編號：CCMP98-RD-044

中醫藥典籍罕用字解決方案之研究

魏林梅
財團法人中文數位化技術推廣基金會

摘　要

貴會電腦系統中現在有衛生署造字集與 貴會自造字集兩種罕用字集，因其編碼不同，在儲存及處理中醫藥典藏資料及資料交換時，會碰到一字兩碼的情形，需建立一套符合 CNS11643 國家標準交換碼之自造字管理系統，解決典藏資料在管理及網頁顯示方面之問題。

本計畫依據下列方法實施：

1.首先比對兩套造字集的字型，建立以 CNS11643 為核心的對照表。
2.在 貴會準備的電腦硬體上安裝全字庫字型及屬性資料與本基金會開發的造字管理伺服器軟體，並技轉 貴會資訊使用者正常使用該造字管理軟體進行國家標準字型查詢、新增該造字至 貴會專屬自造字集中的方法。
3.輔導 貴會資訊系統管理人員在會內指定的個人電腦上安裝 PC 端造字自動更新軟體，能夠快速將伺服器新增造字資料更新到 PC 端使用。
4.在伺服器上安裝圖型產生軟體，整合網頁系統的程式軟體，讓網頁瀏覽者不需安裝罕用字集便可顯示及輸入 貴會收納的罕用字。
5.建立兩套造字檔間的轉碼介面軟體，讓資料交換後能正常顯示所造之罕用字的編碼。

本計畫結果：

1.比對衛生署版(5,107 字)、貴會 94 年版(4,025 字)及 94 年之前版(2,330 字)三套造字集共計 11,462 字的字型，完成以 CNS11643 為核心的對照表。協助建立 貴會與衛生署造字管理員間之協調程序，將 貴會於本案結案前所需要的新字加入衛生署版 Big5 造字集的 FBF0~FE7E 及 C6A1~C8FE 造字碼區，完成 貴會罕用字整合至衛生署字集之目標。
2.完成在 貴會準備的電腦硬體上安裝與測試全字庫字型及屬性資料、本基金會造字管理伺服器軟體。
3.衛生署已應 貴會請求將 貴會所需自造字(506 個外字)於衛生署使用之文鼎造字集造字完成，貴會同仁可直接使用文鼎外字集，已無需於 貴會個人電腦上另行安裝 PC 端造字自動更新軟體。
4.因 貴會典籍資料無線上即時轉碼之需要，完成協助 貴會以批次方式轉換

使用舊碼之 86 本典籍之資料檔成為使用新編衛生署造字集的資料。

5.網頁顯示及輸入自造字的元件軟體已開始於 貴會指定網站完成導入測試。

6.為因應 貴會網頁顯示自造字所需，本會字庫伺服器於保固期內隨時保持最新字集。

關鍵詞：cns11643、造字管理軟體、罕用字

Number: CCMP98-RD-044

Management System of User Fonts

Selena Wei

Chinese Foundation for Digitization Technology

ABSTRACT

CCMP computer system includes two rare character sets, the Department of Health character set and CCMP user-created character set. However, due to encoding variation between the two sets, users often encounter a situation of having two codes for the same character while they are saving and processing collected information and data interchange of Chinese herb medicine; thus such situation results difficulties in the management of collected information as well as website page display. Therefore, it is urgent to establish a set of user-created character management system keeping with CNS11643 Code to address the issue.

The implementation plan based on the following methods:

1. First of all, compare fonts of the two character sets to create a cross reference table centered on CNS11643.
2. Install Master Ideographs Seeker fonts, attribute data and CMEX developed character-creating management sever software in your computer hardware, and provide technical support and method to your users who take normal applications of the font management software to make queries of national standard fonts or add the created characters to your exclusive user-created characters set.
3. Assist your information system management personnel to install character-creation automatic update software in the designated PCs to rapidly update sever add-defined data to PC user end.
4. Develop code conversion interface software for the two character sets so that the codes of the rare characters can be displayed in a normal status after interchanging data.
5. Install graphic software and webpage system integration software in sever so that users are able to display or input rare characters collected by you without installing a rare character set.

Results:

1. CNS11643 core table was set up by comparing the Department of Health version (5,107 characters), CCMP 2005 version (4,025 characters) and the version before 2005 (2,330 characters), and we compare totally 11,462 characters fonts in the three character sets.

   Those additional new characters required by CCMP was also and will be keeping on be added to the Department of Health version of FBF0~FE7E and

C6A1~C8FE character-creating code section through the coordination with the DOH Fonts manager who completed the integration of rare characters into the DOH character sets during the 1 year warranty time period.

