漢字構形資料庫的研發與應用

2009年7月

中研院資訊所文獻處理實驗室

鄧賢瑛 ying0419@iis.sinica.edu.tw

漢字構形資料庫的研發與應用

- 第1章 漢字構形資料庫的研發概要
- 第2章 漢字構形資料庫的部件拆分
- 第3章 漢字構形資料庫的構形編碼
- 第4章 漢字構形資料庫的應用
- 第5章 漢字構形資料庫的展望

第1章
漢字構形資料庫的研發概要

- 1.1字形結構分析與部件檢字
- 1.2構字式與缺字問題
- 1.3異體字表
- 1.4銜接古今文字
- 1.5增收甲骨文、金文及楚系簡帛文字
- 1.6古漢字重文與風格碼
- 1.7漢字構形資料庫的版本沿革（略）
- 1.8漢字構形資料庫的架構及特色

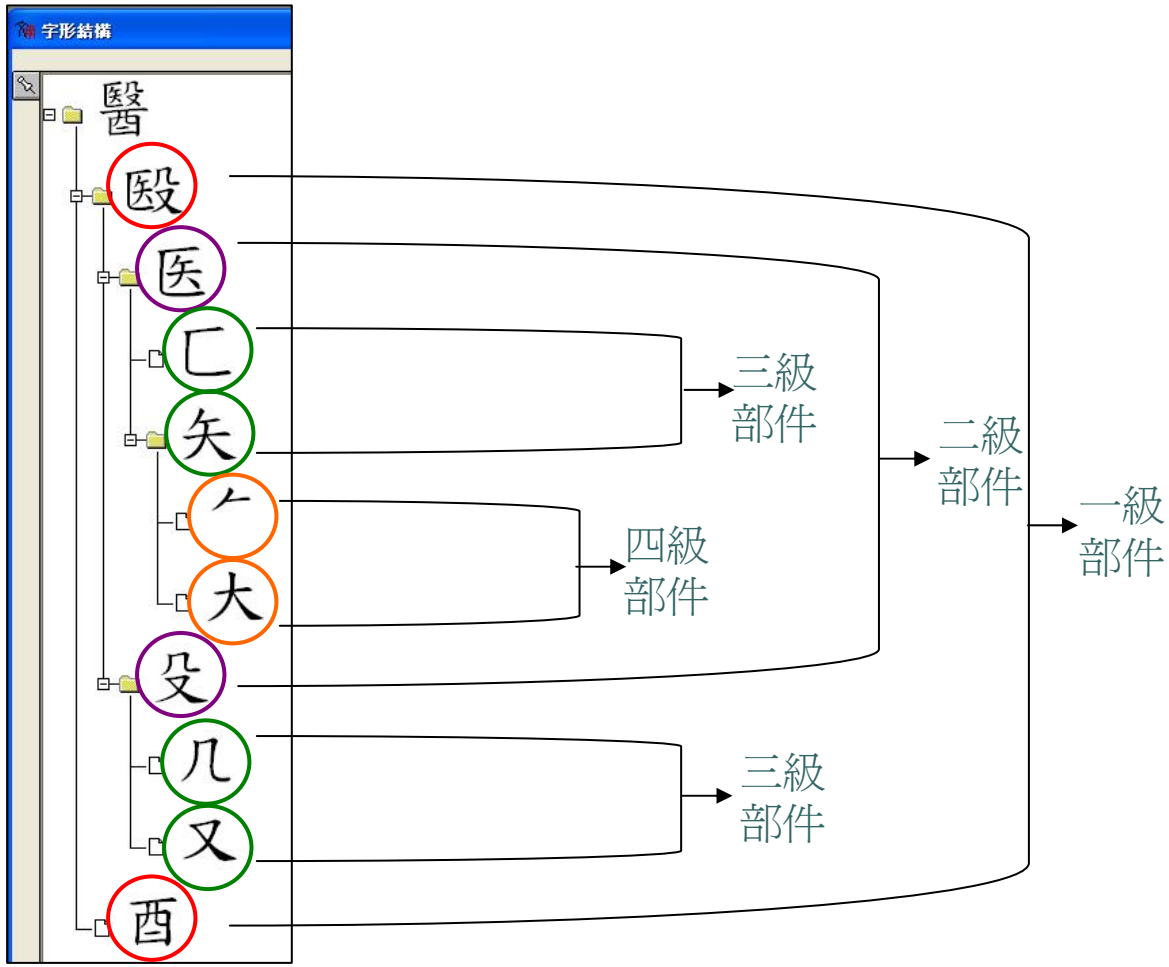
前言

- 漢字構形資料庫，是一個記錄漢字形體知識的資料庫
- 1998年8月推出第一個正式版本
- 研發至今已有超過10年的時間，至2009年6月份爲止，所推出的版本已更新至2.53版，收錄古今文字119,195個及異體字12,208組

1.1字形結構分析與部件檢字

- 1998年8月推出漢字構形資料庫1.0版
- 這是漢字構形資料庫最早的正式版本，收錄五大碼13,051個字形
- 因義構形是漢字的特點，當對漢字進行構形分析時，可將字形依層次拆分爲各級部件
- 各級部件都可用來檢索字形

各級部件



「醫」字的各級部件

部件檢字舉例

圖1.2：以「矢」進行部件檢字的結果

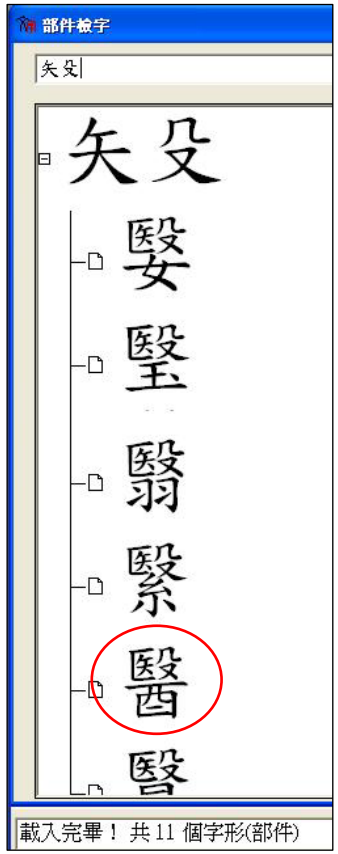


圖1.3：以「矢」和「殳」進行部件檢字的結果

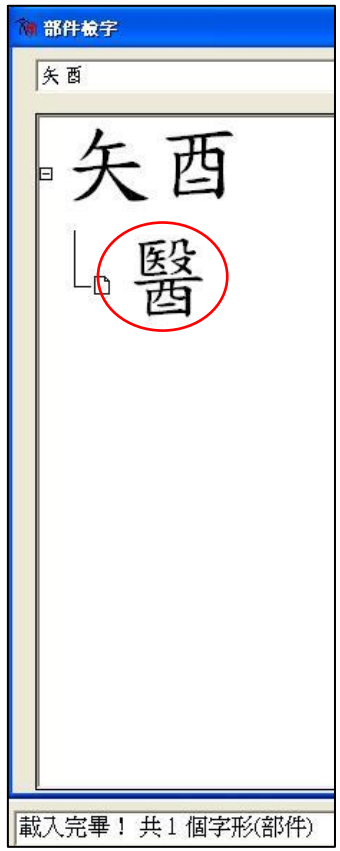


圖1.4：以「矢」和「酉」進行部件檢字的結果

1.2構字式與缺字問題

- 1999年1月20日推出1.1版
- 擴充漢字構形資料庫1.0版的字形。除了原已收錄的五大字集以外，收錄《漢語大字典》的單字，使漢字構形資料庫所收錄的字數擴充至將近5萬字
- 在這5萬字當中，有許多是電腦的缺字，因此需要在計算機中制式表達這些缺字

構字式

- 文獻處理實驗室將漢字部件的組合方式，簡化爲橫連、直連與包含三種組合方式，分別以⿰、⿱、⿴三種連接符號表示
- 使用連接符號連接部件的字形結構表達方式稱作構字式
- 可利用有限的部件表示無窮的漢字

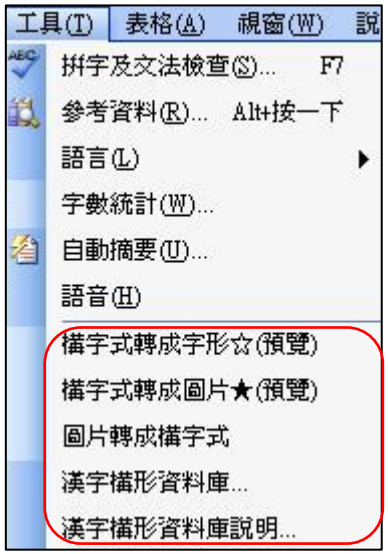
構字式舉例

- 「毆」的構字式可寫作「医殳」
- 「醫」的構字式可寫作「毆酉」
- 「医」的構字式可寫作「匚矢」

缺字預覽巨集

- 2000年10月18日推出的1.2版中，加入了Word缺字預覽巨集
- 使用者可以在Word文件中先輸入缺字字形的構字式，再透過Word缺字預覽，顯示文件中的缺字字形

缺字預覽舉例



Word缺字預覽巨集

正坐之間，忽然土⿰日比皆一道黑氣⿰氵⿰日比中天，須臾不見天日，晡時雖散，仍乃不大明朗。包公心甚疑，其必有⿰宀⿰厶兔柱。是夜左右點起火⿰日比丁燭，包公困倦，伏几而臥。

原始文件

缺字預覽

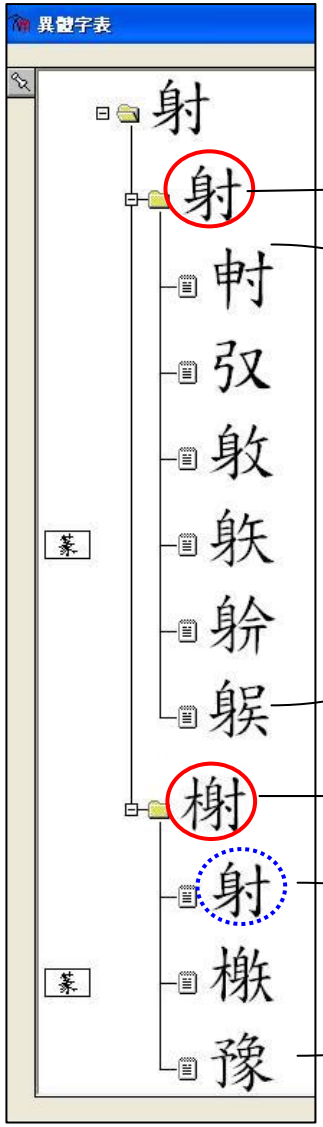
正坐之間，忽然堦前一道黑氣冲天，須臾不見天日，晡時雖散，仍乃不大明朗。包公心甚疑，其必有宽柱。是夜左右點起灯燭，包公困倦，伏几而臥。

缺字預覽文件

Word缺字預覽結果

1.3異體字表

- 2001年1月19日推出1.3版
- 在中文缺字中有大部分是屬於異體字，這和漢字「一字多形」的特點息息相關
- 增添了異體字表的功能，一共收入《漢語大字典》異體字表12,208組



射」字的異體字表

- 在漢字中，主體字與異體字的關係有時候是相對的
- 某個字可能本身是主體字，同時又是其他字形的異體字
- 例如當「射」字作主體字時，它的異體字有7個
- 但「射」字同時又是另一個主體字「榭」字的異體字

1.4銜接古今文字

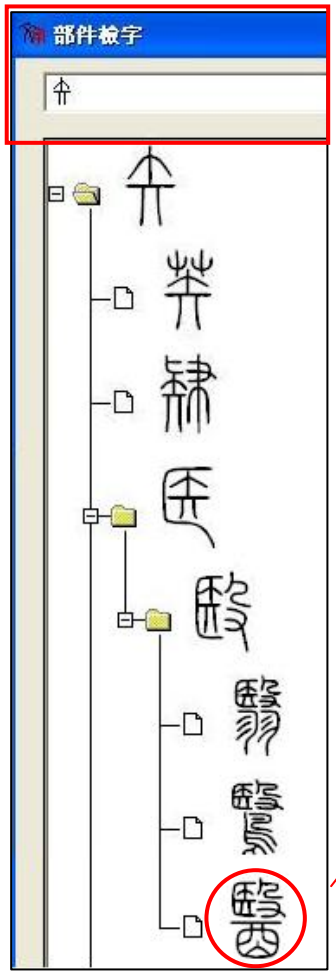
- 2002年7月2日發佈 2.0版
- 漢字構形資料庫2.0版開始收錄古漢字，開始著手收錄《說文解字詁林》中9,831個小篆字形
- 2003年3月17日推出的2.1版時，已完整收齊《說文解字詁林》中的小篆及重文字形一共11,100個

1.4銜接古今文字（續）

- 小篆可以說是研究古文字與今文字的過渡橋樑，《說文》中的小篆不但保存古漢字演變的線索，也是現今漢字尋求字源的重要參考依據
- 在漢字構形資料庫中，檢索小篆字形同樣也是利用部件檢字的方式，並且依小篆的字形進行字形結構分析

