



中央研究院 資訊科學研究所
Institute of Information Science, Academia Sinica

在網路上 解決中文缺字的機制

數位典藏國家型科技計畫 /
數位典藏技術發展組
林德潤

缺字系統

大綱

- 缺字問題
- 因應網路應用的缺字技術
— 漢字構型資料庫
- 技術應用與擴散現況
- 結論與未來規劃



缺字問題

- 中文字具有延展性
 - 透過重組現有的字形或部件可創造新的字形
 - 不同時期的文字於現今編碼無法顯示
- 中文字的編碼空間有限
 - BIG5(13,060字)、GB2312(6763字)、Unicode2.0(20902字)、Unicode5.0(70217字)
- 附帶的異體字問題



因應網路應用的缺字技術

- 缺字檢索系統
- 構字式儲存在應用資料庫的編碼方式
- 網頁呈現的技術
- 漢字構形資料庫尚未新增之缺字處理
- 漢字編碼資訊、漢字拼音資訊



漢字構形資料庫

- 漢字構形資料庫是一個記錄漢字形體的知識資料庫，這些知識包括：
 - 古今漢字的字形演變
 - 古今漢字的字形結構
 - 不同漢字間的使用關係
- 收錄不同時期的字集
 - 小篆、金文、甲骨文、楚系簡帛文字、楷書
- 使用構字式來表達漢字
- 漢字構形資料庫可用來解決文字學數位化的問題，尤其是缺字問題



構字式

- 對於漢字字形結構的制式定義
 - 包含有漢字、部件、字根、連結符號、構字規則
- 利用部件及字根的組合方式來表達漢字
- 定義了三類共計十三個的「構字符號」



構字符號表

類別	符號	Big5	說明	構字式範例
連 接	⋈	8DF2	當部件的連接順序由左至右	順 = 川 ⋈ 頁
	⋉	8DF1	當部件的連接順序由上至下	含 = 今 ⋉ 口
	⋊	8DF3	當部件的連接順序由外至內	圍 = 口 ⋊ 韋
部件序	⊠	8DFC	按部件書寫順序輸入，前後以起始	解 = ⊠ 角 刀 牛
	⊡	8DFD	符號(⊠)和終止符號(⊡)包夾。	⊡
方便符號	⊗	8DF4	二個相同部件直連	炎 = ⊗ 火
	⊘	8DF7	三個相同部件直連	
	⊙	8DF5	二個相同部件橫連	朋 = ⊙ 月
	⊚	8DF8	三個相同部件橫連	
	⊛	8DF6	三個相同部件呈三角狀排列	焱 = ⊛ 火
	⊜	8DFB	四個相同部件橫連	
	⊝	8DFA	四個相同部件直連	
	⊞	8DF9	四個相同部件呈四角狀排列	焱 = ⊞ 火



因應網路應用的缺字技術

- 缺字檢索系統
- 構字式儲存在應用資料庫的編碼方式
- 網頁呈現的技術
- 漢字構型資料庫尚未新增之缺字處理
- 漢字編碼資訊、漢字拼音資訊



缺字檢索系統

- 透過漢字部件檢索
 - 相同部件的相似字
 - 構字式
 - 構字組合
 - 注音(僅Big5字)
- 製作字型圖片
 - 設定字型大小、顏色、字體



Microsoft Internet Explorer

共【107】筆 搜尋關鍵字：【青】 字根組：【青】

構字組合	注音	快速剪貼
青	<一ㄥ	複製
彳青	<一ㄥˋ	複製
小 青	<一ㄥˋ	複製
彳青	<一ㄥˋ	複製
彳青	ㄑㄩㄥˋ	
日青	<一ㄥˋ	
彳青	ㄑㄩㄥˋ	
目青	ㄑㄩㄥˋ	複製
立青	ㄑㄩㄥˋ	複製
米青	ㄑㄩㄥˋ	複製
虫青	<一ㄥ	複製

僅顯示前 10

字形

青

倩

情

清

猜

晴

菁

睛

靖

精

蜻

缺字

情

字型大小 120 更改

字型顏色 black

字體型式 標楷體

注音 <一ㄥˋ

構字式 小 青

檢索符號 小 青

Microsoft Internet Explorer

將 '小 青' 複製到剪貼簿(IE)

