「人文社會科學新領域之開發及研究環境之改善」計畫執行報告

壹、緣起

96 年 2 月 12 日，數位典藏 1、2 月份工作小組合併會議中，何大安委員提出「數位典藏對於缺字處理，應採何種策略？」的討論案，陳克健委員指出資訊所文獻處理實驗室早已提出缺字解決方案，只是缺乏經費及人力，最後劉總主持人決議由史語所、語言所與資訊所提出申請，各所經費不超過 100 萬元，並呈請副院長辦公室批准支應。上述經費核可的批文已於四月底到史語所、語言所與資訊所。五月八日，史語所代表陳昭容老師、資訊所代表陳克健老師、語言所代表盧秋蓉小姐開會討論，決議交由資訊所莊德明先生負責計畫的執行，並統籌經費的使用及人員聘用，而史語所及語言所提供行政上的支援[1]。

貳、計畫提要

一、 研究目的：解決數位典藏計畫網頁缺字的呈現問題
二、 時程：自民國九十六年五月一日起至九十六年十二月三十一日止，共八個月。
三、 經費：共計新台幣 284 萬元整
四、 執行單位：中研院資訊所文獻處理實驗室
五、 執行人員

莊德明　中研院資訊所研究助技師
宋雲鳳　中研院資訊所研究助理
趙苑曲　中研院資訊所研究助理
黃榮順　中研院史語所研究助理
鄧賢瑛　中研院資訊所研究助理
丁玟伶　中研院語言所研究助理
陳建安　中研院史語所研究助理

參、實施方法

缺字問題的解決絕不能簡化成「造字」，各自造字的結果是資料無法交換及共享的主因。中研院資訊所文獻處理實驗室解決缺字問題的方法，是遵從漢字構形的原理，對漢字字形的結構做制式表達與詳

[1] 詳http://www.sinica.edu.tw/~cdp/service/documents/A960508.doc

細的分析。一個字的字形結構式（簡稱構字式），是該字極佳的識別符號與工具；因為字形若不一樣，則字形結構必不相同；反之，字形結構若相同，其形亦必同。構字式的制定及應用其實是奠基於「漢字構形資料庫」。

漢字構形資料庫[2]是從民國八十二年開始研發，經過這十幾年來不斷的擴充，它的主要特色如下：一、銜接古今文字以反映字形源流演變。二、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。三、記錄不同歷史時期的漢字結構，以呈現漢字因義構形的特點。四、使用構字式及風格碼來解決古今漢字的編碼問題

網頁缺字的呈現問題，指的是如何將網頁的構字式轉成字形。中研院資訊所網路與通訊實驗室於民國九十一年已成功使用Java Applet將網頁的構字式轉換成字形圖片，數位典藏技術發展組如今則改用Java Script，轉成圖片的速度也較舊版本快。這兩種技術同樣都可讓網頁瀏覽者在不需額外安裝其他軟體的情況下，直接使用現有的瀏覽器看到缺字。但對網頁的建置者而言，包含構字式的網頁必須加入Java Applet或Java Script的描述指令。[3]

肆、執行經過及成果

一、 資訊所研究助理鄧賢瑛於五月一日到任，語言所研究助理丁玟伶、史語所研究助理陳建安於六月一日到任。

二、 自六月四日起，先後聘用造字工讀生葉淑音等二十五人。

三、 協調計算中心、數位典藏技術發展組，分別於三月十四日、二十二日，四月十二日、二十六日，五月十日召開五次缺字技術會議。[4]

四、 從五月十五日起，訪談缺字單位，瞭解缺字問題。訪談的單位包括史語所中原考古庫房、佛教石刻造像拓本計畫、傅斯年圖書館、地理資訊室及語言所上古漢語文獻標記語料庫計畫、先秦金文簡牘詞彙資料庫計畫、閩客語典藏計畫。[5]