2. We completed the installation and testing of Chinese Fonts Management System on CCMP's server for Master Ideographs Seeker font and attribute data.
3. It is because we had integrated required 506 new characters into DOH's characters set, there is no need to install our PC client automatic update software into CCMP's personal computers.
4. Since CCMP does not have the real-time trans-coding need for crossing fonts set on-line reference data. We complete the conversion of 86 publications reference data from using the old CCMP 2005 version code into the newly version complied the Department of Health character-creating encoding data file in batches.
5. We successfully provided the component software for webpage display and input user-created characters and helping CCMP's contracting software vendor to integrate into website.
6. We will keep our font server with up to date character sets during the 1 year warranty time period for webpage display and input user-created characters prupose.

Keywords: cns11643, font management system, gaiji

壹、前言

中國醫學藥典的內涵豐富，流傳久遠，造福後世，闕功至偉。 貴會將這些資料數位典藏的成果，對資訊化社會的價值非常重大。然而因為古人寫書並沒有要求使用標準字形，容或發生某個字形多一點、少一畫的情形，但是後人還是可以意會出所代表的涵意。

但是若要將所有中國古籍中出現過的字形一一編碼造字，可能會有數十萬字之多，遠超出 貴會目前電腦配備的 Big5 字形編碼(13,053 個字)所能處理的範圍。貴單位負責整理、出版中醫藥古今資訊，亦常發生電腦系統打不出中文罕用字的情形，造成出版業務及資料處理上的困擾。

本計畫分兩階段以漸進式完成，第一階段完成有關自造字基礎架構的整理與建置。第二階段擴充至整頓資料庫系統及造字交換轉碼的應用上，計劃如下：

第一階段(2009/9/8~2010/3/8)：

一、以教育部發佈的國標字為標準，比對 貴會現有中文造字檔，產生問題字、錯別字、重覆字、符號字...等狀況分析表，若為 CNS11643 中文標準交換碼全字庫已有之字，應列出其中文標準交換碼，若為全字庫所無之字，則提供造字服務，並申請 CNS11643 中文標準交換碼。並彙整為一套共用標準自造字集，供 貴會內部電腦使用。

二、協助向戶政、地政、商工主管機關查詢問題字的正確寫法。

三、提供處理數字編號字、符號字的諮詢顧問。

四、提供 Web 化自造字對照碼表管理介面(包含中文造字檔的 Big5 內碼、CNS 11643 國家標準交換碼、Unicode 國際碼)。

(一)於 貴會指定之 Server 上安裝 Server 軟體(Linux 或 Windows 版)。

(二)協助 貴會在所有 Windows PC 上安裝與造字 Server 同步自動更新造字的 Client 端軟體。

(三)提供一年保固服務及 貴會若有新增字時，由本會協助過濾是否全字庫有該字。若有，則提供 CNS11643 交換碼交予 貴會，以利 貴會向衛生署行文新字申請；若無，則請 貴會提供該字出處(那一本典籍)影本及前後段文字，以利本會送請主計處編列國家標準交換碼時之審查工作，本基金會會先行完成新字造字，供 貴會向衛生署行文新字申請。

五、提供使用自造字管理軟體的諮詢顧問服務。

第二階段(2010/3/9~2010/9/8):