確定




因應網路應用的缺字技術

- 缺字檢索系統
- 構字式儲存在應用資料庫的編碼方式
- 網頁呈現的技術
- 漢字構型資料庫尚未新增之缺字處理
- 漢字編碼資訊、漢字拼音資訊



構字式儲存在應用資料庫的編碼方式

- 將構字式中的字根及構字符號轉成 Unicode 跳脫字元，以解決 Big5 資料庫之問題
- 當資料由資料庫取出時，瀏覽器自動解析回原本的構字式

小  青 = 青



因應網路應用的缺字技術

- 缺字檢索系統
- 構字式儲存在應用資料庫的編碼方式
- 網頁呈現的技術
- 漢字構型資料庫尚未新增之缺字處理
- 漢字編碼資訊、漢字拼音資訊



網頁呈現技術

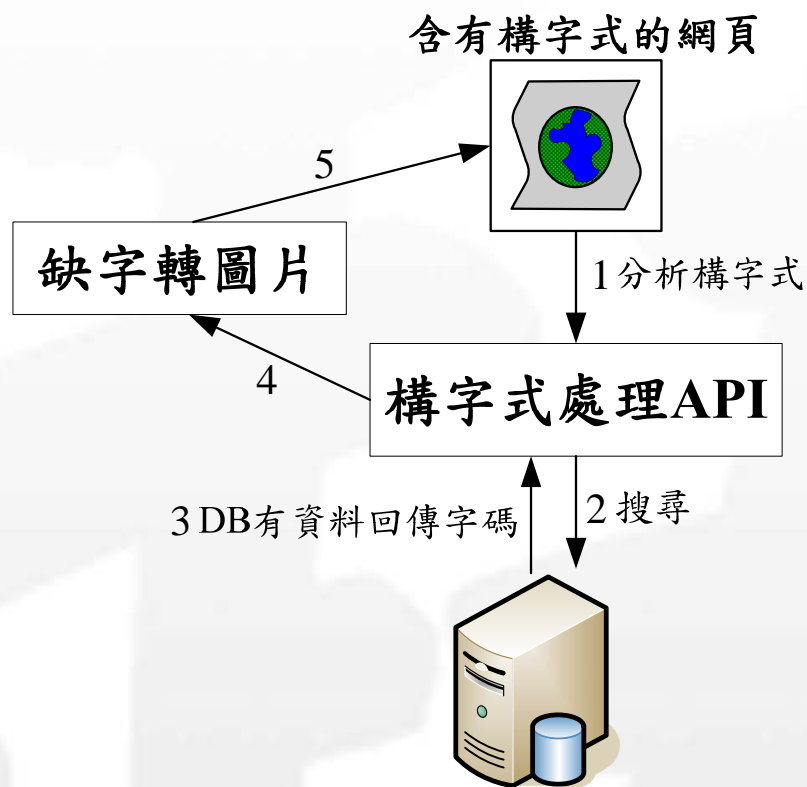
- 解決 網頁 缺字呈現問題
 - 將構字式轉換為缺字圖型
 - 可於任何瀏覽器中使用
 - 使用者端不須安裝缺字字型檔
- 解決方案

	特性
Java Applet (維護，不再更新)	在Client端轉換，速度較慢，使用者必須安裝JRE。
Java Bean (維護，不再更新)	在Server端轉換，速度較快，由程式開發者維護。
Java Script	在Client端轉換，速度較快，由程式開發者維護，步驟簡單好上手。



網頁呈現流程

- 網頁呈現流程



Java Script Example

```
<HTML>
<SCRIPT src="http://char.ndap.org.tw/API/ics.js" language="javascript"></SCRIPT>
<BODY>
<div id="d1">包待制出身源流<br>
詩：世事悠悠自酌量，吟詩對酒日初長·<br>
    韓彭功業消磨盡，李杜文章正顯揚·<br>
    曰：庭下月來花弄影，檻前風過竹生凉·<br>
    不如暫把新編玩，公案從頭逐一詳·<br>
</div>
·
<SCRIPT LANGUAGE="JavaScript">
processObject(document.getElementById('d1'),"Red","12");
processPage("Blue","12");
</SCRIPT>
```



