# 數位典藏系統缺字處理及應用

黃國倫 蕭人豪 李家豪 陳心渝

中央研究院資訊科學研究所

{ kulun, jenhao, chahao, kwakwai8}@iis.sinica.edu.tw

**摘要**

自『數位典藏國家型科技計畫』執行以來，已有大量的典藏品進行數位化工作，並且產出包括 Metadata 資料及多媒體的檔案。然而在 Metadata 建置時，常會遭遇到漢字資料處理時，部份的罕用字或古字的字形是系統交換碼所沒有的，因而無法被著錄至典藏系統中，也就是所謂的「缺字問題」[1]。為了解決缺字問題，中央研究院資訊科學研究所文獻處理實驗室發展出漢字構形資料庫，在資料庫中字形是利用部件及字根的組合方式來表達，故透過有限的部件及構字符號，即可組出無限的字形，此稱為構字式[2]，因而得以解決缺字問題。故本文章將針對典藏系統如何整合構字式之處理技術，並利用漢字構形資料庫中所建立的漢字字形提供數位典藏系統運用在缺字輸入、識別、查詢和後續的處理，以達成典藏資料之著錄、流通、分享、整合與再利用。所以不但可以解決一般自行造字所產生的問題外，也可實際整合至目前既有典藏系統中，因此可做為典藏單位或相關技術人員缺字處理時之參考。

關鍵字：數位典藏、缺字、構字式、漢字構形資料庫

## 1. 前言

在漢字數位化過程中，常會遇到系統字元編碼中並無法包含中文罕用字、古字，原因在於西方文字為拼字方式僅使用有限的字母即可表示所有的字，是屬於封閉式的字母集合（close set），而漢字為表意方式，自古迄今隨著時空環境改變，衍生有字音、字義、字形之變化，且漢字不斷會有新字、新詞產生，而這些字卻無法完全包含在系統交換碼中，因而無法被著錄，此即為缺字問題。尤其在處理古代文獻時，此問題更為嚴重。為了解決漢字數位化問題，常用的方法是利用使用者造字區內自行新增所需之缺字，但此方法在面對資料檢索及交換時會遭遇資料錯誤或無法讀取等問題，故並未徹底解決缺字問題，再加上當使用者造的字越來越多時，複雜的管理問題也隨之而來，例如：

1. 大量的缺字資料整理及造字工作，將大幅增加著錄成本。
2. 造字數量累積越多時，便難以人工化方式管理。
3. 使用者造字區編碼空間有限，缺字數量同樣受到限制。
4. 各典藏單位或系統間自行造的字可能會出現相同編碼，造成資訊分享的困難。

此外，當協助各個典藏單位建置典藏系統時，不同的內容主題對於缺字之需求欄位也會有所差異，例如在動物、器物、書畫、金石拓片、善本古籍、考古主題中的資料庫對於缺字需求欄位如下：

1. **動物-台大動物學生態模式資料庫**：以界、門、綱、目、科、屬、種之生物分類規則為主，生物領域知識為輔，記錄各物種的特性描述、標本、採集地等相關資訊，為具有物種分類架構管理、多媒體管理與呈現、安全控管之動物學數

位典藏系統。

2. **書畫-故宮書畫典藏資料庫**：典藏故宮博物院書畫處所保存之珍貴文史資料、書籍、畫冊之數位化管理系統。
3. **金石拓片-中研院史語所青銅器全形拓資料庫**：針對中研院傅斯年圖書館相關的拓片收藏予以有系統的數位化保存，其中包括了銅器全形拓一千兩百件、國家圖書館七百二十件、以及北京圖書館七百多件。
4. **善本古籍-中研院傅斯年圖書館善本古籍資料庫**：將傅斯年圖書館收藏之宋、元、明及清初刊本、稿本、名人之批校本、手抄本、繪寫本之收藏作品在數位化後以系統化方式管理保存。
5. **考古-中研院史語所漢代簡牘資料庫**：將中研院史語所文物陳列館所收藏之漢代簡牘中所記錄自西漢中晚期到東漢初期，當地軍民之軍事、法律、教育、經濟、信仰以及日常生活情形之內容及簡牘影像予以數位化保存