一、授權使用網頁顯示自造字元件軟體及資料轉碼軟體。

二、掃瞄 貴會資料庫內使用自造字之狀況。

三、提供資料庫存取自造字資料的客制化轉碼介面軟體及轉碼服務。

四、提供與 貴會網頁應用系統整合以上元件軟體的實作與諮詢顧問服務。

貳、材料與方法

一、現有造字字型比對採用之方法：

(一) 如果原造字集有注音或倉頡屬性資料，以電腦程式產生相似字表，加速檢視速度。

(二) 如果完全沒有屬性資料，則以人工比對，貴會提供的兩套字集(『94 年之前使用之版本』及『94 年版本』)皆為採用人工比對作業。

二、現有典籍資料是否使用某字集之方法：

採用本基金會開發的造字自造字掃瞄程式。

三、自造字管理之方法：

本基金會開發之造字伺服器軟體及 Windows PC 端自動更新造字軟體。

四、網頁顯示自造字之方法：

本基金會開發之網頁顯示自造字元件軟體。

參、結果

一、第一階段先對 貴會前後共提供的三套造字檔進行比對，結束如下：

(一)『94 年之前使用之版本』，計有 2330 個造字。

『94年之前使用之版本』自造字比對分析報告

日期 / 類別	字數/百分比		備註
自造字字集	2330	100%	
待查問題字	-113	4.85%	全字庫中沒有納入的字。
重覆字	-311	13.35 %	307組618個重覆字。 資料庫有一字兩碼的現象，有資料檢索不到的風險。
系統字	-221	9.48%	有3個系統字，20個系統數字，25個系統部首字，173個系統日文字。 Windows已內建，不應再造字。
符號	-220	9.44%	有23個符號，197個數字，若文書需要，可以保留，但無法透過CNS11643交換碼進行跨機關間的文字交換。
可保留之正確造字數量	=1465	62.88 %	因卜/戈; 糸/火; 走字部 ;羽字部...等筆誤字形有298個字，可以改用教育部國標字楷體或宋體字型。
剩餘之造字碼位	4752		=6217 - 1465

(二)『94 年版本』，為何技正 94 年所編，計有 4,025 個造字。

『94年版本』自造字比對分析報告

日期 / 類別	字數/百分比		備註
自造字字集	4025	100%	
待查問題字	-1248	31.01%	全字庫中沒有納入的字。
重覆字	-11	0.27%	11組22個重覆字。 資料庫有一字兩碼的現象，有資料檢索不到的風險。
系統字	-9	0.22%	屬於Windows已內建的13053個系統字，不應再造字。
符號	-31	0.77%	建議刪除。
可保留之正確造字數量	=2726	67.73%	因卜/戈; 糸/火; 走字部 ;羽字部...等筆誤字形有289個字，可以改用教育部國標字楷體或宋體字型。
剩餘之造字碼位	3491		=6217 - 2726

(三)『衛生署 98 年版本』，有 5,107 個造字。

『衛生署98年版』自造字比對分析報告

日期 / 類別	字數/百分比		備註
自造字字集	5107	100%	截至2010/02/24止
待查問題字	-278	5.44%	全字庫中沒有納入的字。
重覆字	-64	1.25%	62組126個字。 資料庫有一字兩碼的現象，有資料檢索不到的風險。
系統字	-39	0.76%	屬於Windows已內建的13053個系統字，不應再造字。 資料庫有一字兩碼的現象，有資料檢索不到的風險。
符號與數字	-25	0.49%	建議刪除。若文書需要，可以保留，但無法透過CNS11643交換碼進行跨機關間的文字交換。
可保留之正確造字數量	=4701	92.05 %	
剩餘之造字碼位	1516		(=6217 – 4701)

(四)合併字集的可行性分析。

以上三套字比對的總字數達 11,462 字之多，並分別產生各個字集的比對分析報告。接著經過交叉比對，以一字一碼之原則合併字集，無論以二合一或三合一方式，其所需要的造字碼位皆超過 Big5 用者造字區所能容納的 6,217 個字，無法產生統一的字集。

94年版與衛生署98年版兩套字集合併之分析報告

日期 / 類別	合併後需要之造字碼位	備註
維持一字一碼所需之總造字碼位　(兩套字中若有相同字形者，則歸納為一個字碼)	6813	1.94年版本 ：2,726個正確自造字 2.衛生署98年版本 ：4,701個正確自造字 3.刪除重覆字、問題字及系統字。
待查問題字(全字庫中完全找不到對應或相似的問題字)	1526	1.94年版本： 1,248個問題字 2.衛生署98年版本 ：278個問題字
最大字集總造字碼位	8339	

三套字集合併之分析報告

類別 \ 日期	合併後需要之造字碼位	備註
維持一字一碼所需之總造字碼位 (共比對11,462字，這三套字中若有相同字形者，則歸納為一個字碼)	7460	1.94年版本 ：2,726個正確自造字 2.衛生署98年版本： 4,701個正確自造字 3.94年之前使用版本 ：1,465個正確自造字 4.刪除重覆字、問題字及系統字。
待查問題字(全字庫中完全找不到對應或相似的問題字)	1639	1.94年版： 1,248個問題字 2.衛生署98年版本： 278個問題字 3.94年之前使用版本 ：113個問題字
最大字集總造字碼位	9099	