缺字系統

回首頁 缺字查詢 正

測試 UTF-8 測試 動態組字&正規化測試 Java Script DEMO

包待制出身源流
詩：世事悠悠自酌量，吟詩對酒日初長·
韓彭功業消磨盡，李杜文章正顯揚·
曰：庭下月來花弄影，檻前風過竹生凉·
不如暫把新編玩，公案從頭逐一詳·

包待制出身源流
詩：世事悠悠自酌量，吟詩對酒日初長·
韓彭功業消磨盡，李杜文章正顯揚·
曰：庭下月來花弄影，檻前風過竹生凉·
不如暫把新編玩，公案從頭逐一詳·

包待制出身源流
詩：世事悠悠自酌量，吟詩對酒日初長·
韓彭功業消磨盡，李杜文章正顯揚·
曰：庭下月來花弄影，檻前風過竹生凉·
不如暫把新編玩，公案從頭逐一詳·



因應網路應用的缺字技術

- 缺字檢索系統
- 構字式儲存在應用資料庫的編碼方式
- 網頁呈現的技術
- 漢字構型資料庫**尚未新增**之缺字處理
- 漢字編碼資訊、漢字拼音資訊









漢字構型資料庫尚未新增之缺字處理

- 解決使用者即時新增字形之問題

- 利用動態組字技術

- 剎那搜尋工坊所發展 (Open Source)
 - 依漢字構形資料庫模式，自訂構字規則

符號	說明	範例
	兩個部件關係是以左右組合	順=  川頁
	兩個部件關係是以上下組合	員=  口貝
	兩個部件關係是由外包圍內	圓=  口員



動態組字

組字圖片網頁版測試 - Microsoft Internet Explorer

檔案(F) 編輯(E) 檢視(V) 我的最愛(A) 工具(T) 說明(H)