| 內容主題 | 數位典藏系統 | 整合缺字之主要欄位需求 |
|---|---|---|
| 動物 | 台大動物學生態模式資料庫 | 昆蟲分類、昆蟲中文學名 |
| 書畫 | 故宮書畫典藏資料庫 | 書畫作者、印記擁有者、主要題名、釋文、題跋作者、題跋內容 |
| 金石拓片 | 中研院史語所青銅器全形拓資料庫 | 人名、拓印之青銅器資料、青銅器器名 |
| 善本古籍 | 中研院傅斯年圖書館善本古籍數位典藏系統 | 作者、書目題名、題記、全文、釋文 |
| 考古 | 中研院史語所漢代簡牘資料庫 | 釋文、說明、校記 |

綜合上表，可發現缺字常發生在兩種屬性的內容中，一為「人名」，中國人自古迄今命名一向慎重，除了寓意吉祥、好寫、好記、好念外，尚需配合生辰八字，五行命理，故怪字層出不窮；二則為「藏品本身之文字資訊」，如古書、古物上所記載的罕用字與古字等，皆是目前不常被使用的文字。因此，在缺字處理的過程當中，會有下列的要求：

1. 文字是從古至今演變而來，所以相同一個漢字會因不同時間或空間有相異字形（glyph），如：楷書與小篆為時間之差異、繁體與簡體則為空間之差異，故必須能還原字形的結構，並保留文字間的關係。
2. 缺字資料須如同一般文字，能在典藏系統中輸入、描述、識別、查詢和後續的處理工作。
3. 缺字處理能以最低的系統整合成本，達成在網路環境下的各種系統操作功能。

在本研究中主要以構字式為基礎，利用中央研究院資訊科學研究所文獻處理實驗室所發展的漢字構形資料庫來收錄數位典藏中所遭遇的缺字，並且在典藏系統中透過不同的處理流程，來克服缺字資料如何在系統中著錄、顯示與查詢等問題，即使在網路環境下仍然可以操作包含缺字之網頁。因此可做為典藏單位或相關技術人員缺字處理時之參考依循。

## 2. 文獻探討

早期的電腦系統是發源於美國，因此最早的編碼系統也是發源於此。由於這樣的編碼系統僅包含數字、26 個英文字母 (包括大小寫)、標點與其他特殊符號、外加一些電腦系統的控制碼而已。然而亞洲地區大多為表義文字，尤其漢字的字集依古今的變異、專

業與應用環境的差異等而有字數、字形、字音以及字義上的變化。若就字數而言，即已不適合作固定數量的限定，因此系統的編碼空間對亞洲語系而言是不足夠的。以目前常被使用的繁體中文編碼 BIG5 為例，僅收納了一萬三千多個常用字，所以許多的漢字皆無法被表示，因此需要有更好的方法來解決缺字問題。

### 2.1. Unicode

由於 Unicode[3]在其編碼中同時容納了全世界各種語言的字元和符號，因此已成為國際常用的交換碼標準。目前 Unicode 在漢字的支援方面目前已經定義超過七萬多個字元，收納的字遠多於 BIG5 (約收藏一萬三千多個字)，且收納字的範圍還在繼續增補中，因此也的確解決了某些層次字形編碼不足的問題。並且在許多系統支持下，在資訊交換上也的確有其便利性。因此若在數位典藏中使用 Unicode，可以降低缺字發生的可能性。但當面對大量的古書資料時，由於包含的古文字數量相當驚人，並且在不同的時間存在著同義異形的相異字形(如：楷書、小篆)， 而 Unicode 所收錄的是以常用字為主，因此根本無法包含所有的古文字，即使將每一個字皆擴充進去，那所占的編碼空間將會相當可觀。再著，對於有著同樣意義的字，分別指派給不同的編碼，則未來在進行資料檢索時，將無法搜尋到所有的同義字，因此系統設計上將會增添許多新的難題。