(五)掃瞄　貴會資料庫內使用自造字之狀況。

經　貴會同意，提早先行掃瞄 貴會提供典籍組歷年來所編印的 86 本典籍，確認完全沒有使用「94 年之前使用之版本」字集，經 貴會同意，決定不處理該字集。

同時發現以上典籍曾使用 94 年版本造字集中的字只有 805 個。將此 805 個字與衛生署 98 年版字集比對後，貴會需要而衛生署沒有的自造字計有 486 個字。加上 貴會於 2 月新增字，共計 506 個字。將此 506 個字與衛生署 98 年版本合併後，共需 5,616 個碼位，小於 6,217，代表可以跟衛生署字集進行二合一。

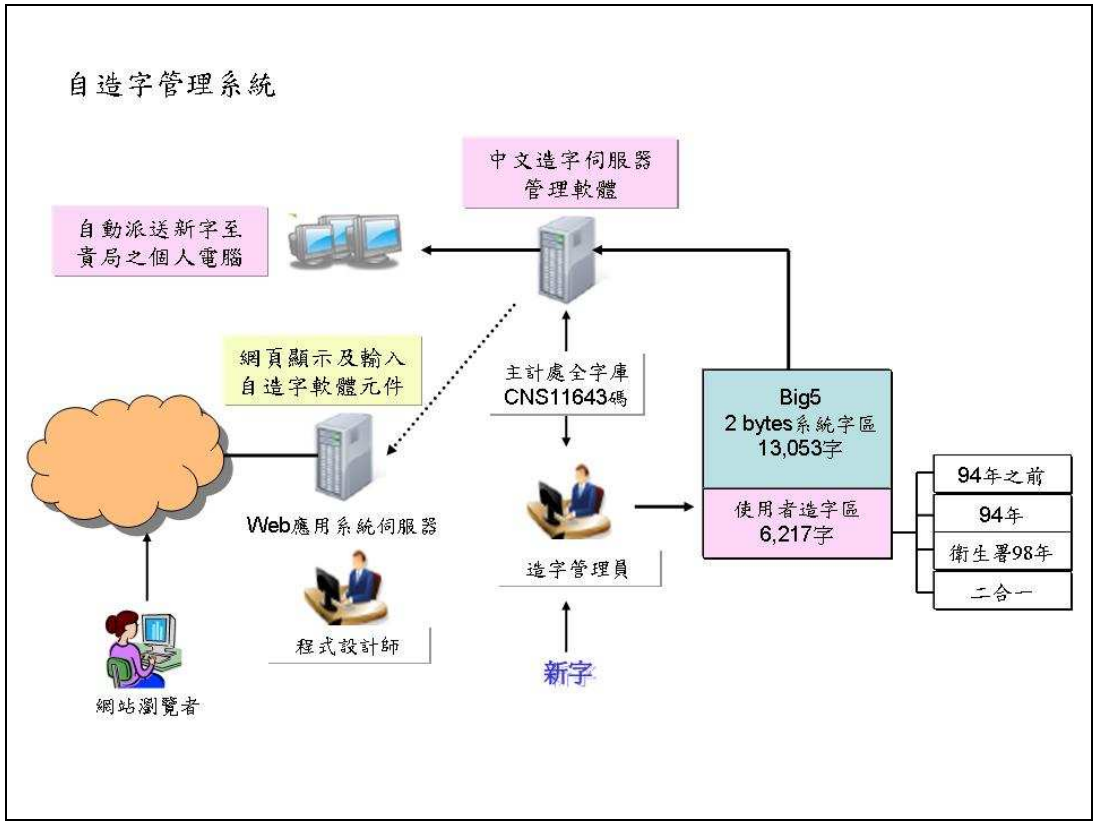
經提供對照表給署方後，於 5 月完成將此 506 字加入衛生署版 Big5 字集的 FBF0~FE7E 及 C6A1~C8FE 中。

貴會與衛生署字碼合併之編碼原則

Big5造字碼區	衛生署98年版	貴會94年版	二合一版 (將貴會所需之造字加入衛生署開放之空碼區)
FA40~FEFE	FA40~ (只使用51字碼)		FBF0~FE7E(365個碼位)
8E40~8FFE	8E40~8DBB	8E40~8FFE	完成二合一後，將提供新舊碼對照表，並指導 貴會工程師完成86本典籍之轉碼工作。 合計509字
9040~9FFE	9040~9FFE	9040~9C48	
8140~8DFE	8140~8DBB		
A041~A0FE	A041~A0FE		
C6A1~C8FE			C6A1~C8FE (408個碼位)