五、 分析以往漢籍電子文獻累積的 5,174 個缺字，其中至少有

[2] 詳http://www.sinica.edu.tw/~cdp/cdphanzi/

[3] 詳http://www.sinica.edu.tw/~cdp/service/documents/T960807.pdf

[4] 詳http://www.sinica.edu.tw/~cdp/service/meeting.htm

[5] 詳http://www.sinica.edu.tw/~cdp/service/problems.htm

1,639 個字必須增補到漢字構形資料庫[6]；分析語言所上古漢語文獻標記語料庫計畫的 3,306 個缺字，其中有 1,850 個字必須增補到漢字構形資料庫[7]。

六、 自六月四日起開始進行造字，需修正的有漢語大字典字形兩千多，增補的新字包括漢籍電子文獻及上古漢語文獻標記語料庫計畫的缺字 3,000 多個外、大正藏缺字約 8,000 個、Unicode 缺字約 14,000 個。本年度計完成 14,085 字，由於每個字需同時造標楷體及細明體，實際造字數為 28,170 字。

七、 史語所助理陳建安自七月下旬投入漢籍電子文獻Big5 缺字轉碼的工作。新版漢籍電子文獻的缺字因同時採用舊版漢籍造字及文獻處理實驗室的構字式，導致部分造字字碼重複定義，無法全部自動轉碼，重碼的造字必須重新比對原書才能確定字形。這項工作是由九十五年底開始進行，截至十二月底，成果如下[8]：

1. 完成《清實錄》56 部書的缺字轉碼，合計約 1 億兩千萬字。
2. 這 56 部書使用舊版漢籍造字 3,959 個，字頻 144,993 次，缺字比率約為萬分之十二；3,959 個缺字有 1,644 個可轉成 Unicode，字頻 104,446 次。由 Big5 轉成 Unicode 後，缺字比率降為萬分之三。
3. 新增缺字 2,944 個需增補到漢字構形資料庫 2.5 版。
4. 技術報告《漢籍電子文獻缺字轉碼—以《宋史》為例》將於九十七年元月出版。

八、 分別於八月八日協調數位典藏技術發展組與史語所金文工作室，九月十二日與史語所中原考古庫房，九月二十七日與傅斯年圖書館解決網頁缺字問題。

九、 八月十三日推出漢字構形資料庫 2.5 版。2.5 版收錄古今漢字 112,533 個；其中楷書字形 62,366 個，小篆及重文 11,100 個，金文 20,069 個，楚系簡帛文字 16,801 個，甲骨文 2,197 個。另收《漢語大字典》異體字表 12,208 組。[9]

十、 八月二十一日公開缺字處理計畫網站及缺字服務電子信箱，正式受理全院缺字問題。缺字處理計畫網站包括行事曆、技術文件、會議記錄、缺字問題、造字服務、缺字轉

---

6 詳http://www.sinica.edu.tw/~cdp/service/documents/B960727.doc

7 詳http://www.sinica.edu.tw/~cdp/service/documents/B960801.doc

8 詳http://www.sinica.edu.tw/~cdp/service/documents/transfer/record961121.xls

9 詳http://www.sinica.edu.tw/~cdp/cdphanzi/

碼、聯絡窗口等項目。[10]至十二月十八日為止，回覆使用者的信件 54 封，包括 22 個人；其中院內有 25 封，包括 9 個人。

十一、 提供造字服務，九月五日完成史語所金文工作室為籌備出版《第一屆古文字與古代史學術研討會論文集》所需缺字 122 個[11]；九月十一日完成數學所網頁人名缺字 1 個[12]；十月十二日完成「閩南語典藏—歷史語言與分布變遷資料庫」網站，戲曲《荔鏡記》俗體字所需造字 14 個[13]；十月十八日完成史語所金文工作室「甲骨文詞彙資料庫」缺字 41 個[14]。