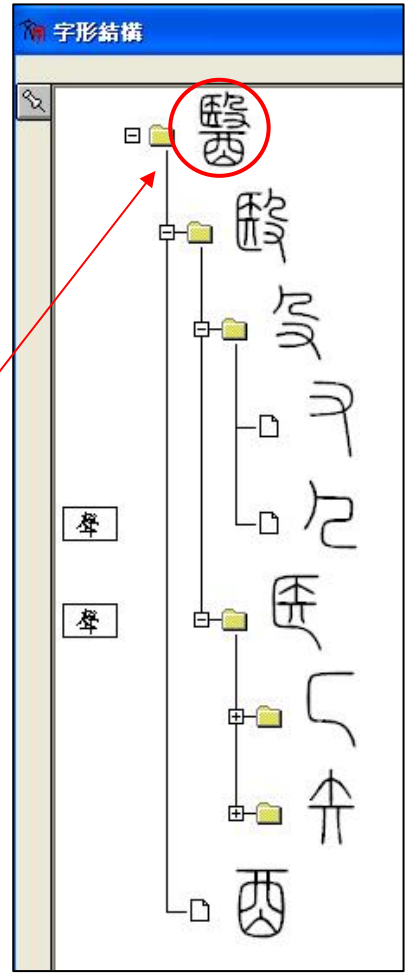
1.4銜接古今文字（續）

先利用「醫」的部件「矢」字檢索而得到「醫」字



以「矢」進行小篆部件檢字的結果

再點選「醫」字後，即可看見它的字形結構



「醫」字的小篆字形結構

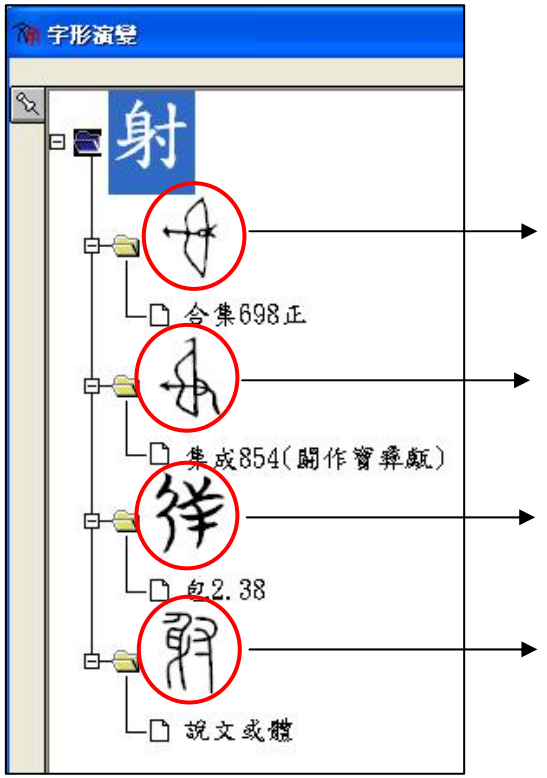
1.4銜接古今文字（續）

射」字的小篆異體字表

射」字的小篆字形

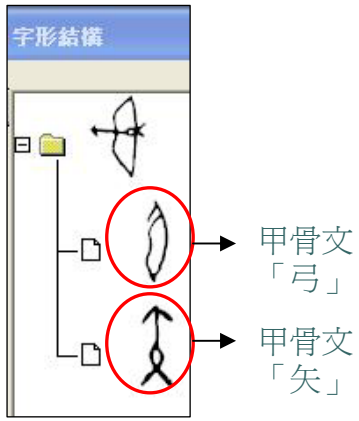
現今楷書的「射」字在《說文》小篆中是「  」字的或體，在漢字構形資料庫的小篆異體字表中，即將「射」 字列在「  」字底下，並在字形演變視窗中顯示小篆「射」字爲說文或體

1.5增收甲骨文、金文及楚系簡帛文字

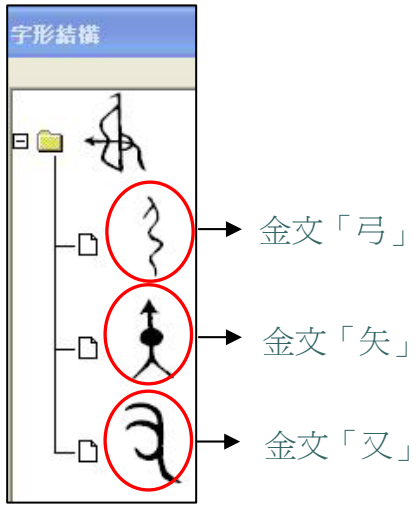


「射」字的字形演變

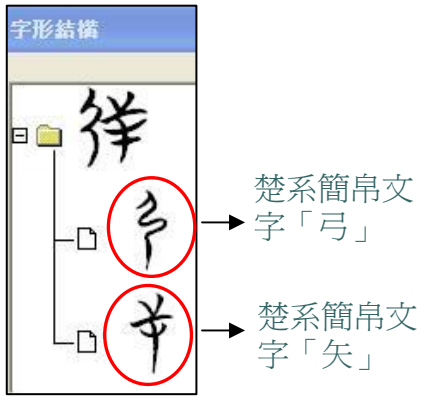
- 2004年12月6日推出的2.2版，開始收錄《金文編》的金文字形
- 2005年8月3日推出的2.3版，則開始收錄《楚系簡帛文字編》的楚系簡帛文字
- 2006年8月2日推出的2.4版，開始收錄《殷墟甲骨刻辭類纂》的甲骨文字形



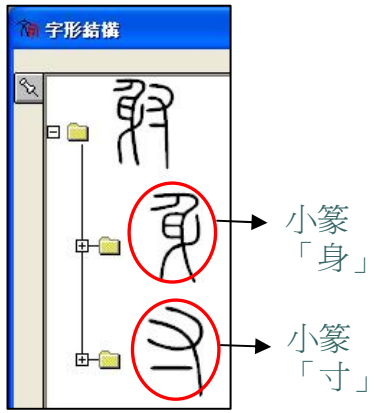
甲骨文「射」字字形結構



金文「射」字字形結構



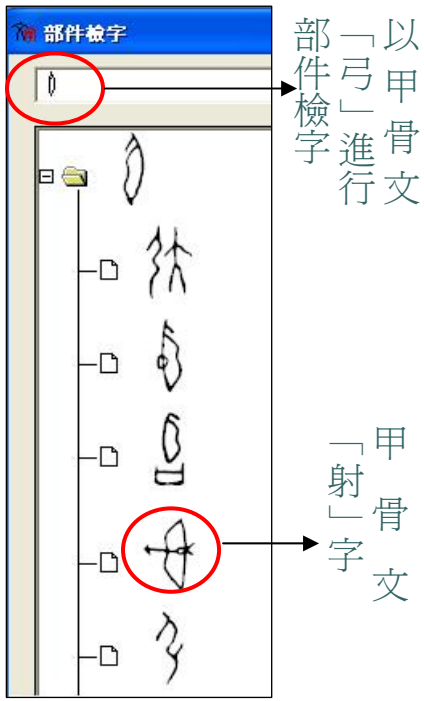
楚系簡帛文字「射」字字形結構



小篆「射」字字形結構

- 現代楷書看來是同一字，但在古漢字卻可能有不同的字形結構
- 在對古漢字進行部件檢字時，也需要依照不同的字形結構，選擇適合的部件

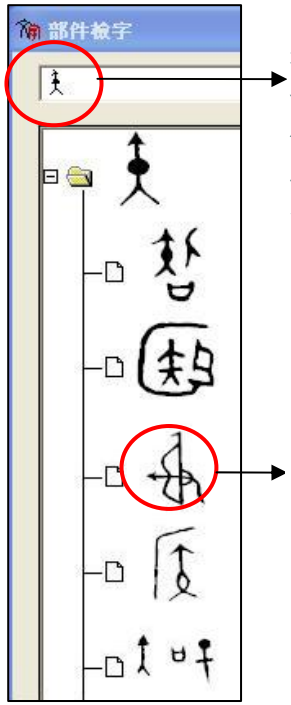
1.5增收甲骨文、金文及楚系簡帛文字（續）



以甲骨文「弓」進行部件檢字

甲骨文「射」字

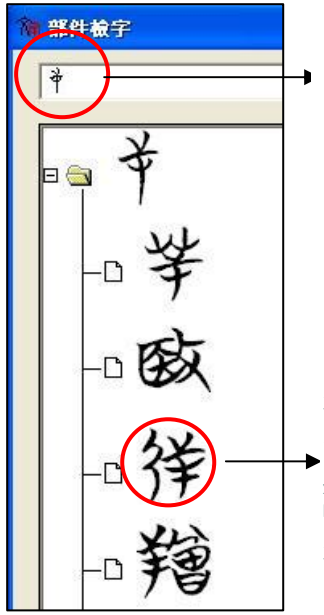
以「弓」進行甲骨文部件檢字



以金文「矢」進行部件檢字

金文「射」字

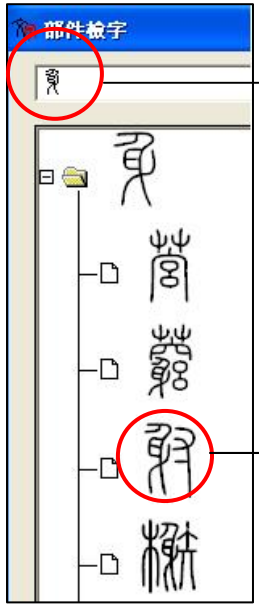
以「矢」進行金文部件檢字



以楚系簡帛文字「矢」進行部件檢字

楚系簡帛文字「射」字

以「矢」進行楚系簡帛文字部件檢字



以小篆「身」進行部件檢字

小篆「射」字

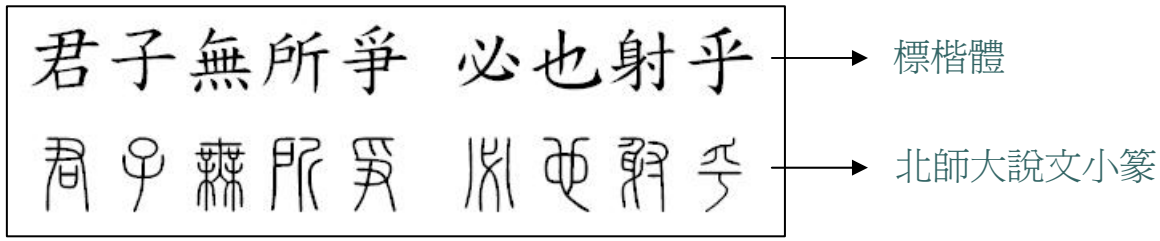
以「身」進行小篆部件檢字

1.6古漢字重文與風格碼

- 2007年8月9日推出2.5版
- 大量增收古漢字，並且開始利用風格碼進行古漢字的編碼工作
- 新增出處檢字功能
- 新增了自動貼圖至Microsoft Office Word的功能

古漢字字型

- 過去在解決古漢字的顯示問題時，通常是先製作古漢字字型，再透過字型切換的方式顯示

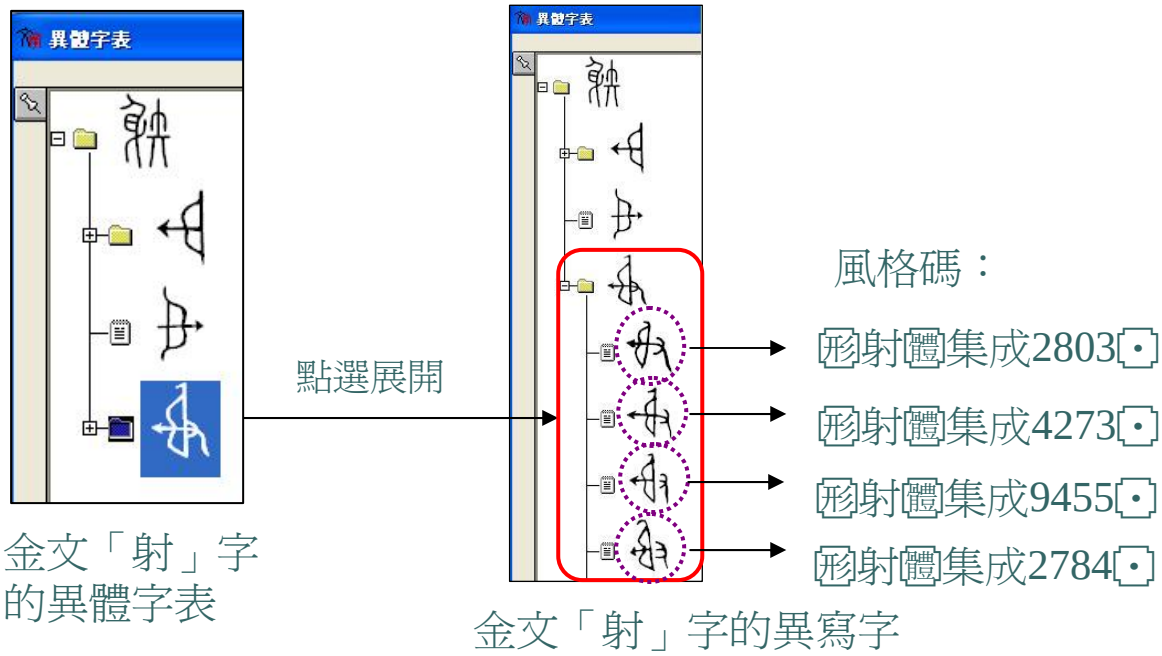


利用切換字型的方式，顯示小篆字形

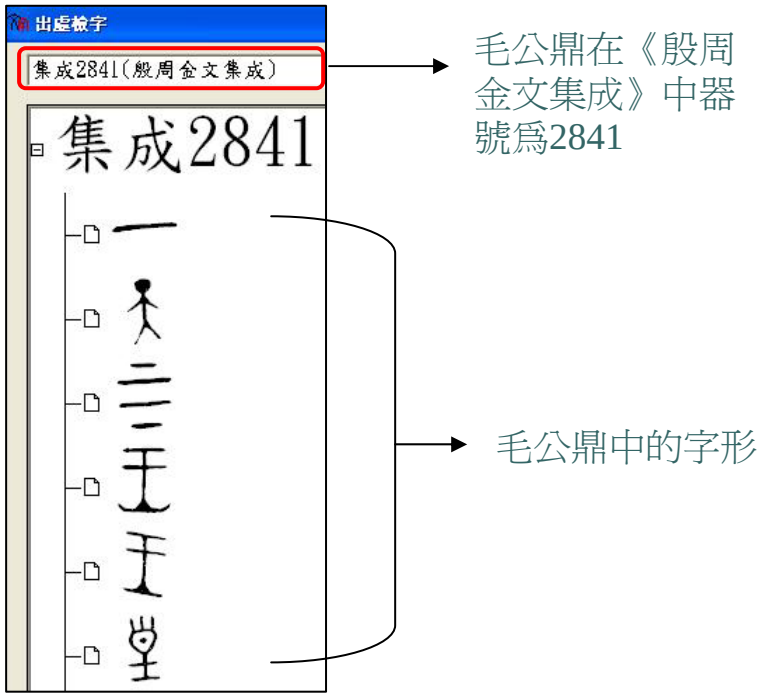
古漢字的編碼工作

- 但形體不同的重文。如果要運用切換字型的方式顯示這些具有大量異體字的古漢字時，就會面臨難以抉擇的困擾
- 有鑑於此，於是漢字構形資料庫2.5版開始採用古漢字本身的源流資訊進行編碼，與楷書運用構字式的編碼方式有所區分

