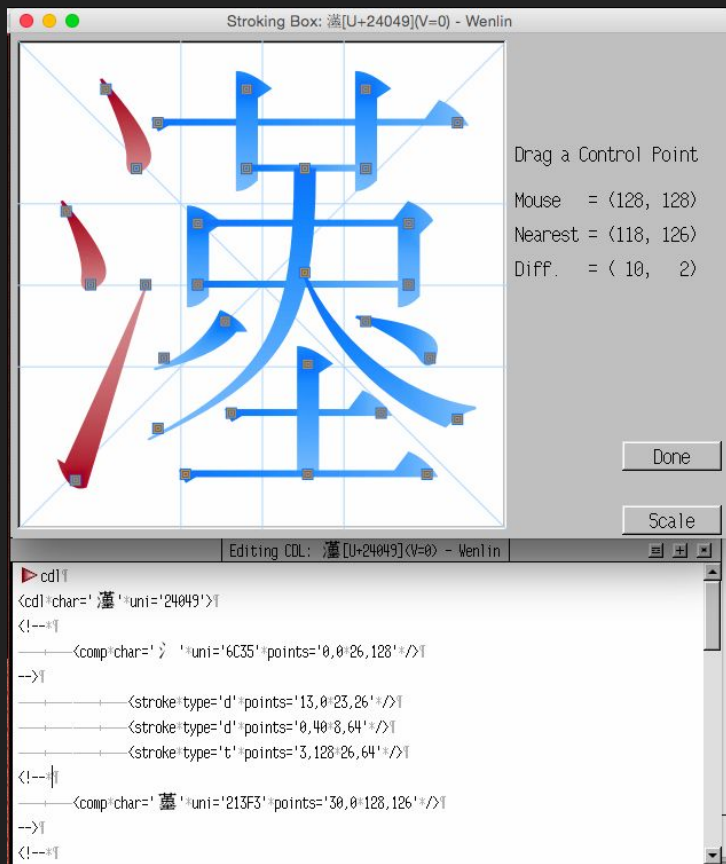
## 張時劍



## 朱邦復



# 文林 (基於XML, 商用)

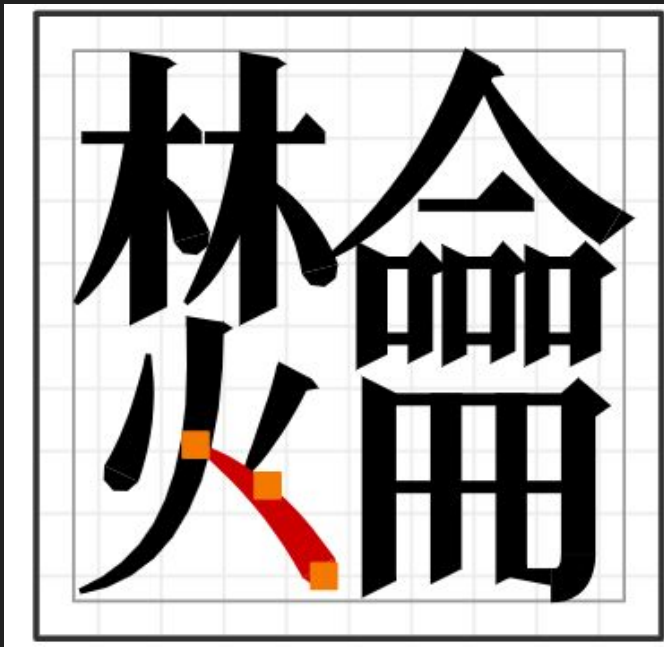


>> CDL with this comp: 墓[U+213F3] <<

- Direct (Recursion Depth=0; Total=35):  
墓 (V=0): 搯 儻 曠 癩 癩 墓 蠆 饑 嚙 塚 廬 殛 燻 溼 燻 燻 燻  
穉 謹 魘 (19/35)  
墓 (V=1): (0/35)  
墓 (V=2): 勤 歎 艱 艱 艱 艱 艱 艱 艱 艱 艱 艱 艱 艱 (13/35)  
墓 (V=3): 艱 艱 (2/35)  
墓 (V=4): 嚙 (1/35)
- Indirect (Recursion Depth=1; Total=12):  
墓 (V=0): 墓 (1/12)  
墓 (V=1): (0/12)  
墓 (V=2): 轟 蕪 蕪 蕪 蕪 蕪 蕪 蕪 蕪 蕪 蕪 蕪 (11/12)  
墓 (V=3): (0/12)  
墓 (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 ㇀ ㇀

一) 系統 調取初的構字信息。即 ㇀ 刀 以及佈局(框)。

二) ㇀ ㇀ 表示將 ㇀ 刀 改為 ㇀ 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 礻

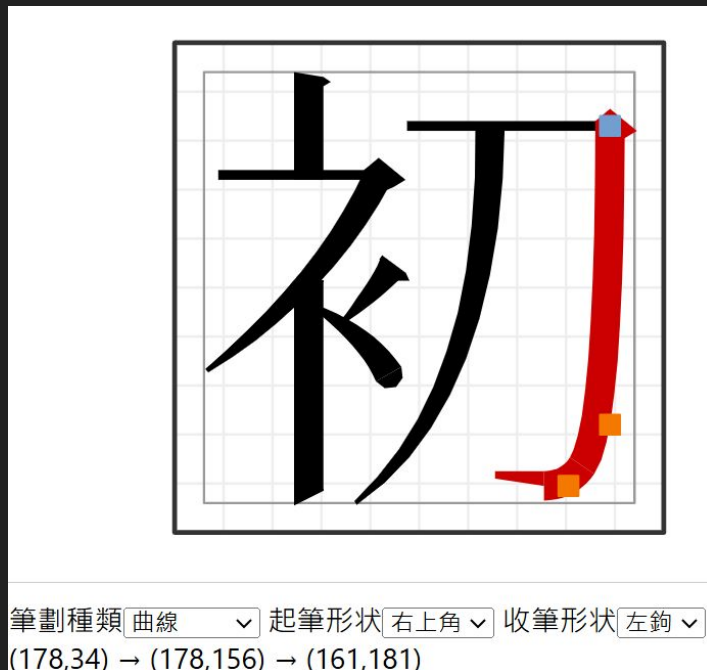
99:0:0:-5:0:205:200:u5200-02:0:0:0 刀

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！