十二、 撰寫缺字技術文件十二篇[15]。

伍、檢討與建議

網頁缺字問題的解決，目前必須透過網頁內容提供單位、網頁建置單位、數位典藏技術發展組及文獻處理實驗室四個單位的配合，才能圓滿達成；若有一個單位來不及配合，則問題依舊無解，以下分別舉例說明：

一、 數學所網頁缺字問題。

1. 九月十一日中研院數學所來信表示，民國六十七年數學所由范㯳先生擔任所長一職，人名「㯳」字為缺字。現因建置數學所簡史網頁[16]以及編輯文件等用途，需要資訊所文獻處理實驗室協助造字，並告知如何處理網頁缺字問題。
2. 九月十一日下午文獻處理實驗室於完成造字，並更新漢字構形資料庫，提供下載。
3. 九月十九日數位典藏技術發展組更新伺服器端的漢字構形料庫及字型。文獻處理實驗室測試確認「㯳」字可在網頁顯示，並通知數學所已可著手更新網頁缺字，並提供技術文件告知如何處理網頁缺字。
4. 十月五日數學所來電詢問如何在網頁編輯軟體上加

[10] 詳http://www.sinica.edu.tw/~cdp/service/
[11] 詳http://www.sinica.edu.tw/~cdp/service/documents/B960905.pdf
[12] 詳http://www.sinica.edu.tw/~cdp/service/documents/B960911.pdf
[13] 詳http://www.sinica.edu.tw/~cdp/service/documents/B961012.pdf
[14] 詳http://www.sinica.edu.tw/~cdp/service/documents/B961018.pdf
[15] 詳http://www.sinica.edu.tw/~cdp/service/tech.htm
[16] 詳http://www.math.sinica.edu.tw/history/history_c.php

入構字式，答覆可至中文缺字查詢網頁[17]直接複製；由於考量到可能會有編輯office文件等後續需求，因此建議直接下載安裝漢字構形資料庫 2.5 版。

5. 十月十二日去信詢問數學所是否已經解決缺字問題，但未得到回覆。之後以電話聯繫，數學所方面表示已經沒有其他問題，然而數學所簡史網頁上的缺字，截至目前，尚未更新。

二、 史語所中原考古庫房「考古資料數位典藏著錄系統」的缺字問題。

1. 九月四日文獻處理實驗室和中原考古庫房開會確認著錄系統的缺字問題類型：(1)倚天中文系統的擴充字引起的亂碼問題。(2)中原考古庫房自行採用的缺字表達式的轉換問題。[18]
2. 倚天字引起的亂碼問題，是數位典藏技術發展組在將資料轉入資料庫時，因程式錯誤所產生的，後來錯誤雖已修正，但並未同時修正資料庫的亂碼問題。[19]
3. 九月十二日數位典藏技術發展組、中原考古庫房及文獻處理實驗室討論倚天字的亂碼問題，中原考古庫房希望數位典藏技術發展組將可能出現問題的資料庫欄位匯出，並交由他們確認字形。
4. 十月十八日追蹤中原考古庫房線上資料庫的缺字處理情形，倚天字相關的錯誤資料正在修訂當中，下一步預計處理考古庫房自訂缺字表達式的問題，並建議考古庫房可以先將需要造字的清單提供給文獻處理實驗室進行造字。
5. 十二月三十一日考古庫房來信表示，著錄系統仍有缺字，並請文獻處理實驗室持續提供協助與諮詢。

三、 語言所「閩南語典藏-歷史語言與分布變遷資料庫」《荔鏡記》的網頁缺字問題。

1. 九月十日文獻處理實驗室與「閩客語典藏計畫」助理開會討論缺字問題，初步選定處理該計畫網站「閩南語典藏—歷史語言與分布變遷資料庫」[20]《荔鏡記》的網頁缺字。
2. 十月十二日文獻處理實驗室完成文本缺字分析，並增

[17] 詳http://char.ndap.org.tw/search/
[18] 詳http://www.sinica.edu.tw/~cdp/service/documents/B960904.pdf
[19] 詳http://www.sinica.edu.tw/~cdp/service/documents/B961015.pdf
[20] 詳http://southernmin.sinica.edu.tw/index.htm