古漢字的編碼工作（續）


異體字表
點選展開
異體字表
風格碼：
形射體集成2803
形射體集成4273
形射體集成9455
形射體集成2784
金文「射」字的異體字表
金文「射」字的異寫字

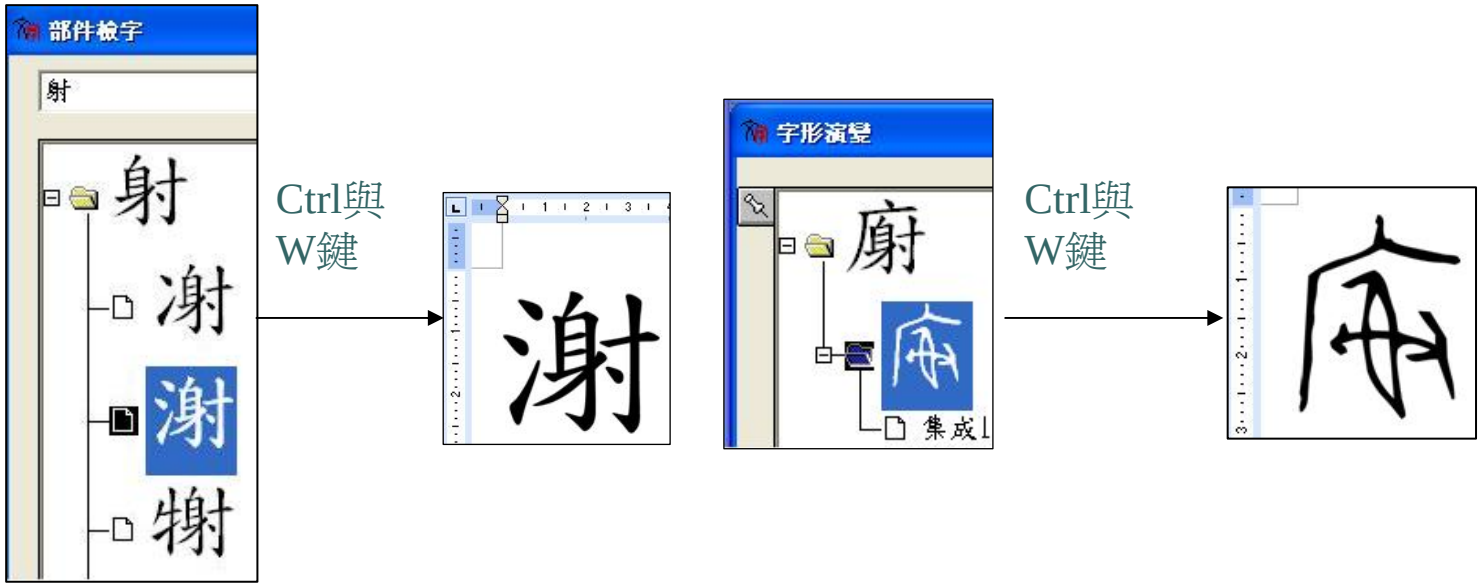
新增出處檢字功能



利用出處檢字檢索毛公鼎上的字形

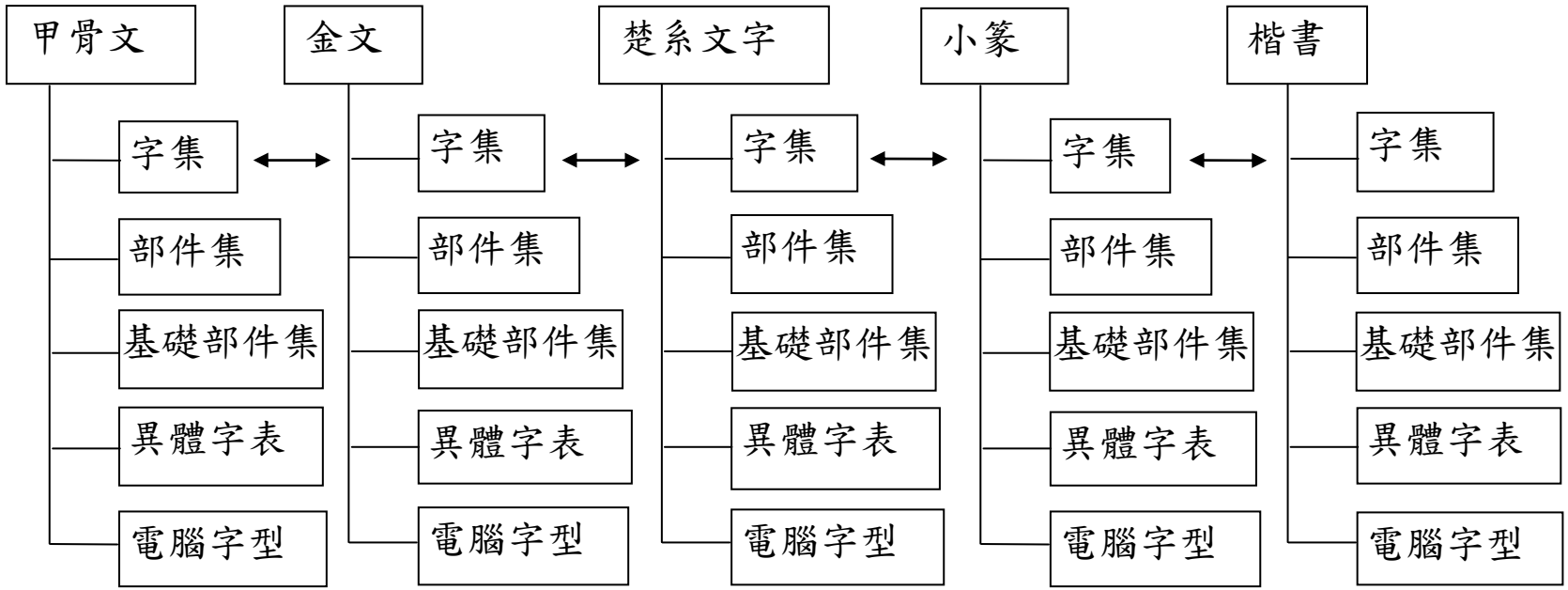
- 例如，使用者若想檢索毛公鼎（《殷周金文集成》器號2841）上的所有字形，即可利用毛公鼎的器號在出處檢字上進行檢索
- 可提供檢索的條件：甲骨文編號、金文器號，或楚系簡帛文字的簡號

新增了自動貼圖至Microsoft Office Word的功能



在Microsoft Office Word中貼入字形圖片

1.8漢字構形資料庫的架構及特色



漢字構形資料庫的組成

1.8漢字構形資料庫的架構及特色（續）

- 銜接古今文字以反映字形源流演變
- 收錄不同歷史時期的異體字表，以表達不同漢字在各個歷史層面的使用關係
- 記錄不同歷史時期的文字結構，以呈現漢字因義構形的特點
- 使用構字式及風格碼來解決古今漢字的編碼問題

第2章
漢字構形資料庫的部件拆分

- 2.1名詞釋義
- 2.2基礎部件的規範
- 2.3部件拆分的一些問題
- 2.4五大字集部件表
- 2.5簡化字集部件表

2.1名詞釋義

- 部件：由筆畫組成具有組配漢字功能之構字單位，當一個形體被用來構造其他的字，成爲所構字的一部分時，是爲所構字的部件
- 例如「毆」、「酉」、「医」、「殳」皆爲「醫」的部件

2.1名詞釋義（續）

- 成字部件：可以獨立成字的部件稱成字部件，當它不作爲其他字的部件時，本身就是一個完整的字
- 例如「醫」字的部件「殹」、「酉」、「医」、「殳」、「匸」、「矢」、「大」、「几」、「又」皆可獨立成字，是爲成字部件

2.1名詞釋義（續）

- 非（成）字部件：不能獨立成字的部件稱非字部件，非字部件不能獨立存在，必需依附於其他部件之上
- 例如「醫」字的部件「𠂤」無法獨立成字，是爲非字部件

2.1名詞釋義（續）

- 基礎部件：最小的不再拆分的部件稱基礎部件
- 例如「醫」字的基礎部件有「匚」、「𠂉」「大」、「几」、「又」、「酉」，這些部件都無法再繼續往下拆分

2.1名詞釋義（續）

- 合成部件：由兩個以上的基礎部件組成的部件稱合成部件
- 例如「醫」字的合成部件有：「殹」、「医」、「殳」、「矢」，這些部件都是由兩個以上的基礎部件所組成，可以繼續往下拆分至基礎部件爲止

2.1名詞釋義（續）

- 漢字結構：部件構成漢字的方式和規則
- 結構理據 ：根據字源或參考字源，從漢字的部件組合分析出的造字意圖，稱結構理據
- 例如「醫」字，根據《說文》：「殹，病聲。酉，所以治病也。」可見「醫」字的造字意圖是由「殹」與「酉」而來。根據字源解釋「醫」字造字意圖，即爲「醫」字的結構理據

2.1名詞釋義（續）

- 部件拆分：將漢字拆分爲部件稱部件拆分
- 有理據拆分：根據結構理據所進行部件拆分，稱有理據拆分
- 例如「醫」字，根據《說文》：「殹，病聲。酉，所以治病也。」因此「醫」字可以拆分爲「殹」與「酉」

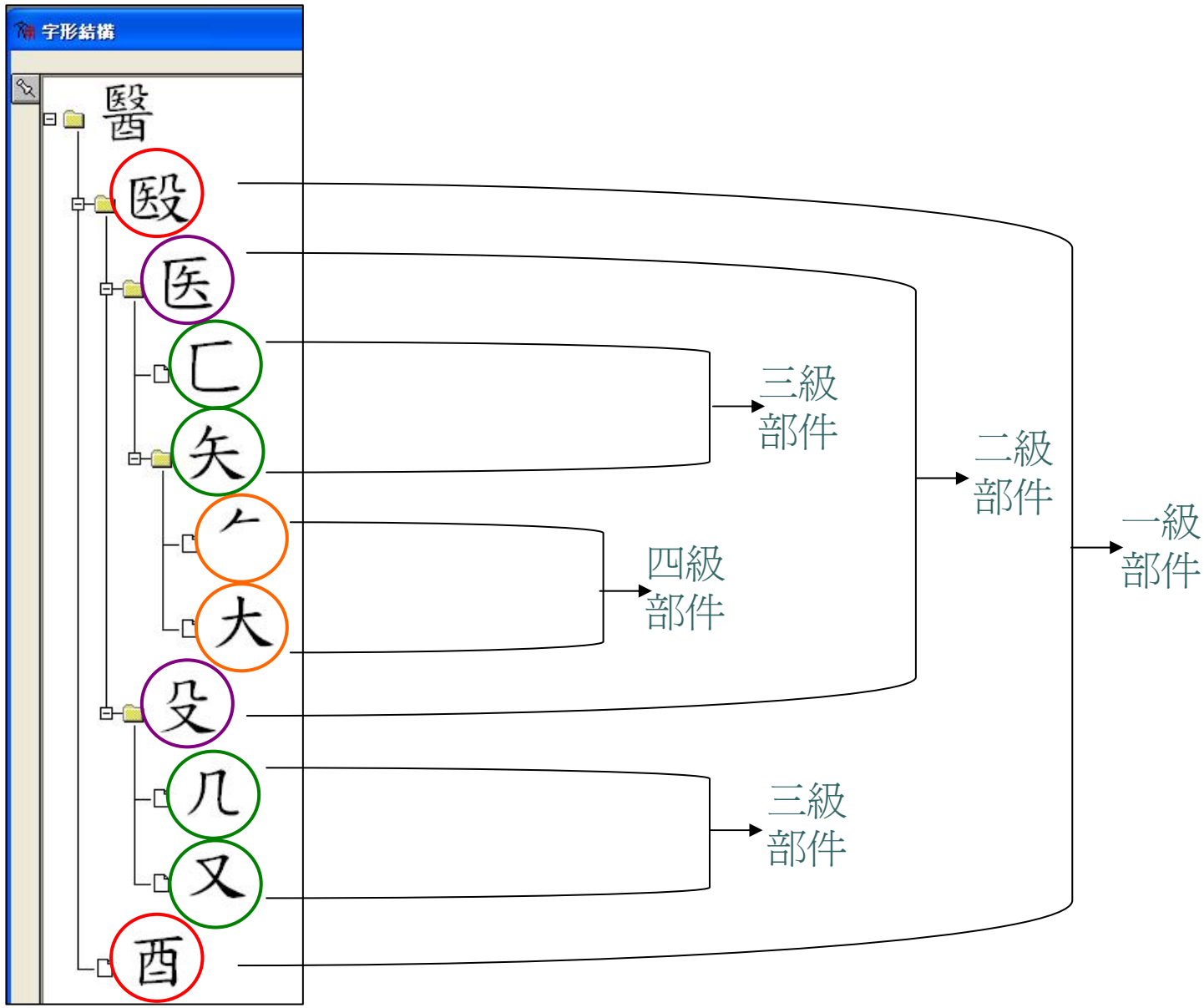
2.1名詞釋義（續）

- 無理據拆分：當無法分析理據或理據與字形發生矛盾時，依照字形所進行的部件拆分，稱無理據拆分
- 例如「矢」往下拆分爲「𠂉」、「大」卻沒有任何字源解釋，因此在「醫」的部件當中，「矢」的拆法爲無理據拆分