### 2.2. 使用者造字

一般解決缺字常見的做法是在使用者造字區中，建立缺字的字形及設定輸入方式，待完成後，即可在系統中進行著錄之動作，此步驟及使用上算是簡易。但隨著缺字數量越來越多，此種方法的問題也就越多，例如：要整理出所需的缺字及為每個缺字建立字形、編碼即須投入很高的人力成本，再加上系統缺字資料的著錄，以及未來造字區可容納的字數限制所造成的擴充問題；再者若要將包含缺字的文件傳給另一系統讀取或修改時，則該人員亦須先匯入同一份的造字檔才能正確顯示造出來的字。故整體看來，就使用上的確能解決了缺字在文件上的著錄及顯示的應用，但在數位典藏系統中，就面臨到資訊分享及交換的困難。

## 3. 缺字處理

數位典藏的缺字處理主要可依照使用者應用方面的需求及操作上的流程來加以區分，可將其分為構字式、著錄及顯示等三大部份。完整的流程，如圖1所示，首先典藏單位(Content Creation)將其相關的研究資料著錄到典藏系統 (Content Provision)中，之後使用者可經由瀏覽器來閱覽這些資料。在這樣的流程中，主要是在使用者輸入資料至系統及使用者透過瀏覽器來檢索資料時，須針對構字式加以處理，使缺字能被儲存至資料庫中，並且正確的顯示在使用端，而解決缺字問題。以下將詳細說明整個數位典藏系統缺字處理運作原理。

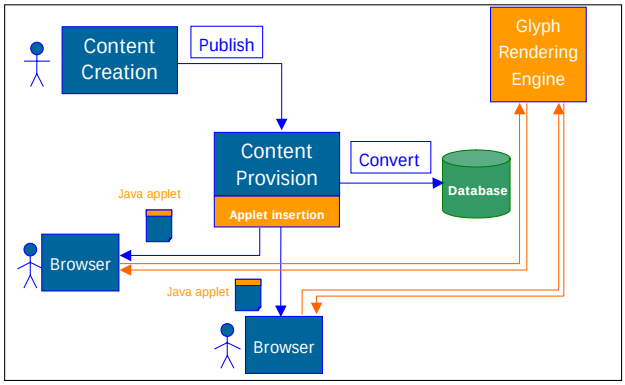


圖 1　數位典藏系統處理流程圖

### 3.1. 構字式

為了徹底解決現行漢字交換碼不足所造成的缺字問題，中研院資訊所文獻處理實驗室從漢字字形結構的拆分與分析中，利用有限的部件及字根的組合方式來表達任一漢字，此稱為構字式。例如『顯』，以構字式拆解的話，可拆分成「景」與「頁」兩個部件，其中為了表示部件與部件的連接關係，故定義了三類共計十三個的「構字符號」，故『顯』的構字式為『景⿰丆頁』。因此構字式是由部件和構字符號組成，且「構字符號」也是一般文字和缺字的辨識依據。

因此以構字式為基礎下，將收錄的漢字建立以Big5為編碼系統的『漢字構形資料庫』。目前漢字構形資料庫已收錄了楷體字形57,820個、小篆11,100個、金文932 個，異體字12,271組，所以當各典藏單位面對數位化所遭遇到缺字問題時，若使用漢字構形資料庫做為缺字的解決方案未嘗不是個成本較低、功能又較完備的好方法。

### 3.2. 典藏系統缺字著錄

在缺字著錄部份，最重要的是解決構字式輸入問題，目前採取的方法可以在使用者端安裝漢字構形資料庫工具[4]即可輸入構字式；或是直接由網路連結至缺字查詢主機[5]進行缺字查詢。此兩種查詢方式皆允許使用者透過筆畫(stroke)或部件作為條件來查詢出符合的缺字，而查詢結果頁面會將符合這部件的相關字全部列出，如圖 2 所示。若以網路缺字查詢為例，在結果中包括了以圖片顯示的缺字字形、以圖片顯示的部件和構字符號，以及完整的構字式內容。因此即使是未安裝過字型檔或輸入法工具的使用者皆可以輸入構字式。

當使用者將查詢出來的缺字構字式複製到系統著錄欄位後儲存即可完成著錄程序。目前已有許多數位典藏系統，連結到缺字查詢主機，方便使用者在著錄資料時即時進行缺字查詢，且不須固定在某些有安裝使用者端程式的電腦才可使用構字式輸入功能，故大大的提升這些典藏單位在著錄資料工作時的便利性。