← 上一頁 → 搜尋 我的最愛

網址(1) http://140.109.18.243/char/index.jsp 移至 連結 >>

構字式: 虎

字體大小: 100

字體型式: 圓體

圖片格式: GIF

顏色: black

送出

註: 包含組僅限「門、口、囗、冂、凵、廾」

特殊字: 無

測試頁: Demo output

字型大小: 100 更改

字型顏色: black

字體型式: 標楷體

注音: ㄏㄨˇ

構字式: 虎

檢索符號: 虎

虎虎

動態產生的字

缺字

虎

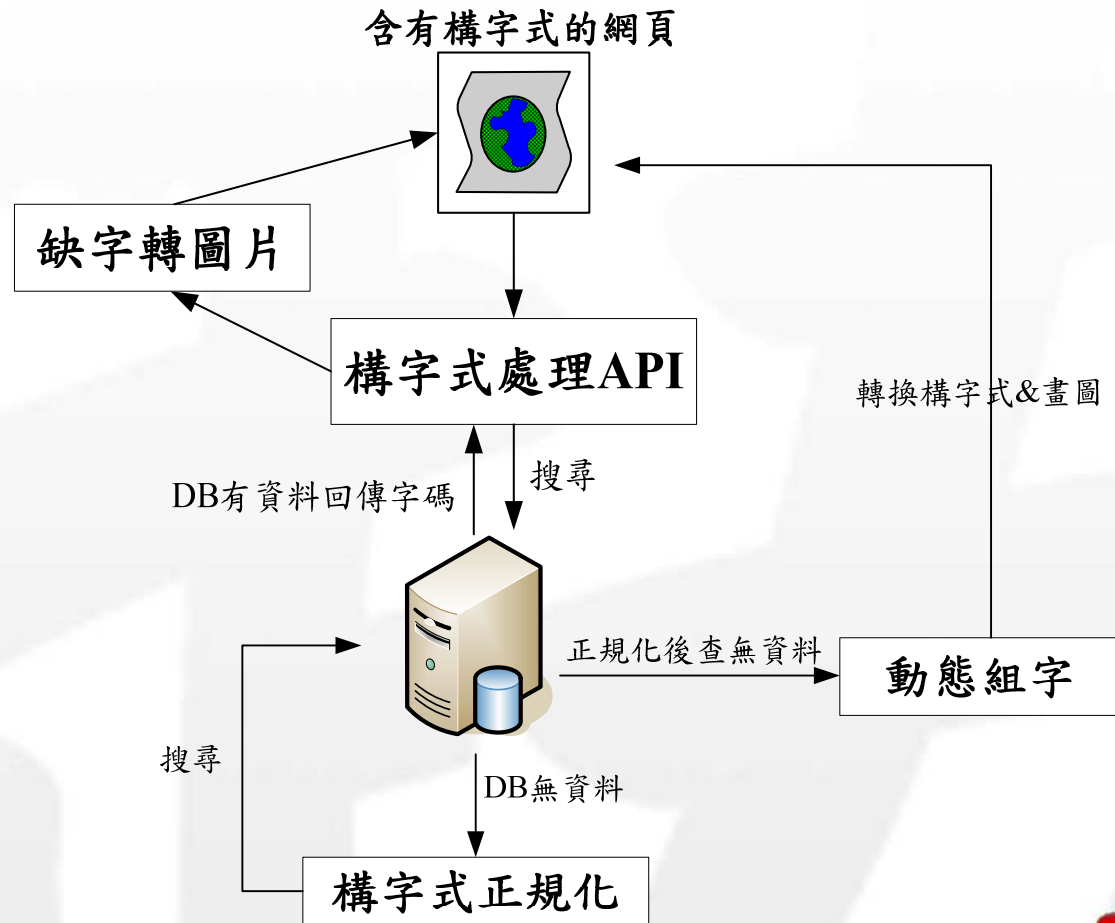
漢字構形資料庫產生的字

字

院 資訊科學研究所
Information Science, Academia Sinica

動態組字與漢字構形資料庫整合

- 整合後流程



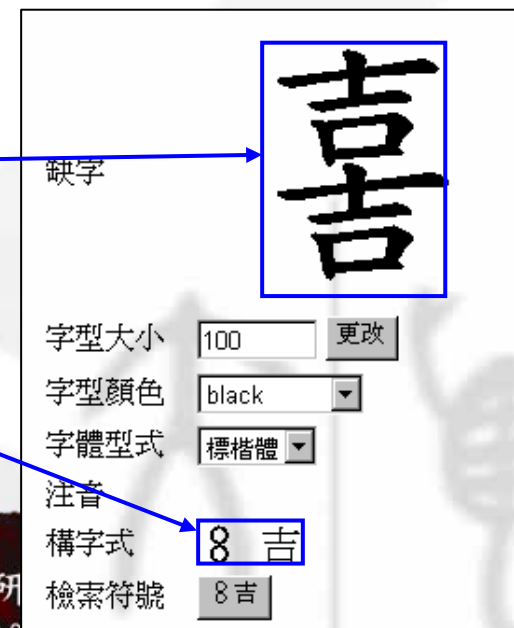
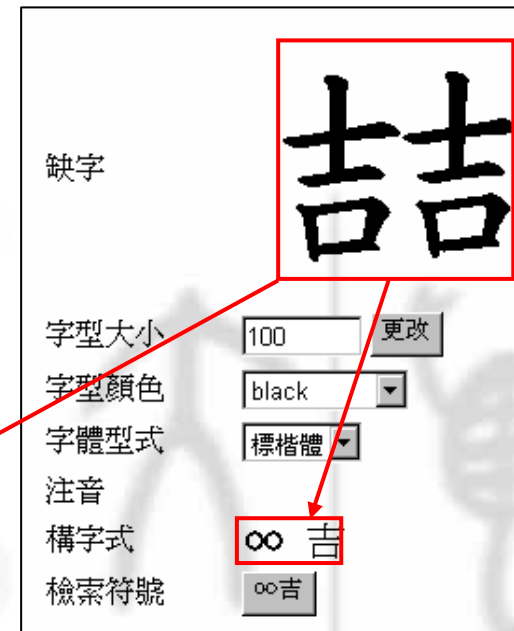
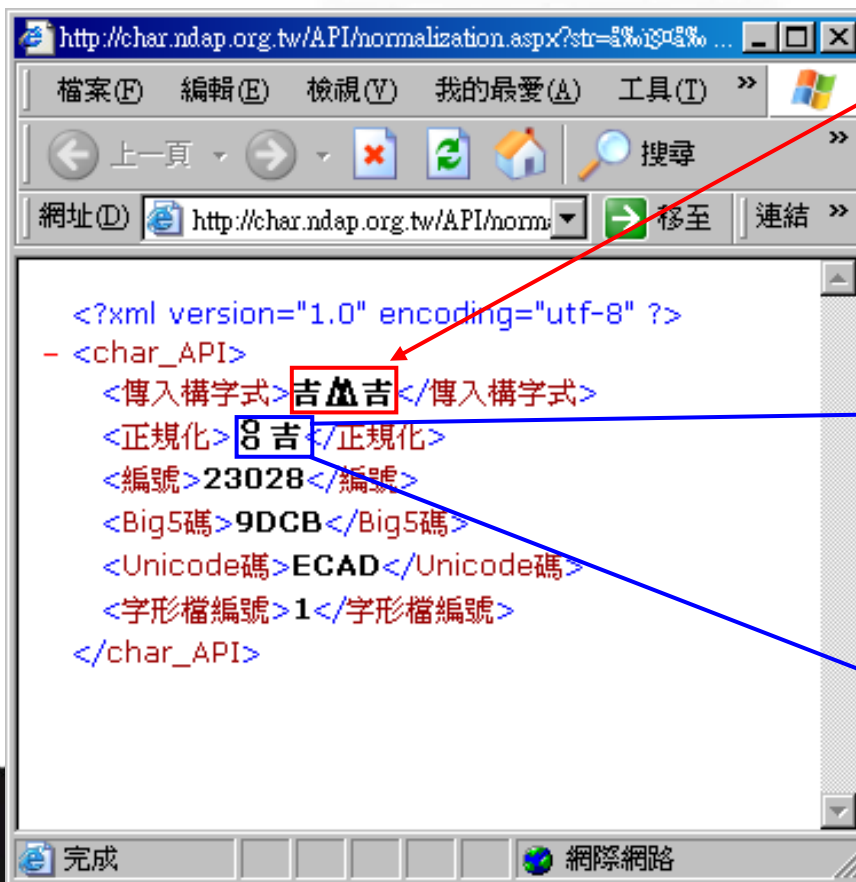
構字式正規化

- 使用者欲查詢“**龍**”