2.1名詞釋義（續）

- 部件的層級：依層次拆分的漢字中，部件是有層級的。以「醫」字爲例，含有以下四個層級的部件：

一級部件：殹、酉 → 直接部件
二級部件：医、殳
三級部件：匸、矢、几、又
四級部件：𠂉、大



「醫」字的各級部件

2.2基礎部件的規範

- 現行中文楷書的拆分原則，可參考兩份中文字的基礎部件標準
- 「中文字基礎部件及部件屬性」（編號CNS 11643-2，以下簡稱CNS 11643-2）
- GB 13000.1字符集「漢字部件規範」（編號GF3001-1997，以下簡稱GF3001）

GF3001

- 於1997年發佈
- 是對GB13000.1「信息技術通用多八位編碼字符集（UCS）第一部分：體繫結構與基本多文種平面」中的20,902個中文字進行拆分後得出的基礎部件表以及使用原則
- GF3001有560個基礎部件

CNS 11643-2

- 於2007年發佈
- 是對CNS11643「中文標準交換碼」第1及第2字面的13,051個中文字進行拆分後得出的基礎部件表及使用原則
- CNS 11643-2共有517個基礎部件

漢字構形資料庫的基礎部件

- 漢字構形資料庫在制定基礎部件時，主要是依據CNS 11643-2，兼以GF3001爲輔助參考
- 其與CNS 11643-2相異處爲：採取GF3001的認同使用原則

何謂認同使用

- 當部件因爲在字中所處的部位不同而產生了筆畫變形或比例變化，例如「土」當作字形偏旁時，經常寫作提土旁「扌」，若將「扌」與「土」視爲相同的部件，是爲認同使用
- CNS 11643-2中將這類變化後的部件稱爲附部件，視爲與主部件不同的部件，是採取不認同使用

採取GF3001的認同使用原則

- 附部件的存在會增加檢索之困難，例如部件「木」，分出位置在字形左半邊的附部件「木」和位置在字形下半邊的附部件「木」
- 漢字構形資料庫採取了GF3001的認同使用原則

2.3部件拆分的一些問題

- 漢字構形資料庫主要根據字形理據來進行部件拆分。當字形符合理據的，進行有理據拆分；無法分析理據或理據與字形矛盾的，依字形進行無理據拆分

2.3部件拆分的一些問題（續）

- 對多部件的漢字進行拆分時，應先依漢字組合層次做有理據拆分，直至不能進行有理據拆分而仍需拆分時，再做無理據拆分

有理據拆分與無理據拆分

- 有理據拆分—根據字源
  - 絕大多數楷書字形拆分後的部件都和小篆相同
  - 例如「醫」字拆成「殹」、「酉」

有理據拆分與無理據拆分（續）

- 有理據拆分—參考字源
  - 楷書和小篆的部件差異不只是變形，而是由另一個部件所替代
  - 例如小篆「  （奠）」拆成「（酋）」、「  （丌）」，楷書「奠」字拆成「酋」、「大」，「丌」爲「大」所替代

有理據拆分與無理據拆分（續）

- 無理據拆分
  - 無法分析理據或理據與字形矛盾時採用
  - 例如小篆「  （易）」爲象形字，不再拆分，楷書「易」字則依CNS11643-2拆分成「日」、「勿」

非字部件

- 在上述「有理據拆分—參考字源」的字形中，部分楷書的部件已由另一個部件所替代，這些替代的部件有些爲非字部件
- 例如小篆「  (唐)」拆成「  (庚)」、「  (口)」，楷書「唐」字中替代部件「庚」的「庚」，即爲非字部件

非字基礎部件與非字合成部件

- 非字部件由於不是字，電腦的中文字集不會收錄
- 在漢字構形資料庫中，基礎部件是不可或缺的，因此非字基礎部件絕對要收錄；至於非字合成部件，由於數量較多，基於構字的需要，則可適量收錄

非字合成部件收錄原則

- 目前只要是具有特定構意的非字合成部件，若在《漢語大字典》有兩個（含）以上的單字用到，漢字構形資料庫即予收錄
- 例如「徽」的非字合成部件「微」，構意爲「微聲」，並可構成「黴」、「徾」等字，因此可收錄；至於「唐」的非字合成部件「庚」，構意爲「庚聲」，但僅構成「唐」字，可予取消

無理據拆分的原則

- 在不增加部件的情況下，使用最少的部件來拆分
- 例如「兵」拆分成「丘」、「八」；「易」拆分成「日」、「勿」

2.4五大字集部件表

- 五大字集：其中含常用字5,401個，次常用字7,652個，合計13,053個字，其中有2個字重複編碼，因此實際收錄13,051個中文字，即本文所稱之五大字集
- 依照漢字構形資料庫的部件拆分原則，五大字集拆分後的部件總數爲2,297個，其中基礎部件爲441個，合成部件爲1,856個

2.4五大字集部件表（續）

- 五大字集基礎部件表，共441個（見報告p.33-39）
- 五大字集基礎部件組字頻率表（見報告p.40-47）
- 五大字集合成部件表，共1,856個（見報告p.48-71）

2.5簡化字集部件表

- 簡化字集指的是《簡化字總表》收錄的2,235個簡化字
- 依照漢字構形資料庫的部件拆分原則，簡化字集拆分後的部件總數爲1,122個，其中基礎部件爲367個，合成部件爲755個

2.5簡化字集部件表（續）

- 簡化字集基礎部件表，共367個，扣除和五大字集基礎部件重複的326個外，共有41個（見報告p.72-73）
- 簡化字集合成部件表 ：共有755個，扣除和五大字集合成部件重複的530個外，共有225個（見報告p.74-78）

第3章
漢字構形資料庫的構形編碼

- 3.1部件的組合及識別
- 3.2構字式的制定及使用
- 3.3構字式的處理技巧
- 3.4風格碼的制定及使用

3.1部件的組合及識別原則

- 漢字是由有限的部件所組成，除了不同的部件可組成不同的漢字外，相同的部件也可利用相對位置或部件的個數來組成不同的漢字

3.1部件的組合及識別原則（續）

- 以五大字集的13,501個字形爲例，其中有12,817個字使用的部件均不相同，部件相同但相對位置不同的字共有115組，234個字
  - 例如「架」、「枷」、「栜」三個字均由部件「力」、「口」、「木」所組成，不同的是部件的相對位置

3.1部件的組合及識別原則（續）

- 五大字集還有67個字僅由單一部件重複組合而成，例如「多」、「朋」、「林」、「炎」
- 綜合上述，絕大多數的漢字皆可由其組成的部件來識別，少數字形則須再描述部件的相對位置

3.1部件的組合及識別原則（續）

1. 漢字係由一或多個部件依層次逐級組合而成，絕大多數字形可透過各級部件的組合來識別，例如「謝」字的各級部件組合「言射」、「言身寸」都可用來識別「謝」字
2. 少數漢字的差異僅在於部件的相對位置不同，要識別此類字形，除部件本身外，還須描述部件的相對位置，例如「暉」、「暈」

3.1部件的組合及識別原則（續）

3. 少數漢字是由單一部件重複組合而成，此類部件的組合方式通常爲由左至右，由上至下，或呈三角狀、四角狀排列。
   - 例如「林」字由單一部件「木」由左至右排列，「棗」字由單一部件「朿」由上至下排列，「轟」字由單一部件「車」呈三角狀排列，「燚」字由單一部件「火」呈四角狀排列

3.1部件的組合及識別原則（續）

4. 大多數漢字部件的相對位置，可由前後部件之性質判斷得知。
   - 例如部件「言」與其他部件組合字形時，部件的相對位置通常爲左右，而「言」在左方，因此「言」、「射」組合成「謝」
   - 部件「雨」與其他部件組合字形時，部件的相對位置通常爲上下，而「雨」在上方，因此「雨」、「相」組合成「霜」
   - 部件「口」和其他部件組合字形時，部件的相對位置通常爲內外，而「口」在外，因此「口」、「員」組合成「圓」。

3.1部件的組合及識別原則（續）

5. 基礎部件爲構形識別的最小單位，不得再拆分成其他基礎部件來識別字形。
   - 例如部件「貝」爲CNS 11643-2規定之基礎部件，不得再拆分成「目」、「ハ」，因此「目ハ」無法用以識別「貝」字

3.2構字式的制定及使用

構字符號		說　明	構字式
形標	㊀ ⊡	㊀為字形標示的起點，⊡表示結束，部件則包夾於其中。	寶＝㊀宀王缶貝⊡
定位符號 連接符號	⩓	橫連—當部件的組合順序為由左至右	謝＝言⩓射
	⩔	直連—當部件的組合順序為由上至下	霜＝雨⩔相
	△	包含—當部件的組合順序為由外至內	圓＝口△員
定位符號 重疊符號	㇀	兩個相同部件的組合方式為由左至右	林＝㇀木
	㆓	兩個相同部件的組合方式為由上至下	棗＝㆓束
	㇇	三個相同部件的組合方式為由左至右	孫＝㇇子
	㆔	三個相同部件的組合方式為由上至下	尛＝㆔小
	㆕	三個相同部件的組合方式為三角狀排列	轟＝㆕車
	㇈	四個相同部件的組合方式為由左至右	𠔃＝㇈丨
	㆖	四個相同部件的組合方式為由上至下	𠂤＝㆖一
	㆕	四個相同部件的組合方式為四角狀排列	燚＝㆕火

3.2構字式的制定及使用（續）

1. 構字式可用於識別字形，它係由部件及構字符號組成
2. 構字符號分爲：
   - 形標：不涉及部件的相對位置
   - 定位符號：可用於描述部件的相對位置

3.2構字式的制定及使用（續）

3. 形標爲字形的標示，由構字符號「  」及「  」構成，其中「  」爲字形標示的起點，「  」表示結束，用以識別字形的部件則包夾其中
   - 例如構字式「  宀王缶貝  」可用以識別「寶」字

3.2構字式的制定及使用（續）

4. 定位符號分成
   - 連接符號：主要用以描述不同部件的組合情形
   - 重疊符號：僅用以描述單一部件之重複組合
5. 連接符號有三個：
   - 「  」：橫連，部件的組合順序爲由左至右
   - 「  」：直連，部件的組合順序爲由上至下
   - 「  」：包含，部件的組合順序爲由外至內

3.2構字式的制定及使用（續）

6. 在構字式中，不同的連接符號不可併用，而相同的連接符號可連續使用。
   - 例如「寶」字的構字式不可寫成「宀  王  缶貝」，應寫成「  宀王缶貝  」

3.2構字式的制定及使用（續）

7. 採用形標的構字式，目前採用之排序原則爲部件間的連接順序：先假設不同的連接符號可併用，再抽離連接符號，留下的部件順序即是。
   - 例如先假設「寶」字的構字式可寫成「宀王  缶  貝」，抽離連接符號後，構字式爲「  宀王缶貝  」

3.2構字式的制定及使用（續）

8. 重疊符號共八個，常用的有「〇〇」、「𠔃」、「𠔃𠔃」、「𠔃𠔃」四個，其餘四個不常用的重疊符號爲「〇〇〇」、「𠔃」、「〇〇〇〇」、「𠔃」。構字式中，重疊符號應置於部件之前。
   - 例如構字式「〇〇木」可用以識別「林」字，「𠔃束」可用以識別「棗」字

3.2構字式的制定及使用（續）

9. 在構字式中，形標和連接符號不可併用。
   - 例如「謝」字的構字式可寫成「言射」、「言身寸」或「  言射  」、「言身寸」，但不可寫成「言射」或「言身寸」

3.2構字式的制定及使用（續）

10. 在構字式中，重疊符號可和形標或連接符號併用。
    - 例如「堯」字的構字式可寫成「垚  兀」，也可寫成「  土兀」

3.2構字式的制定及使用（續）

11. 由於一個字可透過各級部件的組合來識別，所以同一個字也可能有數種不同的構字式，但僅有一級部件所組合的構字式最貼近造字意圖 。
    - 例如「謝」字的構字式可寫成「言⿰射」或「言⿰身⿰寸」，其中「言」、「射」爲一級部件，而「身」、「寸」爲二級部件，因此「言⿰射」比「言⿰身⿰寸」更貼近「謝」的造字意圖

3.3構字式的處理技巧

構字式的類型可依構字符號分爲五類

類型	說明	字例	構字式	備註
1	由部件、橫連符號或重疊符號構成	顯	桑⿻頁	只要出現橫連符號即屬於此型
		掰	手⿻分⿻手	
		散	⿰貝⿻攵	
2	由部件、直連符號或重疊符號構成	霾	雨⿱㚔郭	只要出現直連符號即屬於此型
		簠	⺮⿱㚔甫⿱㚔皿	
		鷖	⿰⿰殳⿱㚔鳥	
3	由部件、包含符號或重疊符號構成	麑	鹿⿴兒	只要出現包含符號即屬於此型
		侖	侖⿴⿰口口	
4	由部件、形標或重疊符號構成	邍	㊊辶备条⊙	只要出現形標即屬於此型
		豐	㊊山⿰丰豆⊙	
5	只由重疊符號構成	虩	⿰虎虎	只有出現重疊符號
		戈	⿰戈戈	
		贔	⿻貝貝	
		燚	⿻火火	

構字式的擷取

- 構字式可夾雜在字串中使用，由字串擷取構字式，首先得先找出起始和終結字元。例如「大桑  頁虯髯骨相奇」中構字式的起始和終結字元分別爲「桑」、「頁」
- 構字式的擷取類型可分爲三類，依第一個構字符號是連接符號、形標或重疊符號來判斷

構字式的擷取（續）

類型	符號	字例	構字式	起始	終結	備註
1	⿲	顙	桑⿲頁	桑	頁	當第一個構字符號為連接符號時，則起始字元為⿲、⿳、⿴的前一個；終結字元為連續兩個非構字符號的第一個，例如「大桑⿲頁虮髯骨相奇」之「頁」
		掰	手⿲分⿲手	手	手
	⿳	霩	雨⿳郭	雨	郭
		簠	𥫱⿳甫⿳皿	𥫱	皿
	⿴	麑	鹿⿴兒	鹿	兒
		侖	侖⿴⿰口	侖	口
2	⿸	邍	⿸辶备枭⊡	⿸	⊡	當第一個構字符號為⿸時，則起始字元為⿸，終結字元為⊡
		豐	⿸山⿰丰豆⊡	⿸	⊡
3	㊀	斀	㊀貝⿲攵	㊀	攵	當第一個構字符號為重疊符號時，則起始字元為重疊符號，終結字元為連續兩個非構字符號的第一個
	〇〇	虩	〇〇虎	〇〇	虎
	㊎	贔	㊎貝	㊎	貝
	㊐	燚	㊐火	㊐	火