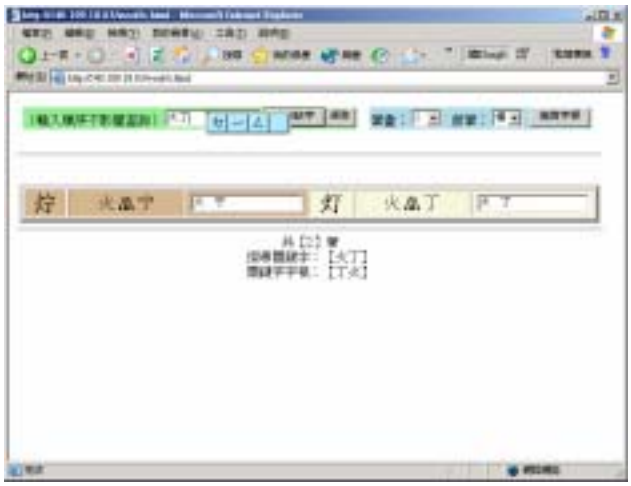

圖 2 缺字查詢網頁

當含有構字式的資料被著錄時，典藏系統中需要一個轉碼元件將構字式的資料進行編碼轉換，如圖 3 所示，這是由於構字式中的構字符號無法對應至 Big5。例如當使用者輸入『火⿰丆丁』時，須要將『⿰丆』符號轉換為 HTTP 跳脫格式(escapes)的表示法，所以在資料庫中會以『火丁』的方式被存入。如此當構字式在網頁顯示時，『』將會被瀏覽器還原成『⿰丆』，便能正確地被另一個負責缺字顯示的元件轉換。

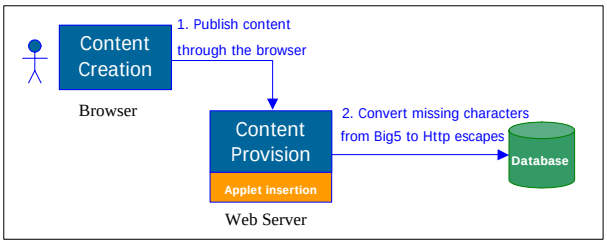


圖 3 使用者著錄流程圖

### 3.3. 網頁缺字顯示

在網頁顯示部份，須將資料中包含構字式內容轉換成圖片格式的缺字字形。當使用者開啟了含有缺字之網頁時，此網頁所包含一個元件，我們稱為 LiveConverter[6]，待資料載入完畢後， LiveConverter 會啟動並對資料內容進行判斷，將構字式轉換成 HTML 的圖片標籤(IMG TAG)，而圖片的 URL 連結到字形解譯引擎(Glyph Rendering Engine)同時傳入該字形相關參數，如字元編碼、大小及顏色。最後在使用者端觀看到的網頁，即是以圖片顯出的缺字資料，其流程如圖 4。目前已開發出 Java 版本的 LiveConverter 元件，故可利用 Java Applet 方式置於網頁中將所有構字式資料進行轉換，或是以 Java Bean 方式針對特定欄位資料進行轉換。

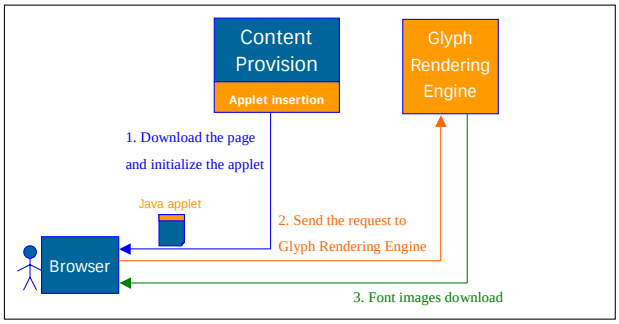


圖 4 網頁缺字顯示流程圖

以下兩張圖示分別顯示了在缺字轉換前後字體顯示上的差別。圖 5 可以看到缺字部份有提供讓使用者自行選擇缺字的大小及顯示顏色的功能，以藍色線圈起之處為轉換之前的缺字，在圖 6 則是缺字轉換後的顯示方式。