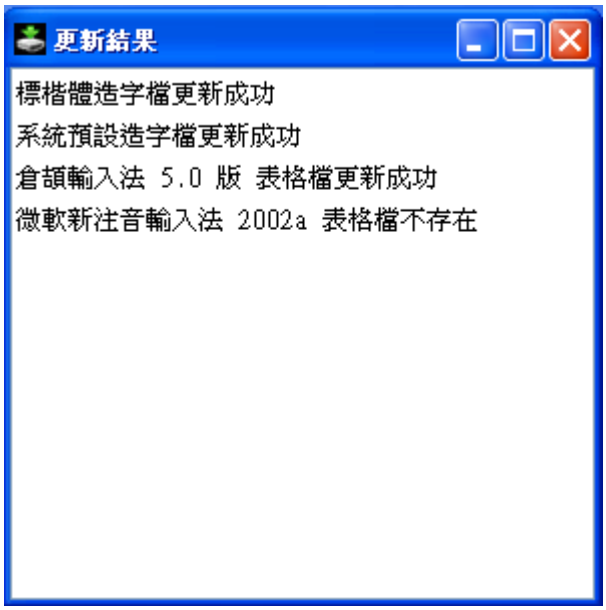
(六)安裝造字伺服器軟體及 Windows PC 端自動更新造字軟體。

第一階段末期之今年 2 月，完成安裝本基金會提供之造字伺服器軟體及 Windows PC 端自動更新造字軟體，經測試可以正常運作。造字系統架構圖及造字管理員管理介面如下圖所示。



PC 端使用者可以自行設定選項，包含指向那一個 Font Server 的 IP 及多久更新一次...等。

自動更新完成之後的更新結果訊息如右。



茲因　貴會決定直接自衛生署造字伺服器下載 貴會所需字型使用，本套系統僅處於備用狀況，但已確實完成高可行性之研究目標。

『系統使用手冊』燒錄於交付　貴會之光碟中。

二、第二階段之資料轉碼軟體，因 貴會決定採用衛生署版新字集，已無線上即時轉碼及提供轉碼介面軟體及轉碼服務之需求，並經與 貴會及負責維運 貴會資料庫之資通公司討論確認該資料庫中皆未儲存造字檔，不需進行轉碼。故提供服務完成協助 貴會以批次方式轉換「使用舊碼之典籍資料」成為「使用新編衛生署造字集的資料」。

三、最後於 貴會網站上加入本基金會自行開發的「網頁顯示及輸入自造字的元件軟體」，並邀請承辦 貴會網站的系統服務廠商進行將元件植入網站系統中的測試，完成驗證之畫面如下(網址：http://www.ccmp.gov.tw/)，達成本研究案的目標。


回首頁 | English | 網站地圖 | 問答集 | 為民服務 | 雙語辭彙 | RSS
中醫藥資訊網
行政院衛生署·中醫藥委員會
全文檢索
活動報導
徵人啟事
招標公告
新聞稿
最新消息
會員登入
帳號
輸入您的帳號
密碼
登入
查詢
申請
業務服務窗口
民眾查詢服務
電子報
委員會介紹區
中醫藥業務區
中藥藥品許可證查詢
研究發展區
資訊典籍區
法令規章區
中醫藥機關團體
醫藥知識區
本會出版品
本會公開資訊區
其他區
2010.08.20 【新聞稿】香港產製之『星洲海狗油』中成藥，國內無販....
【新聞稿】衛生署就中藥摻含西藥把關之具體措施
2010.08.04 【新聞稿】有關監察院糾正中藥摻含西藥嚴重，本會將持....
2010.07.30 【新聞稿】香港零售商「永和興」所販售『蒼朮』中藥材....
2010.07.30 【新聞稿】香港衛生署指令回收「黃澤記有限公司」所販....
前往觀看
中草藥
用藥安全資訊網
教學醫院
中藥臨床試驗中心
中醫評鑑
中醫不良反應
通報系統
行政院衛生署中醫藥委員會
10453 台北市雙城街6號 地圖 電話：(02)2587-2828 傳真：(02)2587-2121
瀏覽人次：1775233
隱私權保護政策 / 資訊安全政策 / 無障礙宣告 / 著作權聲明
無障礙 AAA

肆、討論

進行本研究案之前， 貴會業務單位分別使用衛生署版及自行造字編碼兩種版本的造字系統來進行文書作業及典籍編輯作業，本案緣起於簡化為一套字集。因為衛生署為上級單位，且已於七月接受本基金會建議收納 貴會典籍資料中已經使用的 506 個字及本案提供的衛生署最新版造字對照表，本案初步已經完成造字比對及與字集合一的工作目標。