構字式在資料庫中的表達

- 構字式可採用關聯式資料表儲存，並以「構字符號」及「部件序」兩個欄位來表達
- 「構字符號」用來表達構字式的類型，「部件序」儲存部件
- 不再拆分的基礎部件可視爲型◯，「構字符號」爲0，「部件序」爲基礎部件本身

構字式在資料庫中的表達（續）

- 構字式另有兩個相關的欄位「基礎部件序」、「基礎部件組」
- 基礎部件序可由部件的基礎部件序組合而得
- 而基礎部件組則由基礎部件重新依筆畫、筆順排序而得

構字式在資料庫中的表達（續）

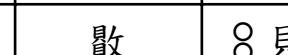
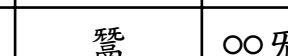
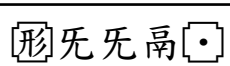
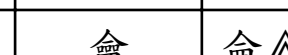
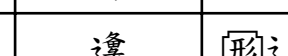
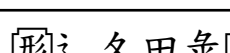
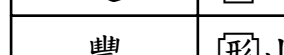
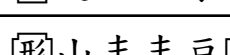
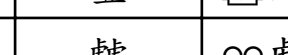
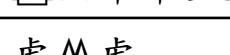
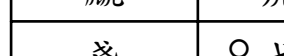
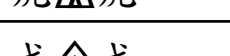
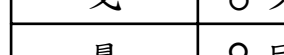
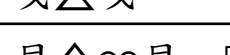
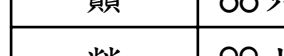
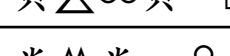
類型	字例	構字符號		部件序	基礎部件序	基礎部件組
1	顯	⿱	1	桑頁	又又又木一自八	一八又又又木自
	掰	⿱	1	手分手	手八刀手	八刀手手
	斂	⿱	1	⿰貝攵	貝貝攵	攵貝貝
2	霩	⿳	2	雨郭	雨亠口子阝	亠口子阝雨
	簠	⿳	2	竺甫皿	竺甫皿	皿竺甫
	鬻	⿳	2	○○乐鬲	乐乐鬲	乐乐鬲
3	麤	⿵	3	鹿兒	广丗匕匕臼儿	儿匕匕广丗臼
	龠	⿵	3	侖○○口	人一冊口口口	一人口口口冊
4	邍	⿸	4	辶备枭	辶夂田厶氷	夂厶辶氷田
	豐	⿸	4	山○○丰豆	山丰丰一口丷	一口山丷丰丰
5	虩	○○	5	○○虎	虍儿虍儿	儿儿虍虍
	戛	⿰	5	⿰戈	戈戈	戈戈
	贔	⿱	5	⿱貝	貝貝貝	貝貝貝
	燚	⿱	5	⿱火	火火火火	火火火火

構字式的正規化

- 在實際應用上，使用者表達的構字式可能和儲存在資料庫的不同，此時可利用基礎部件序或基礎部件組進行構字式的正規化
- 大多數的構字式皆可透過基礎部件序進行正規化，少數的構字式則需透過基礎部件組

構字式的正規化（續）

記錄在資料庫的構字式及使用者表達的構字式

類型	字例	資料庫的構字式	使用者的構字式
1	散		
2	鷙		
3	龠		
4	邍		
	豐		
5	虩		
	戛		
	贛		
	燚		

構字式的正規化（續）

字例	構字式	基礎部件序	基礎部件組	正規化欄位
贁	8貝⟁攵	貝貝攵	攵貝貝	
	[形]貝貝攵⊡	貝貝攵	攵貝貝	基礎部件序
鬵	○○兂⟁鬲	兂兂鬲	兂兂鬲	
	[形]兂兂鬲⊡	兂兂鬲	兂兂鬲	基礎部件序
龠	侖⟁○○○口	人一冊口口口	一人口口口冊	
	亼⟁○○○口⟁冊	人一口口口冊	一人口口口冊	基礎部件組
邍	[形]辶备彔⊡	辶夂田彑氺	夂彑辶氺田	
	[形]辶夂田彔⊡	辶夂田彑氺	夂彑辶氺田	基礎部件序
豐	[形]山○○丰豆⊡	山丰丰一口丷	一口山丷丰丰	
	[形]山丰丰豆⊡	山丰丰一口丷	一口山丷丰丰	基礎部件序

構字式的正規化（續）

字例	構字式	基礎部件序	基礎部件組	正規化欄位
虩	⿰虎虎	虍儿虍儿	儿儿虍虍	
	虎⿰虎	虍儿虍儿	儿儿虍虍	基礎部件序
戛	⿰戈戈	戈戈	戈戈	
	戈⿱戈	戈戈	戈戈	基礎部件序
贔	⿱⿰貝貝貝	貝貝貝	貝貝貝	
	貝⿱⿰貝貝	貝貝貝	貝貝貝	基礎部件序
	⿴形貝貝貝⿴囗·	貝貝貝	貝貝貝	基礎部件序
燚	⿱⿰火火⿰火火	火火火火	火火火火	
	炎⿰炎	火火火火	火火火火	基礎部件序
	⿰火⿰火	火火火火	火火火火	基礎部件序
	⿰火⿱⿰火	火火火火	火火火火	基礎部件序

3.4風格碼的制定及使用

- 風格碼是構字式的延伸，構字式是利用字形結構來區分字形的字形碼，而風格碼則是利用出處來區分同一個字形而風格迥異的字體碼
- 構字式適用於楷書，風格碼則適用於金文、甲骨文等古漢字，制定風格碼的目的在於解決古漢字重文的編碼問題

風格碼的形式

- 「形字形體出處⊙」
- 「形字形體出處;n⊙」

n即表該字重複出現時的編號

風格碼舉例

形中體說文古文

形中體說文籀文

形員體集成6432

形員體集成6432;2

古漢字出處

古漢字	出　　處
甲骨文	合集(甲骨文合集)、屯(小屯南地甲骨)、英(英國所藏甲骨集)、懷(懷特氏等所藏甲骨集)。
金　文	集成器號(約12,000件)
楚系簡帛文字	牌406、仰25、常2、望1、望2、天卜、天策、雨21、馬1、磚370、秦1、秦13、秦99、范27、滕1、包2、信1、信2、曾、帛甲、帛乙等出土墓號及簡號。[3]
小　篆	說文、說文或體、說文古文、說文籀文、說文篆文、說文俗字、說文奇字

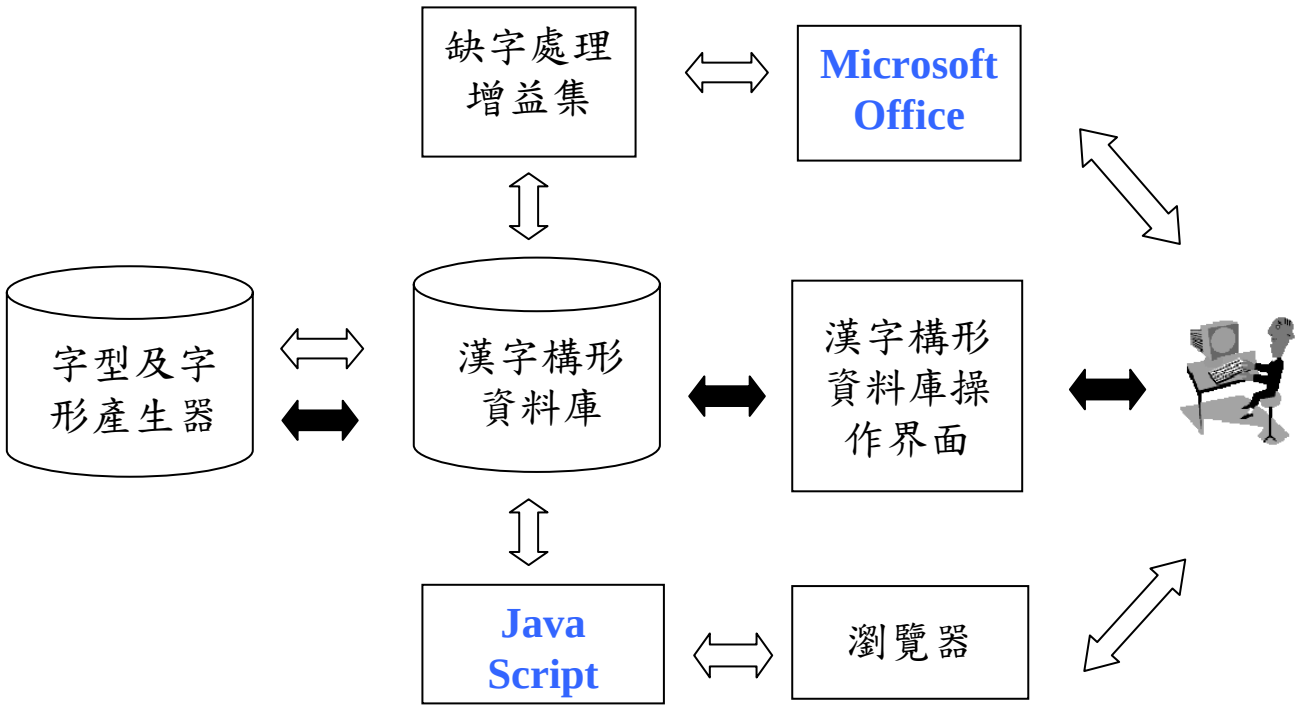
第4章
漢字構形資料庫的應用

- 4.1系統下載及安裝
- 4.2系統架構
- 4.3漢字構形資料庫
- 4.4缺字字型
- 4.5使用界面
- 4.6缺字增益集
- 4.7網頁缺字

4.1系統下載及安裝

- 中研院漢字部件檢字系統是以漢字構形資料庫爲核心，下載網址爲 http://cdp.sinica.edu.tw/cdphanzi/
- 詳細的安裝程序可參考 http://cdp.sinica.edu.tw/cdphanzi/setup.htm

4.2系統架構



應用漢字構形資料庫來解決缺字問題

缺字處理流程

1. 使用者可透過漢字構形資料庫的操作界面，查詢所需缺字之構字式
2. 構字式亦可透過字形產生器直接產生缺字

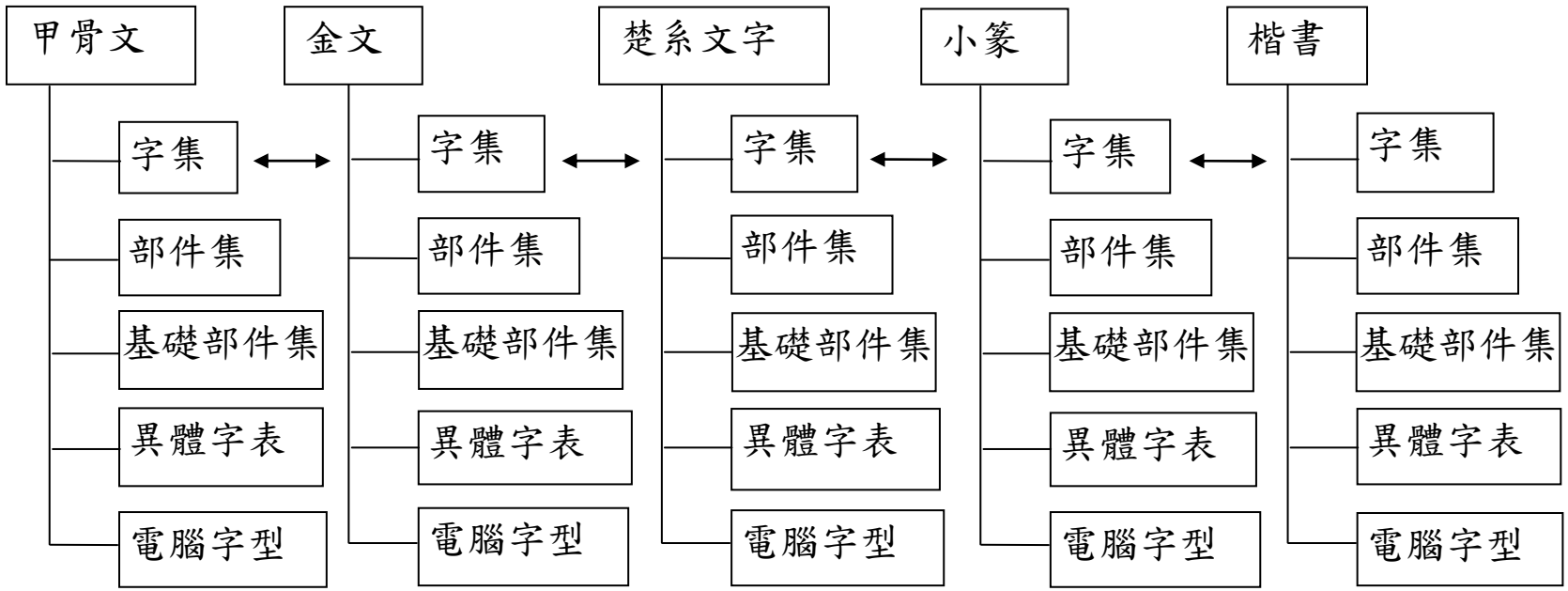
缺字處理流程（續）

3. 使用者可在Microsoft Office文件中嵌入構字式，再透過缺字處理增益集在文件中顯示缺字
4. 網頁管理者亦可在網頁中嵌入構字式與JavaScript程式碼，在使用者開啓並讀取網頁時，即連至主機的缺字處理API，並由主機傳出缺字的字形圖片至網頁