圖 5 缺字顯示轉換前

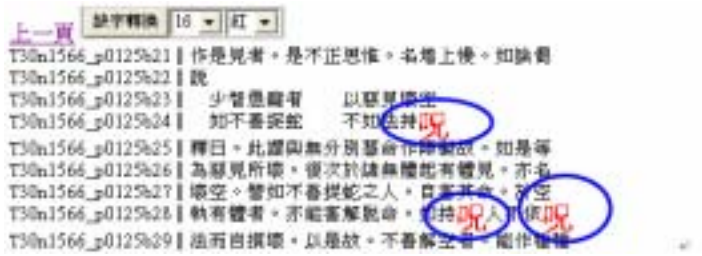

圖 6 缺字顯示轉換後

在網頁以圖片的方式來顯示缺字的最大好處是，使用者端並不需要再額外的安裝任何的軟體及字形檔，如此一來即大幅提升了使用上的便利性和使用者的使用意願。

## 4. 數位典藏系統缺字整合及應用

『傅斯年圖書館藏善本古籍數位典藏系統』[7]是數位典藏計畫下所建置之典藏系統。其開發的目的在將傅斯年圖書館藏善本圖籍有妥善的保存環境，以及提供讀者更方便、更迅速的檢索。由於傅圖所藏善本圖籍主要為宋、元、明及清初、稿本，以及名人批校本、寫繪本等，因此在數位化的過程中也遭遇到相當多古字的缺字問題。以下我們將以善本典藏系統導入缺字技術為例，來探討本論文中技術的可行性，以及分享整個缺字系統的使用者經驗。

### 4.1. 使用者『著錄』流程

當使用者在數位化善本古籍時，遇到缺字的著錄問題時，可以藉由輸入該缺字的構字式而獲得解決。比方說當使用者想要輸入『王  』這兩個字的時候（  是古文缺字），則可以在善本典藏系統中輸入『王龍廾』(如圖 7 所示)。而當含有構字式的資料被著錄時，典藏系統即會自動判別並將構字式中的構字符號進行對應轉換成跳脫字元，所以當使用者輸入『王龍廾』時，經過缺字系統的處理後，在資料庫中將會以『王龍廾』的內容被儲存，若無此步驟，

會造成構字符號因無法對應至 Big5 而轉換成「?」。這種缺字著錄方式的最大優點在於使用者只需要知道缺字的構字式，其他的著錄流程與系統操作與之前是完全相同的，因此完全沒有增加任何額外的負擔。如果使用者不清楚某個缺字的構字式 (在本例中即  這個古文缺字) 時, 則只需要使用缺字構字式查詢工具即可解決此問題。

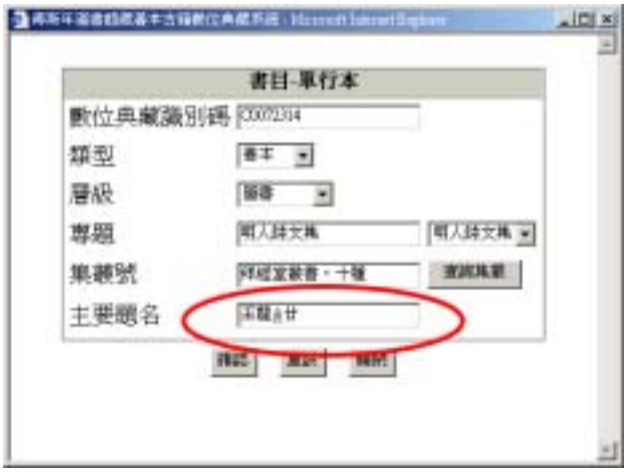


圖 7 缺字著錄

### 4.2. 整合缺字的『查詢』與『顯示』

為了增加典藏資料使用的方便性，『傅斯年圖書館藏善本古籍數位典藏系統』也提供了檢索功能讓使用者可以快速的從中獲取需要的相關資訊及內容。而當使用者輸入的檢索條件包含了缺字，比方說使用者想要查詢善本典藏系統的資料庫中是否存在題名為『王  』的資料時，依然可以透過前面相同的構字式著錄方式，使用者直接輸入包含構字式的內容做為檢索條件，如『王龍廾』。在檢索系統收到使用者的檢索條件後則將查詢條件轉換成包含跳脫字元的『王龍廾』直接到資料庫中進行搜尋即可找到符合的資料，如圖 8 所示。