補新字[21]及更新漢字構形資料庫。

3. 十月十九日文獻處理實驗室與語言所資訊人員、「閩客語典藏計畫」助理討論如何將網頁的缺字改為構字式，但由於網站系統開發人員已離職，目前網頁仍無法更新。

四、 史語所「殷周金文暨青銅器資料庫」的網頁缺字問題。

1. 「殷周金文暨青銅器資料庫」的網頁缺字問題堪稱最為典型。史語所金文工作室為內容建置單位，資料庫系統為計算中心開發，缺字採用文獻處理實驗室的構字式，網頁缺字呈現技術則由數位典藏技術發展組提供。
2. 八月八日金文工作室、數位典藏技術發展組及文獻處理實驗室討論金文網頁缺字顯示過慢的原因，主要是由於採用早期的 Java Applet 技術，新版 Java Script 的缺字顯示速度，金文工作室表示可以接受。
3. 八月二十九日計算中心表示，新版資料庫已改用 Unicode，並可搭配 Java Script 技術顯示網頁缺字，然因中心工作繁忙，目前只先提供測試版。
4. 十二月二十八日計算中心來信表示，有部份構字式未能顯示字形，將繼續追蹤、記錄並回報。
5. 九十七年一月三日金文工作室來信表示，目前網頁缺字呈現的字形不夠清晰，並希望能接著解決缺字的搜尋問題。

為了有效解決網頁缺字問題，我們提出以下的建議：

一、 請多利用缺字服務電子信箱。由於解決網頁缺字問題，需要四個單位的配合，各單位若能透過缺字服務電子信箱告知目前缺字處理情況，問題才能及早解決。

二、 中文字碼全面改用 Unicode。自從 Microsoft Windows 2000 中文版改用 Unicode，Unicode 目前已成中文字碼主流。中研院資料庫的建置，早期均以 Big5 為主，目前已逐漸改用 Unicode。採用 Unicode 可降低缺字頻次，降低網頁缺字處理成本，然而 Unicode 仍然有缺字問題。Big5 轉成 Unicode 並非難事，難的只是如何將原先的 Big5 缺字轉成 Unicode，並處理 Unicode 的缺字。文獻處理實驗室將於 97 年研發 Unicode 版本的漢字構形資料庫及構字式，並協助處理

[21] 詳http://www.sinica.edu.tw/~cdp/service/documents/B961012.pdf

Unicode 缺字。

陸、結語

本年度承蒙劉副院長撥下經費來解決本院網頁的缺字問題，然而本院的缺字問題由來已久，絕非短時間內可解決。單就漢籍電子文獻而言，新版資料庫目前上線字數已超過二億八千七百萬字，目前缺字轉碼字數雖達九千六百萬字，依目前的進度看來，剩下的工作仍需一、兩年才能完成。

在五月八日討論今年經費及工作分配的會議中，陳克健老師希望今年新聘的三個助理，自明年起能編入各所，以利長期缺字問題的解決。這三位助理的工作及近況如下：

一、 陳建安先生負責史語所缺字，而史語所的缺字需求主要在於漢籍電子文獻，陳建安明年將建安納入漢籍電子文獻計畫。
二、 鄧賢瑛小姐擔任缺字聯絡窗口，規畫缺字計畫網站、回覆使用者問題、撰寫使用說明及技術文件，鄧賢瑛明年將納入數位典藏技術分項。
三、 丁玟伶小姐負責語言所缺字，明年將納入語言所的員額，持續處理語言所語言典藏、文獻語料庫等相關計畫的缺字問題。

本年度的預算雖已執行完畢，然而聘用的三位助理明年仍編在不同的計畫繼續處理缺字問題。謹在此感謝劉副院長、何大安委員、林富士委員、陳昭容老師、陳克健委員等人對於缺字計畫的支持。