4.3漢字構形資料庫

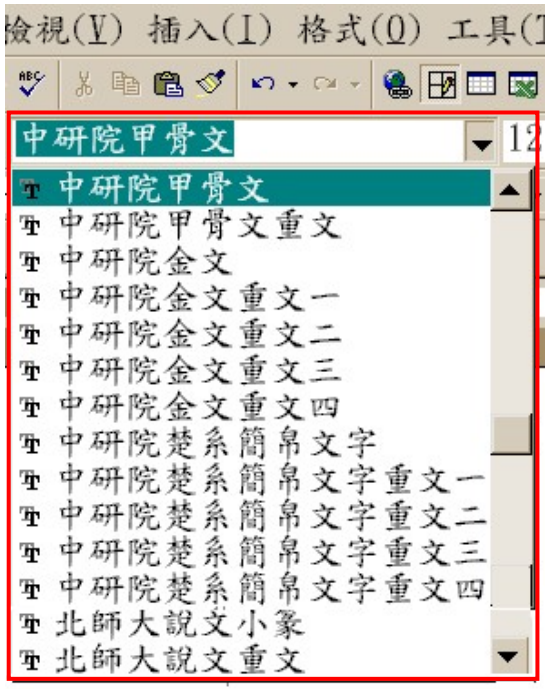
- 漢字構形資料庫2.53版是由甲骨文、金文、楚系簡帛文字、小篆及楷書構形資料庫組合而成
- 每個資料庫都至少包含「檢字表」、「異體字表」兩個資料表

4.3漢字構形資料庫（續）

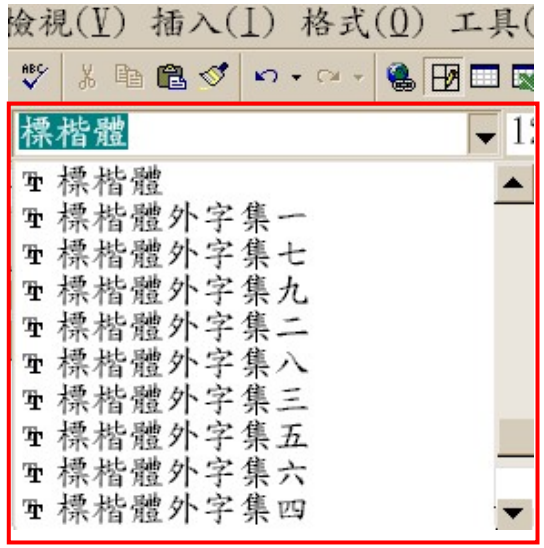


漢字構形資料庫的組成

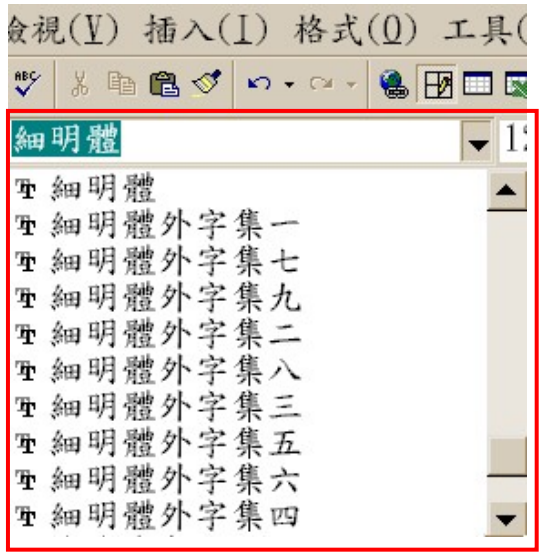
4.4缺字字型



古漢字字型



標楷體外字集



細明體外字集

古漢字字型及缺字字型

4.4缺字字型（續）

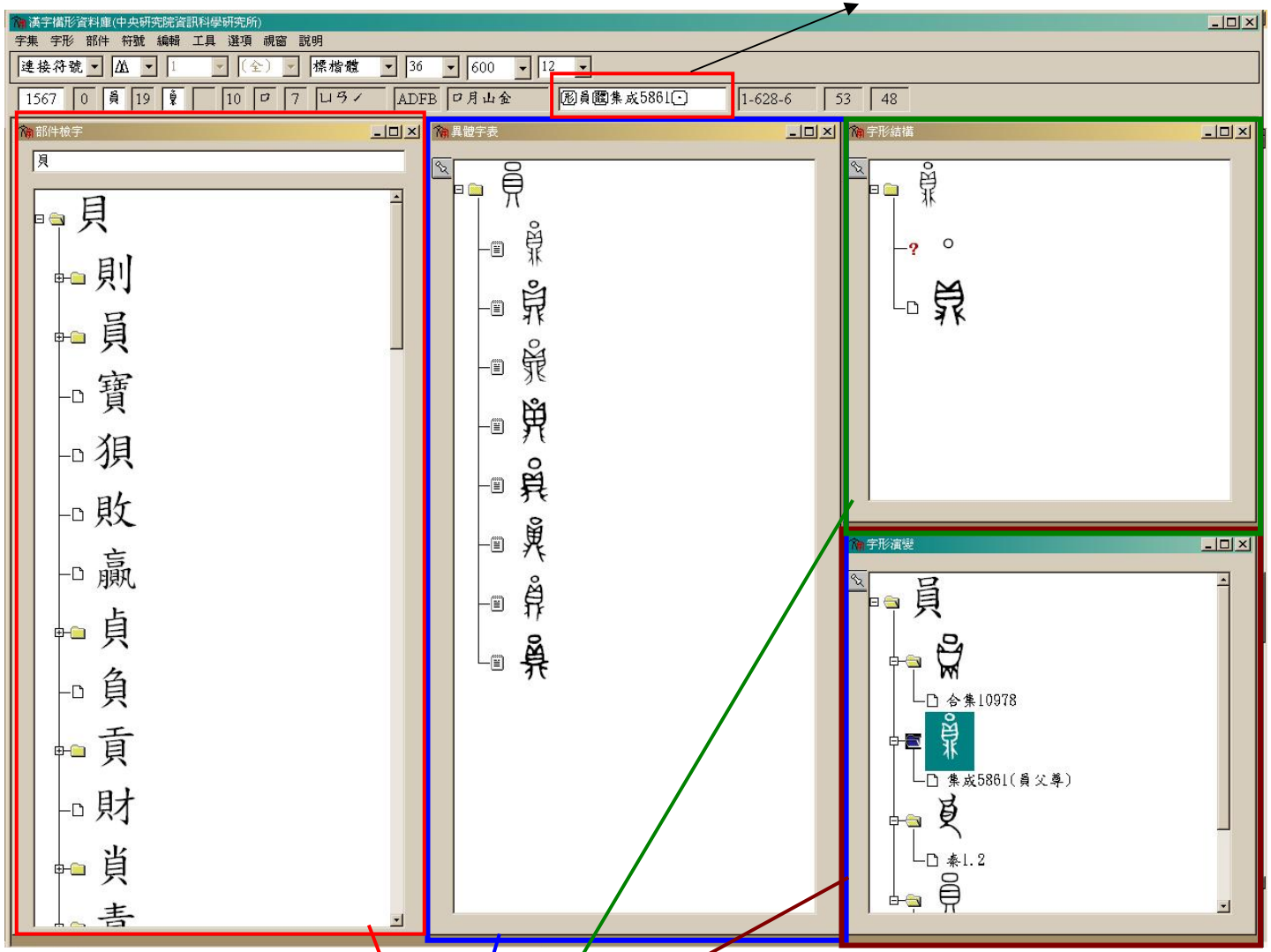
- 每套缺字字型只用到五大碼的外字區，最多只能編到6,217個字
- 細明體部件外字爲系統預設外字集，標楷體部件外字目前只對應到標楷體

4.5使用界面

- 系統安裝成功之後，在「開始」功能表中可找到「缺字公用程式」，並由此啓動「漢字構形資料庫」
- 漢字構形資料庫使用界面的主功能表分成「字集」、「字形」、「部件」及「編輯」等次功能表