- 可能輸入構字式，與資料庫中的構字式不同(龍龍)
 - 龍龍△龍龍
 - 龍龍△龍龍
 - 形龍龍龍龍□
- 將輸入之構字式拆解成依筆劃順序排列的字根組
 - 龍龍龍龍
- 再到資料庫找尋相同字根組
- 反查詢出構字式



構字式正規化(cont')

- 不先做正規化的原因



因應網路應用的缺字技術

- 缺字檢索系統
- 構字式儲存在應用資料庫的編碼方式
- 網頁呈現的技術
- 漢字構型資料庫尚未新增之缺字處理
- 漢字編碼資訊、漢字拼音資訊



漢字編碼資訊

- 進階開發人員需要構字式的相關資訊
 - 正規化表示法
 - 漢字資料庫漢字編號
 - Big5碼
 - Unicode碼
 - 字型檔編號
- 以標準XML格式傳回查詢結果

```
<?xml version="1.0" encoding="utf-8" ?>  
- <char_API>  
  <傳入構字式>方龠方龠土</傳入構字式>  
  <正規化>𠄎方龠土</正規化>  
  <編號>22987</編號>  
  <Big5碼>BAD0</Big5碼>  
  <Unicode碼>F4AB</Unicode碼>  
  <字形檔編號>0</字形檔編號>  
</char_API>
```