圖 8 以構字式檢索

### 4.3. 缺字資訊的『交換』

從前為了解決缺字問題大部分都是使用系統造字來解決，但是這樣做的重大缺點在於限制了含有缺字文件的資訊交換。尤其數位典藏計畫的目標是將國家重要的文物典藏數位化，建立國家數位典藏，以促進人文與社會、產業與經濟的發展。因此資訊的交換絕對是非常重要的考量因素。

在整合缺字處理的『傅斯年圖書館藏善本古籍數位典藏系統』中，則完整的解決了上述的問題。我們利用漢字構形資料庫來避免掉各單位使用者自行造字的成本，而且構字式與原本的系統編碼是相容的，因此只要使用構字式為交換碼的系統，不但非缺字的內容能夠交換，即使是缺字的資料也能夠正確在系統間進行交換工作，並不會產生編碼重覆而產生互相覆蓋的問題。除此之外，利用構字式部件及字根的組合方式來表達所有的缺字，因此也避免掉因不斷的造字而使得可用編碼空間耗盡的問題。所以在缺字問題上，字的數目可以是無限擴充的。所以數位典藏計畫的傳播與應用這一深層的意義也得以實現。

## 5. 結論

在數位化過程中，凡是遇到漢字的人名、地名、史料等等，都有相當嚴重的缺字問題，若不能有效的克服此一問題，則整個數位化的成果勢必受到極大的影響。在本文中所提出的缺字系統整合架構，無疑給了這個難解的問題一個答案，它的優點及克服的問題如下：

1. 解決了典藏系統的著錄、顯示與查詢缺字問題，並且缺字的處理並不會改變原來典藏系統的運作流程。
2. 使用漢字構形資料庫所收錄之字形，減少典藏單位自行造字之成本，而且以構字式為交換碼的缺字表示方式並不用擔心因交換碼空間不足而造成缺字數量的限制。
3. 使用者端並不需要額外安裝任何軟體，即可透過網路觀看含有缺字之網頁，且以圖形的方式讓缺字資訊的得以順利流通。

整個缺字處理的方法也成功的整合在傅斯年圖書館藏善本古籍數位典藏系統中，因此也證明了它的可行性。未來仍有幾項重要的工作需要進行，第一，在缺字的編碼方式上，目前漢字構形資料庫是以 Big5 作為字根及部件的編碼格式，未來文獻處理實驗室將規劃採取 Unicode，到時候典藏系統必須改變資料處理方式，並且對於已經著錄的資料必須經過轉換對應到 Unicode 後才能使用；第二，在字形支援上，目前僅使用到漢字構形資料庫中 5 萬多筆的楷體字形，故對於小篆、金文或是甲骨文皆無法使用，雖然目前典藏單位中對於這些字形並無大量的著錄需求，但若能將這些字形都整合進來，那對於漢字處理的缺字解決方法將會更加完整。


## 誌謝

1. 行政院國家科學委員會，數位典藏國家型科技計畫-技術研發分項計畫，NSC93-2422-H-001-0003
2. 行政院國家科學委員會，數位典藏國家型科技計畫-技術研發分項計畫-典藏系統建置與相關技術研發計畫，NSC 93-2422-H-001-0004
3. 數位典藏技術發展組(DAAL)
4. 中央研究院資訊科學研究所電腦系統與通訊實驗室研發成果
5. 中央研究院資訊科學研究所文獻處理實驗室技術協助
6. 中央研究院歷史語言研究所傅斯年圖書館合作


## 參考文獻


1. 謝清俊，“電子古籍中的缺字問題”，1996.08
2. 莊德明，謝清俊，林晰，“中央研究院古籍全文資料庫解決缺字問題的方法”，1998.05
3. Unicode Standard, http://www.unicode.org/unicode/reports/tr28/
4. 漢字構形資料庫, http://ckip.iis.sinica.edu.tw/CKIP/tool/
5. 缺字查詢系統, http://140.109.18.63/word/s.html
6. Chen-Yu Lai, Jan-Ming Ho, You-Qiao Wang, Zhi-Zhueng Huang “A composite approach to handle missing characters on Web interface”, ICDAT2004
7. 傅斯年圖書館藏善本古籍數位典藏系統, http://nddemo.iis.sinica.edu.tw/rarebook/