異體字表

構字式或風格碼顯示欄位



部件檢字

字形結構

字形演變

所有視窗中的字形資訊都是連動的

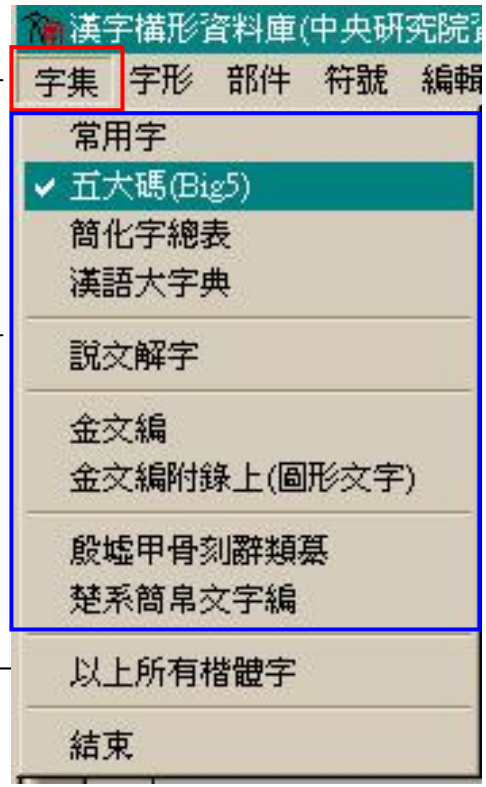
漢字構形資料庫的使用界面

4.5使用界面（續）

開啓「字集」次功能表

可由字集選單中選擇想要檢索的字集

亦可檢索所有的楷體字形

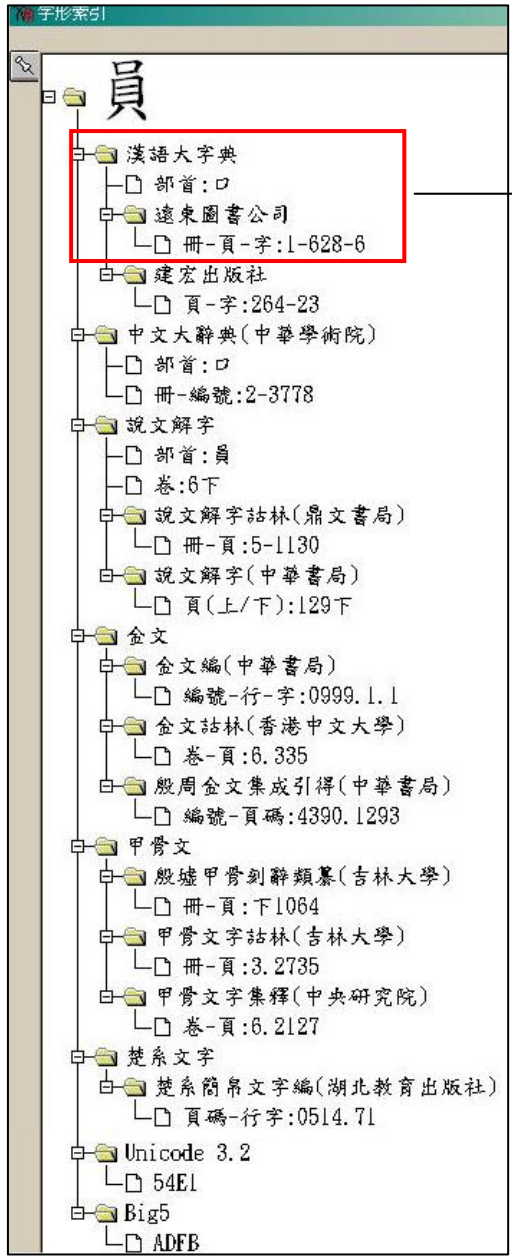


漢字構形資料庫字集次功能表

開啓字形次功能表

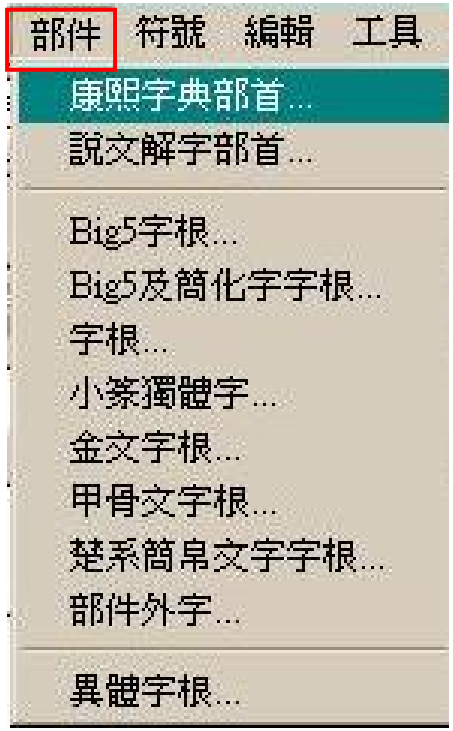
開啓字形索引視窗

漢字構形資料庫字形次功能表

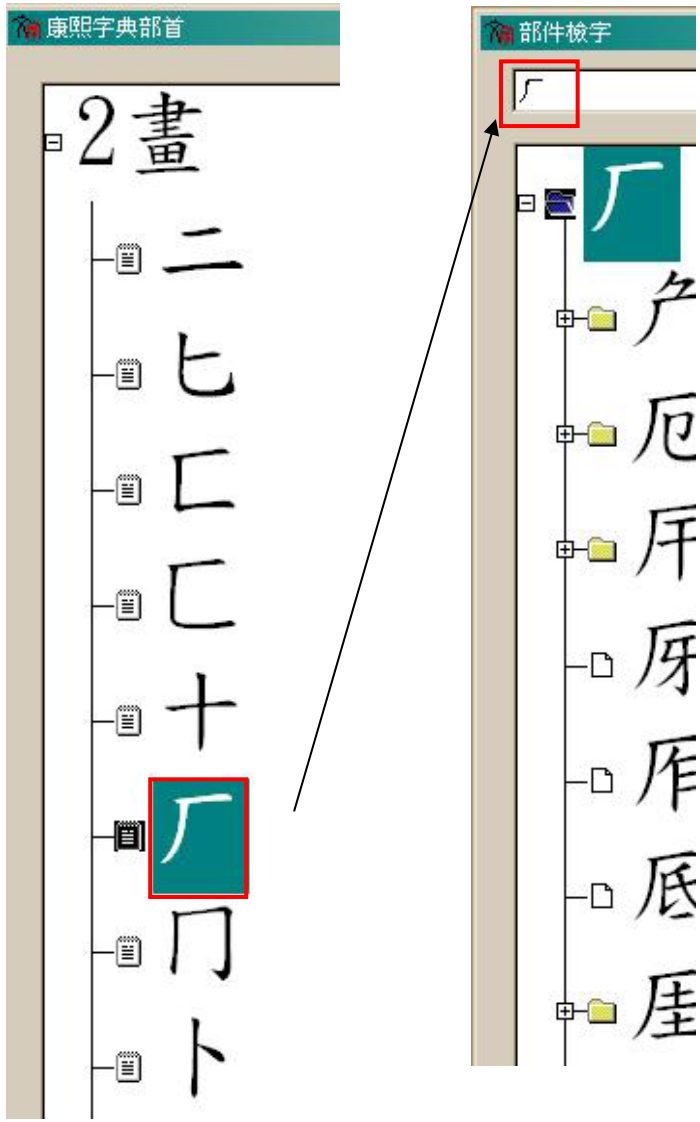


「員」字在遠東圖書公司出版的《漢語大字典》中的索引資料

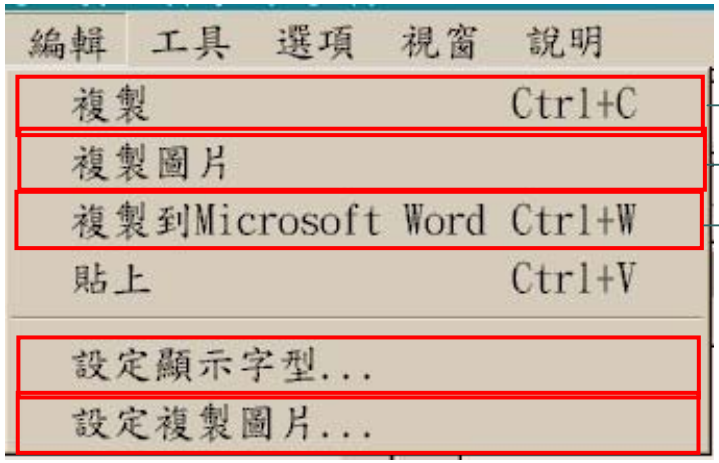
漢字構形資料庫的字形索引



漢字構形資料庫部件次功能表



從「康熙字典部首」視窗中尋找適合的檢索部件用以檢索

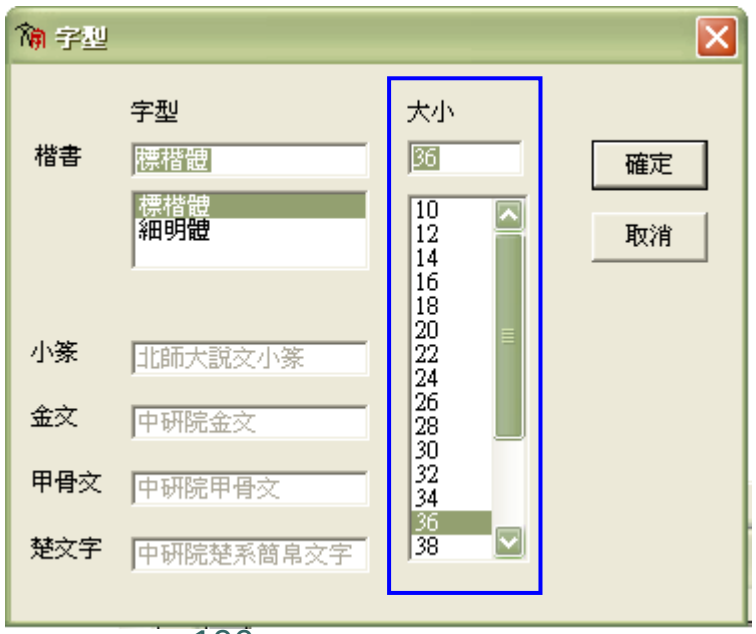


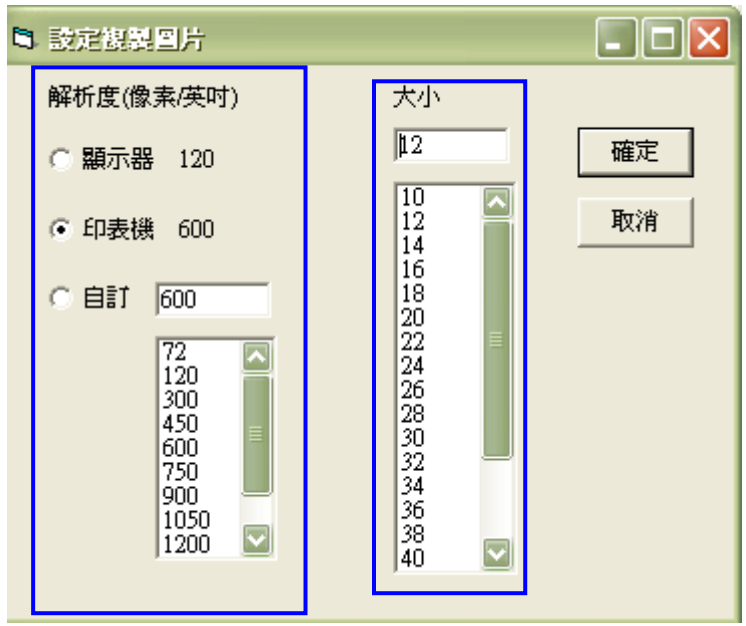
只複製字形或構字式

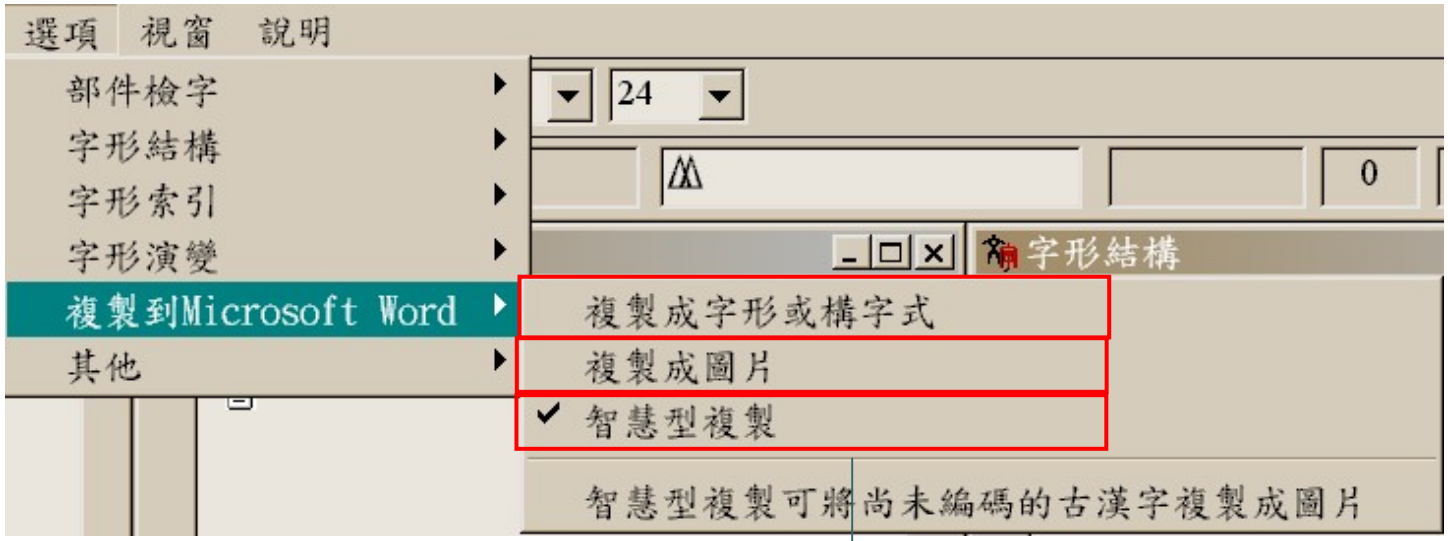
只複製字形圖片

將點選的字形直接貼至Word（需配合選項中的功能使用）

漢字構形資料庫編輯次功能表







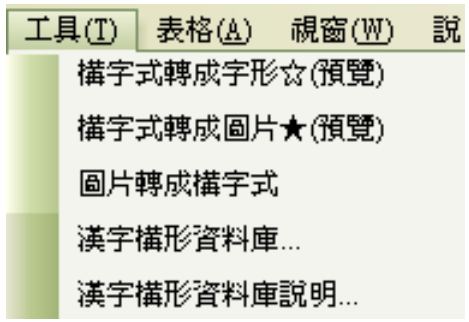
漢字構形資料庫選項次功能表

電腦字集已收錄的字，按Ctrl-W會複製字碼，缺字則複製圖片

4.6缺字增益集

- 安裝漢字構形資料庫之後，即會在Microsoft Word建立缺字增益集，這個增益集包含「構字符號」工具列及缺字處理次功能表

構字符號工具列

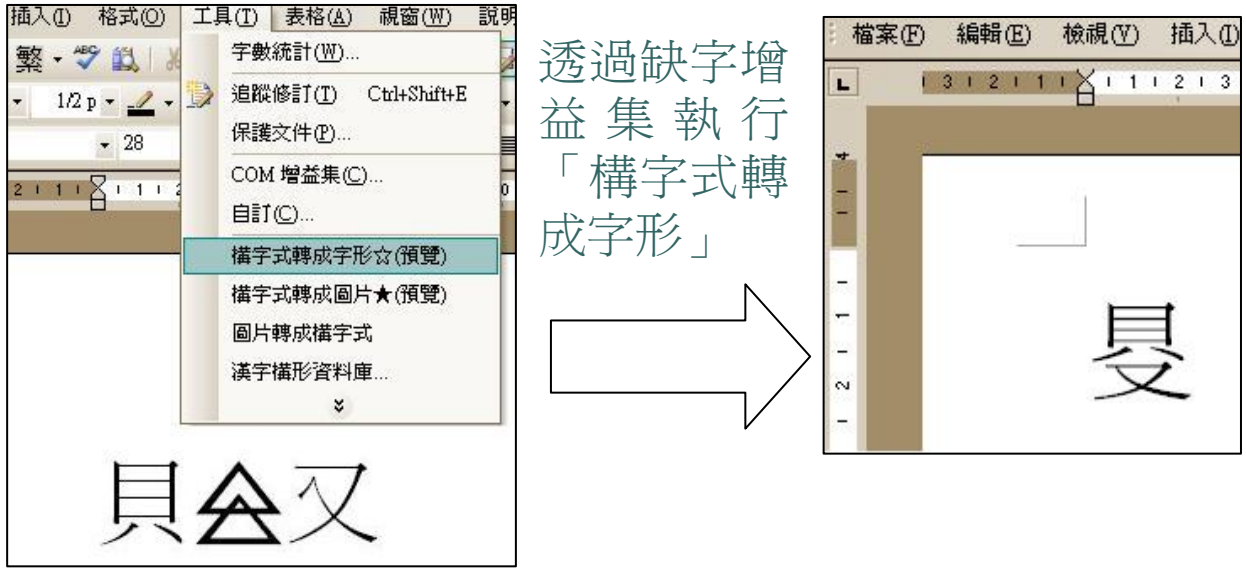


缺字處理功能表

4.6缺字增益集（續）

- 在Word中缺字的輸入，可以構字式或風格碼來表達。
- 使用者輸入構字式或風格碼的文件，稱爲「原始文件」。
- 執行「構字式轉成字形」或「構字式轉成圖片」，可將構字式或風格碼轉成字形或字形圖片，並存爲新增的文件，這個新增的文件稱爲「缺字預覽文件」