漢字拼音資訊

- 供進階開發人員使用
 - 使用者輸入中文字(BIG5字集)
- 依不同需求，得到不同結果包含
 - 國音第一式
 - 國音第二式
 - 通用拼音
 - 漢語拼音
 - 韋式
 - 雅禮
 - 國語羅馬
- 以標準XML格式回傳

```
<?xml version="1.0" encoding="utf-8" ?>  
- <拼音轉換>  
  <傳入字串>測試</傳入字串>  
  <國音第一式>ㄘㄨㄛˋ ㄕㄩˋ </國音第一式>  
  <國音第二式>tse shr</國音第二式>  
  <通用拼音>ce shih</通用拼音>  
  <漢語拼音>ce shi</漢語拼音>  
  <韋式>ts`e shih</韋式>  
  <雅禮>tse shr</雅禮>  
  <國語羅馬>tseh shyh</國語羅馬>  
</拼音轉換>
```



構字式相關API

- 構字式轉缺字圖形
 - 如Java Script
- 漢字編碼資訊
- 漢字拼音資訊



技術應用與擴散現況

- 目前已應用之系統(2008/01/08統計)
- 常用構字式統計



目前已應用之系統(2008/01/08統計)

系統名稱	單位	使用構字式字數	URL
傳圖人名權威	史語所	635	http://ndweb.iis.sinica.edu.tw/fsnpeople
拓片典藏系統	史語所	10419	http://ndweb.iis.sinica.edu.tw/rub
傅斯年典藏系統	史語所	351	http://ndweb.iis.sinica.edu.tw/rarebook
漢代簡牘系統	史語所	12718	http://ndweb.iis.sinica.edu.tw/woodslip
故宮書畫數位典藏系統	故宮	574	http://ndweb.iis.sinica.edu.tw/npm_public
故宮人名權威	故宮	476	http://ndweb.iis.sinica.edu.tw/people
漢籍電子文獻3.0版	史語所 近史所 台史所 文哲所	767601	http://dbj.sinica.edu.tw:8080/handy/
殷周金文暨青銅器資料庫	史語所	7890	http://db1n.sinica.edu.tw/textdb/bronze
故宮數位典藏器物子計畫	故宮	約13000	http://antiquities.npm.gov.tw
數位典藏 聯合目錄	網路核心 平台計畫	9129	http://catalog.ndap.org.tw/dacs5/System/Main.jsp



常用構字式統計

構字式	字碼	Count
爲	%uF40F	17653
倉	%uEFEF	8328
厶	%uF65D	8311
〇口△言	%uF1A3	8185
宀	%u9FD6	7961
墻	%uF057	6211
譽	%uF025	6125
泉	%uF53C	5241
桀	%u9FE4	5144
垂	%u85D0	5019

構字式	字碼	Count
昷	%uF091	4667
𠂇	%uF649	4608
段	%uEE98	4258
屈	%uEEE5	3887
段	%uF5AE	3861
胃	%uEEDC	3529
血△厶	動態組字產生	3485
许	%uEF9B	3479
金△戍	%uF846	3139
郑	%uEF85	2981



結論

- 整合性呈現缺字
 - 使用圖片呈現缺字
 - 一般使用者無須安裝造字字型檔
 - 可自定字型顏色、大小
 - 以構字式為中心，漢字構型資料庫為主軸，動態組字技術為輔
 - 漢字構形資料庫尚未更新前也不會有缺字問題產生
- 相容性高
 - 相容於任何Web Server平台
 - 相容於任何程式語言開發的Web Page
- 已經累積豐富的內容，使用單位包括
 - 中央研究院史語所、民族所、近史所、台史所、文哲所
 - 故宮博物館



未來規劃

- 漢字構形資料庫線上化
 - 針對不同字集透過部件檢索古今文字
 - 提供異體字表、字形結構、字形演變、字形索引等資料
- 開發漢字檢字入口網站
 - 透過部件、部首、筆畫、字音來檢索古今文字
 - 透過出處來檢索古文字或字帖
 - 針對檢索之字，列出其他字典網站的連結
- 持續推動缺字處理技術



感謝您的聆聽

Q & A



中央研究院 資訊科學研究所
Institute of Information Science, Academia Sinica