因現階段因為衛生署所使用之文鼎造字伺服器軟體不會與本基金會提供之 Font Server 以 P2P 方式結合在一起進行同步更新。仍有下列問題需要解決：

一、如果選擇直接自衛生署現有的 Font Server 下載新字至 PC 端使用，就必須與衛生署之間建立一套新字申請的行政作業流程，以確保衛生署能即時將 貴會所需新字加入其 Font Server 中。

二、促請文鼎公司與本基金會合作完成兩種不同架構 Font Server 之間同步更新字集的系統介面。

因為衛生署的 Font Server 不會分享其電腦資源來支援 貴會進行網頁顯示及輸入自造字的功能，這部份的工作還是需要使用到於本案中所建置的中推會版 Font Server 來運作。

本案提供之新版造字管理員軟體與 PC 端自動更新軟體皆較衛生署多年前所使用的文鼎軟體為佳之處包括：

(一)可以即時查詢行政院主計處全字庫網站找到所需新字後，於管理者介面直接下載主計處提供之國家標準正楷體及正宋體(明體)向量字型(True Type Font)及注音、倉頡、筆畫、部首…等相關屬性資料。節省造字管理員處理這些工作的時間(目前衛生署造字管理員僅以組字工具產生一種楷體字型)。

(二)新增字可以透過 Web Based 管理介面輸入 CNS11643 的編碼及專屬內碼後，直接將該新字字型檔下載至 Font Server 上，並透過支援 Windows OS(2000/XP/2003/Vista/7)之 PC 端同步更新程式以單字更新方式加入 PC 中使用。

(三)PC 端同步更新程式可以設定多字集(例如本案設計 PC 端可選用衛生署 98 年版、中醫委 94 年版、中醫委 94 年之前版及 Default 的中醫委與衛生署 98 年版二合一之版本)來開啟文件。

(四)本案並完成包含二合一版本之衛生體系 Big5 內碼、Unicode 碼及 CNS11643 交換碼之轉碼對照表，未來衛生署若由 Big5 碼轉換至 Unicode 碼時，可以用此對照表輕鬆的進行資料轉碼之工程。

(五)網頁顯示及輸入自造字的實作證明可以讓網頁瀏覽者在不安裝任

何造字的條件下，依然可以用正常的注音或倉頡輸入法在網站上輸入查詢字(該字為造字)，並能正常顯示文件中的造字字型。

本案除解決 貴會處理中文自造字之問題外，亦同時幫衛生體系建置與驗證出未來可行之方案，貴會可以此實作之研究成果建議衛生署未來整合擴建中文自造字處理系統之參考。

伍、結論與建議

一、Big5 提供 13,053 個字型及 6,217 個使用者造字碼位，面對戶政 10 多萬字及中醫藥典籍中還有許多異體字，造字需求若不加以控管，使用 Big5 編碼的軟體很快就會面臨造字空間不足的問題。

二、本基金會已經同時建議 貴會及衛生署針對 Client/Server 架構的應用系統需逐漸改採以 Unicode 3.0 以上版本(提供 27,484 個字)的字集，處理中文造字時，才能遊刃有餘。

三、本案將 貴會最近十年來所產生的字集做了一次清查動作，並將大部份的自造字做好與 CNS11643 交換碼的對照表，甚至擴大到將衛生署的造字集一併做完 Big5、Unicode 與 CNS11643 的對照表，為 貴會與衛生署未來與其他跨機關之間的造字交換機制奠定楚石。

四、貴會未來新增字時，建議先自全字庫查出 CNS11643 交換碼後，轉請衛生署造字管理員查詢衛生署版字集是否已建有該字。若無，則編新碼通知 貴會使用；若已有，則通知 貴會使用何碼，不必再新增該字。

五、針對本案提供之軟體繼續提供一年之保固服務。

六、保固期間若新增字於全字庫網站(www.cns11643.gov.tw)上查不到時，敬請 貴會提供該字清晰圖檔、出處(那一本典籍)影本及前後段文字，以利送請編列國家標準交換碼時之審查工作，本基金會會先行完成新字造字，供 貴會向衛生署行文新字申請。

誌謝

本研究計畫承蒙行政院衛生署中醫藥委員會計畫編號CCMP98-RD-044提供經費贊助，使本計畫得以順利完成，特此誌謝。

陸、參考文獻

行政院主計處全字庫網站(www.cns11643.gov.tw)