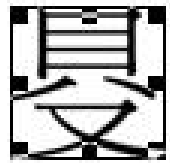
「構字式轉成字形」是將構字式表達的缺字以字型檔的方式顯示

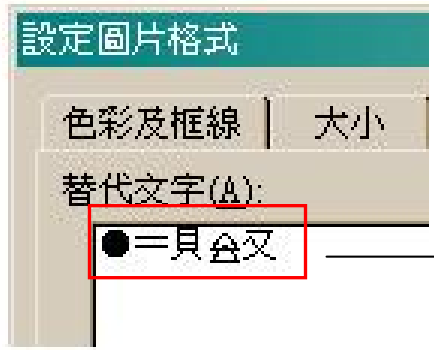


透過缺字增益集顯示缺字字形

「構字式轉成圖片」是將構字式表達的缺字以圖片的方式顯示

透過缺字增益集將構字式轉換成字形圖片



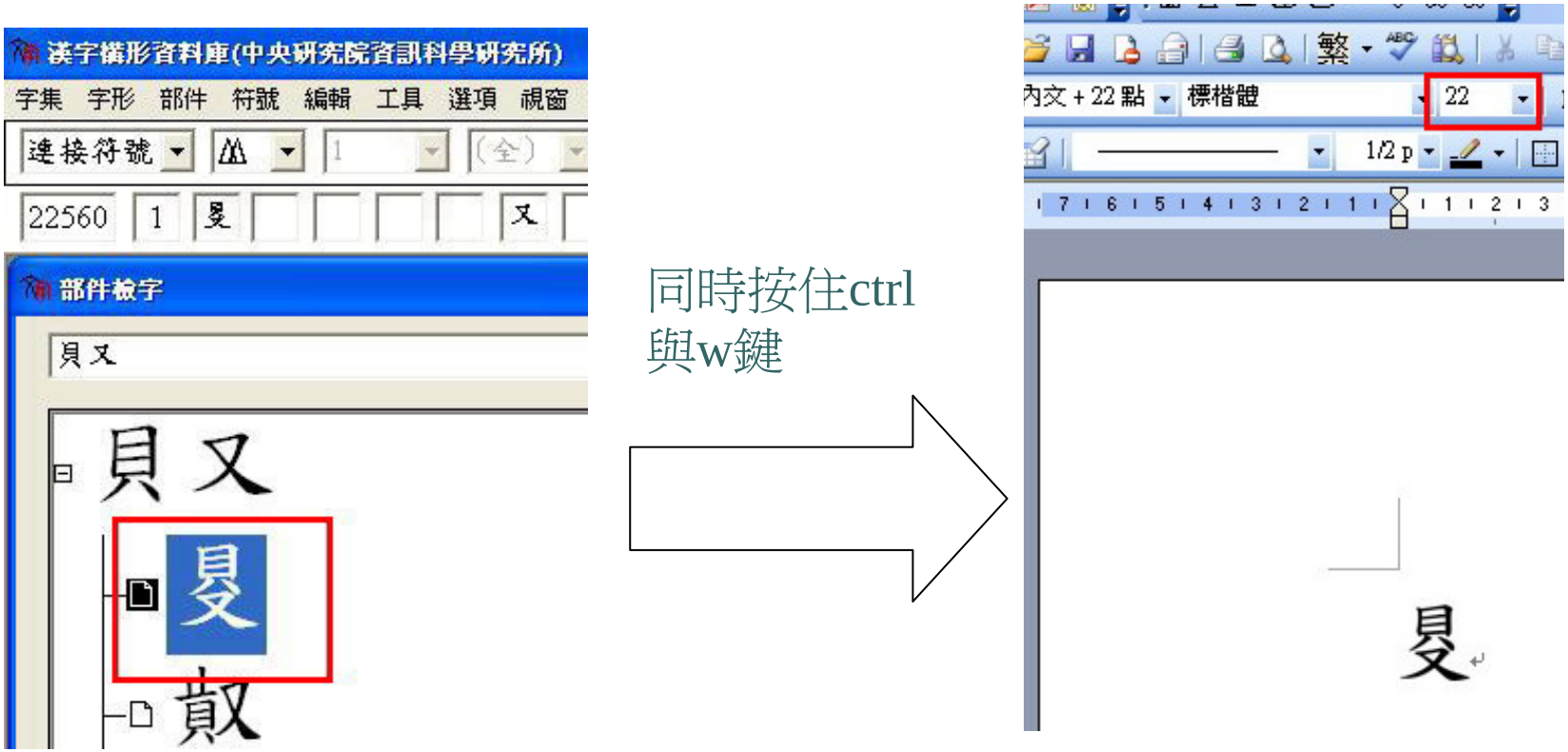


字形圖片內嵌有對應的構字式

透過缺字增益集顯示缺字圖片

自動貼圖功能

- 由於構字式可當交換碼，所以原始文件適合交換及修改。缺字預覽文件只供閱讀及列印，修改時必須回到原始文件進行
- 在2007年8月釋出的漢字構形資料庫2.5版中，新增了自動貼圖至Microsoft Word的功能，在漢字構形資料庫中看到的任何字形，都可以利用快速鍵Ctrl-W，將字形圖片自動貼至Microsoft Word中

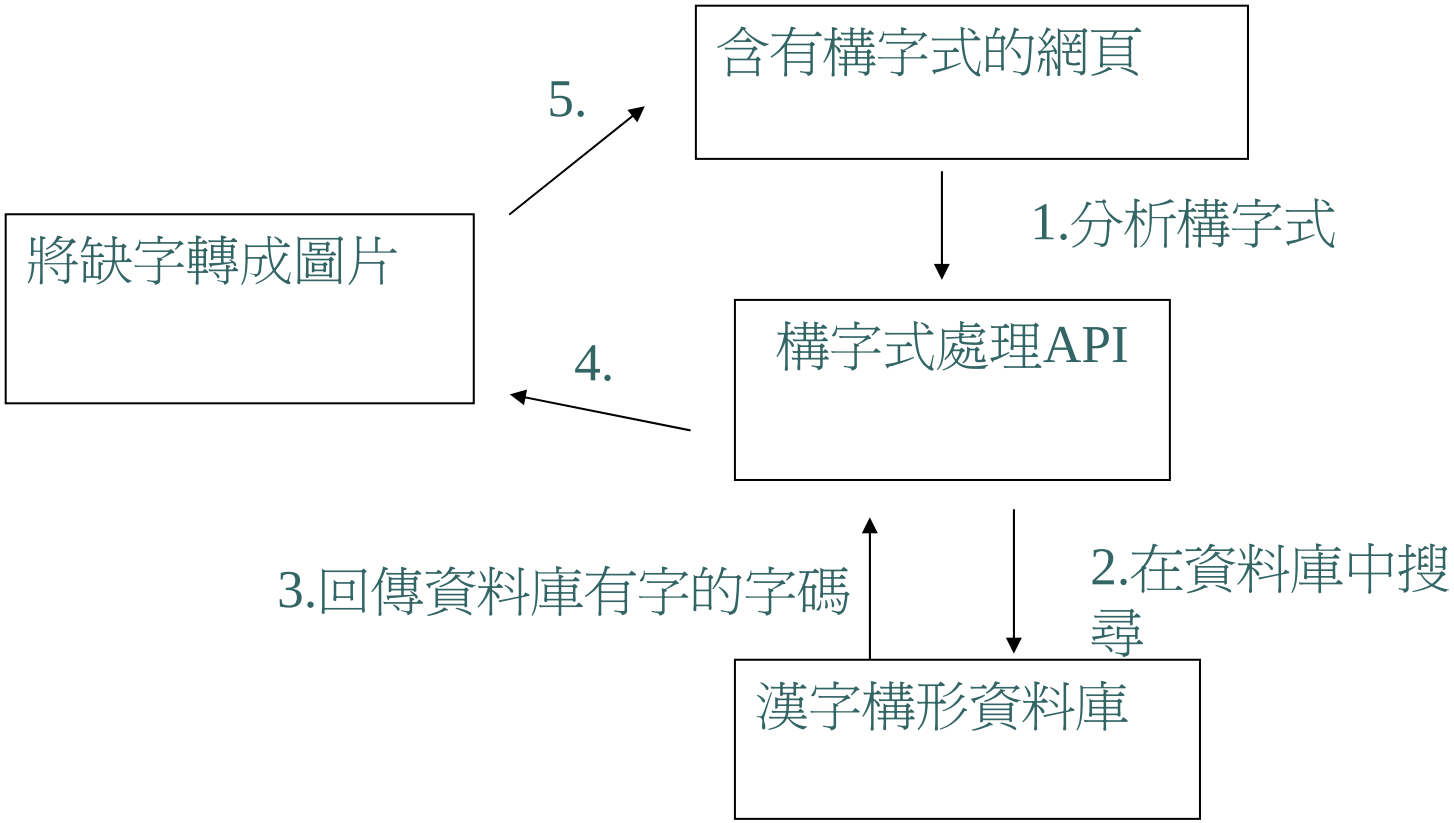


利用快速鍵自動貼圖至Microsoft Word

4.7網頁缺字

- 網頁缺字技術目前由數位典藏技術發展組支援，詳細的說明可參考數位典藏技術發展組設計的缺字系統網站（http://char.ndap.org.tw/index.htm）

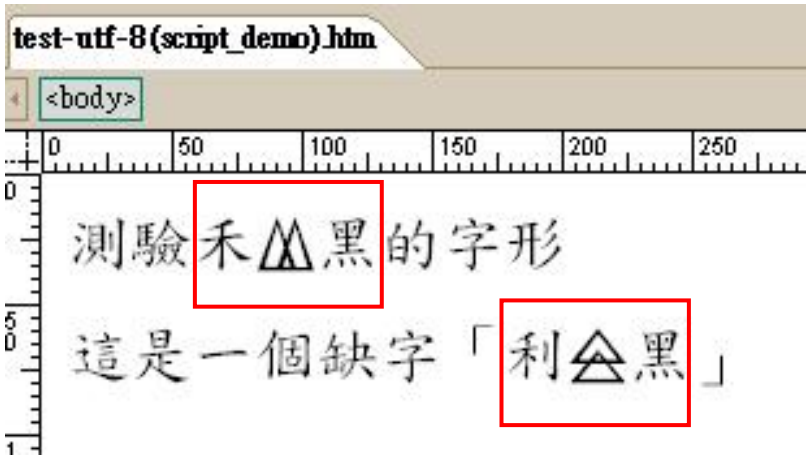
4.7網頁缺字（續）



網頁缺字呈現技術

具體的建置網頁缺字方式

- 在輸入網頁資料時，必須先以構字式表達電腦缺字



以構字式表達網頁上的缺字字形

具體的建置網頁缺字方式（二）

- 接著在網頁的程式碼中加入JavaScript程式碼

```
1  <html>
2
3  <head>
4
5  <meta http-equiv="Content-Type" content="text/html; charset=utf-8">
6
7
8  <SCRIPT src="http://char.ndap.org.tw/API/ics.js" language="javascript"></SCRIPT>
9
10 <title>test</title>
11 </head>
```

網頁缺字需加入之程式碼

```
67 </body>
68 </html>
69
70
71
72 <SCRIPT LANGUAGE="JavaScript">
73 <!--
74
75 processPage('red','20','','','DFKai-sb');
76 //-->
77 </SCRIPT>
78
```

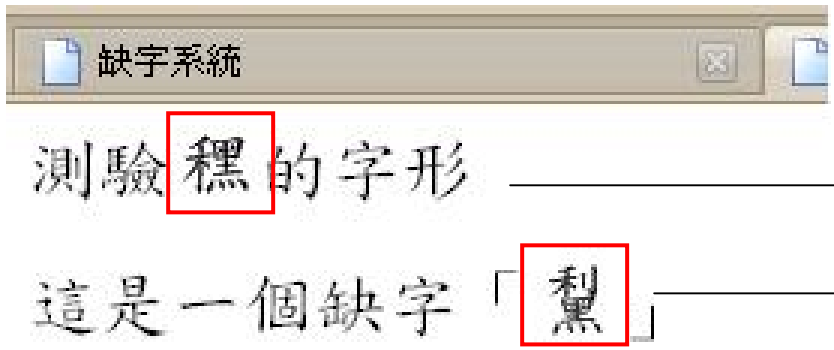
轉換整個頁面的構字式

此函式可供使用者變化整個頁面或是區塊上欲顯示缺字的顏色、字形大小及字體

```
67 </body>
68 </html>
69
70
71
72 <SCRIPT LANGUAGE="JavaScript">
73 <!--
74 processObject(document.getElementById('t4'),'black','18','','','DFKai-sb');
75 processPage('red','20','','','DFKai-sb');
76 //-->
77 </SCRIPT>
78
```

轉換區塊的構字式

具體的建置網頁缺字方式（續）



由缺字處理API傳出的缺字字形圖片，前後文則是正常的文字

網頁缺字的顯示結果

第5章
漢字構形資料庫的展望

- 5.1漢字構形資料庫的應用現況
- 5.2建立文字學入口網站

5.1漢字構形資料庫的應用現況

1. 目前數位典藏與數位學習國家型科技計畫至少已有23個資料庫採用，其中規模最大的爲中央研究院史語所漢籍電子文獻資料庫
2. 以公眾授權模式提供「中研院漢字部件檢字系統」的原始程式碼及相關資料，釋出給大眾使用
3. 推動漢字構字標準，撰寫「中文字構形識別序列」標準草案，並函送標檢局審議

改善空間

1. 漢字構形資料庫目前只著重在字形知識的整理，建立一個形、音、義俱備的漢字知識庫，仍是我們長遠的目標
2. 漢字構形資料庫的缺字解決方案，採用單位仍侷限於中央研究院及數位典藏與數位學習國家型科技計畫，仍有推廣的空間

改善空間（續）

3. 部件檢字雖然比部首檢字便利，比筆畫檢字有效率，但仍應再增加部首、筆畫等其他檢字方式，以求完備
4. 下載人次增加有限。目前每天的下載人次約10次，總下載人次爲15,788次（截至2009年7月27日止）。若要增加使用人次，漢字構形資料庫應再開發中文簡體、英文、日文、韓文等各國語言版，或是網路應用版

5.2建立文字學入口網站

- 文字學入口網站主要是利用漢字的形、音、義知識來檢字，並提供漢字字典網站的連結
- 這樣一個入口網站不同於Google、Yahoo等用關鍵詞來找網頁的入口網站，而這樣的需求也非Google、Yahoo等入口網站所能達成

建立文字學入口網站應有的特色

1. 多語的使用界面：除了繁體、簡體中文、英文外，至少也須再兼顧其他漢字區的使用者，因此日文及韓文界面也須列入考量
2. 多樣的檢字方式：
   - 字形檢字包括部首、筆畫、部件等
   - 字音檢字包括注音、拼音、聲韻等
   - 字義檢字包括英文、日文、韓文等
   - 字碼檢字包括Big5、Unicode、CCCII、CNS11643等

建立文字學入口網站應有的特色（續）

3. 可檢索不同字集：這些字集包括常用字、簡化字，甚至甲骨文、金文、小篆等古漢字
4. 可解決缺字問題：無論使用者電腦中文字碼收錄字數的多寡，即使未安裝漢字字型，入口網站都應能正常顯示古今漢字

建立文字學入口網站應有的特色（續）

5. 提供字典網站連結：入口網站應在使用者找到所要的字後，提供現有的字典網站連結，以節省重複檢字的時間，並提供使用者更多的漢字參考資訊

文字學入口網站之效益

1. 推動字書上網：入口網站提供的檢字及缺字技術可降低字典上網的門檻，而入口網站提供的連結又可增加字典的能見度，增加現有字典上網的意願，同時也有助於線上字典的編纂

文字學入口網站之效益（續）

2. 建立漢字知識庫：透過網際網路集合眾人之力共同推動並予以整合，以建立完善的漢字知識庫
3. 便利漢字知識的擷取：目前的入口網站主要透過關鍵字及關鍵詞來搜尋網頁，以漢字爲主的網頁不易突顯；相形之下，文字學入口網站則只提供這些網頁的連結，因此更有利於使用者擷取到這些漢字知識

文字學入口網站之效益（續）

4. 闡揚漢字文化：入口網站所提供的連結，除了字典網站外，也包括其他和漢字相關的文化、藝術網站，這些連結及網站可用來闡述及發揚漢字文化