

可攜式字集資源-用以解決缺字問題

謝育平*

台灣大學資訊工程研究所

arping@turing.csie.ntu.edu.tw

吳政泓

台灣大學資訊工程研究所

white@turing.csie.ntu.edu.tw

項潔*

台灣大學資訊工程研究所

hsiang@csie.ntu.edu.tw

摘要

缺字問題是指電腦內碼對應文字的不完備造成無法順利瀏覽及輸入未收錄的文字，無法順利傳播及查詢內含該文字的文件，難以在典藏過程是描述新遇到的缺字及完整統一造字的問題。缺字問題對於報業、圖書館或典藏機構來說，是一個重要的問題。本文提出可攜式字集資源架構用以自動解決瀏覽、輸入、傳播及查詢等問題，並結合部件組字法及統一造字法提出一個可行的缺字處理流程。

可攜式字集資源(Portable Word-Set Resource)是一種可攜式文字資源，其定義文字識別碼的方式是以“字集名稱+內碼擴充碼”，而在字集名稱中則暗示該字集資源在網路上的位址。可攜式字集資源架構處理缺字問題除了典藏造字者外，瀏覽者、寫作者、查詢者、傳播者對缺字幾乎都是不知不覺，其具有“只要瀏覽過的就可以使用輸入法輸入，只要寫完文件就可以傳播，只要傳播就可以瀏覽”的特性且支援各式支援字型的應用程式。

關鍵字

缺字問題、可攜式文字資源、可攜式字集資源、缺字碼靜態指定技術。

1. 簡介

文字在計算機中是以數位的方式儲存，透過一些協定來解釋數位資料所代表的文字。這些協定，我們稱之為內碼。例如台灣常用的BIG5編碼及國際常用的Unicode[7]編碼皆屬之。因為各個國家或地區大都會採取其慣用的

編碼法，而各個編碼法內所含文字也各有差異，所以造成編碼系統的不協調。例如：“發”的編碼是BIG5：B56F和Unicode：767C，而“塋”的編碼是Unicode：5803但在BIG5中則沒收錄。所以在現行內碼系統不協調的情況下，交換碼應運而生。交換碼主要是用來處理某領域下的內碼不協調問題。主要是將各內碼中的文字對應到交換碼中，然後再進行內碼系統的轉換。CNS11643[1]就是一個漢字領域的交換碼。例如我們要將某份含“發”的文件從BIG5編碼轉成Unicode編碼，我們就可以先查到BIG5：B56F在CNS11643中的編碼1-6075，然後再轉成Unicode:767C。不管是內碼或是交換碼，都有找不到文字(未收編)的時候，所以在內碼設計時皆會提供幾塊區域作為動態指定未收編字碼的區段。例如BIG5提供FA40 - FEFE、8E40 - A0FE、8140 - 8DFE、C6A1 - C8FE等區段，Unicode提供E000-F8FF區段。對於某個內碼系統來說，我們稱那些未收編的文字為缺字，或詳述之為某內碼系統的缺字。而我們稱這些提供缺字使用的區段為擴充字區或擴充區，而稱擴充區內的字碼為缺字碼或缺字內碼。例如在BIG5編碼中，“塋”並不存在，此時機關或個人使用者可以指定擴充字區的某一缺字碼給“塋”使用，例如BIG5：FA40。但是因為將“塋”指定為BIG5：FA40並不是一個廣泛性的約定，所以在不同的機關或個人可能會指定BIG5：FA5E給“塋”使用或是將BIG5：FA40指定給其他缺字使用，所以此時就發生了缺字不協調問題，簡稱為缺字問題。缺字問題在本質上分成兩類，一是同碼不同字，另一個是同字不同碼。因而衍生出六類應用上的問題：瀏覽、輸入、傳播、查詢、描述及造字問題等。所以缺字問題主要

*感謝國科會計畫 NSC 92-2811-E-002-027 及 NSC 92-2213-E-002-005 的支持。

發生在不同電腦之間，我們稱之為“跨系統”問題，有時缺字問題也會發生在同一台電腦系統中，我們稱之為“系統內”或“系統內部”問題。不管是在跨系統或在系統內部，同碼不同字遠比同字不同碼要來得嚴重許多。瀏覽問題是缺字問題的源頭，但解決瀏覽問題並不能讓缺字的使用方便，還有一個重要的問題是輸入法，輸入法是一種有系統的文字的檢索方法，通常有一個基本的文字與鍵串的對應表，然後使用者將一個文字或多個文字的鍵串輸入輸入法程式中並與輸入法程式做額外的互動，最後輸入法會將該文字或文字串輸出。我們稱該文字與鍵串的對應表為輸入碼表，而稱鍵串為輸入碼。例如 Microsoft 公司出的新注音輸入法，也有一張輸入碼表，如“發”“ㄔㄩ”，“塋”“ㄎㄩ”等，我們稱“ㄔㄩ”為“發”的輸入碼或更準確地稱為注音輸入碼。傳播問題大致上與瀏覽問題是一對的，因為傳播後遇到的第一個問題就是瀏覽問題，所以傳播問題處理的是傳播前的準備工作，而瀏覽問題處理的則是傳播後的瀏覽問題。描述及造字問題大多發生在典藏過程中，是缺字最重要的來源。至於查詢問題對整個缺字問題來說，是值得研究的，但其重要性明顯比其他問題要來的低。

本文與參考文獻[11]中的討論與架構大致相同，最大的差別在於引用“字集+內碼”作為文字識別碼來解決[11]的方法中有關系統損毀則缺字文件無法復原的致命缺點，雖然說與[11]只有些許概念上的不同，但其造成整個系統在實作上及缺字處理的各個面向上卻非常不同。

在第二節中，我們對缺字問題進行研究及討論，在第三節中，我們介紹目前主要的缺字解決方案，在第四節中，我們介紹可攜式字集資源架構，在第五節中，我們提出可行的缺字處理流程，而第六、七、八節分別是討論、誌謝及參考文獻。

2. 缺字問題

缺字問題主要是因為電腦內碼對應文字的不完備所引發的瀏覽、輸入、傳播、查詢、描述、造字等問題。以下我們就本質問題及六個應用面上的問題提出討論：

本質問題：指目前內碼系統設計的限制。

N1. 同字不同碼問題：這是因為缺字與擴充區字碼的對應並沒有一定的規範，所以在不同電腦有可能發生同一個字放在不同的缺字碼中。而在同一台電腦系統內，也有可能因為管理不善而有同字不同碼的問題。

N2. 同碼不同字問題：同樣地，這也是因為缺字與擴充區字碼的對應並沒有一定的規範，所以在不同電腦間有可能發生同一個缺字碼放置不同的缺字。而在同一系統內部，就內碼系統來說，這個問題應該不會發生。(雖然可攜式字集資源方法發生系統內同碼不同字的問題。)

N3. 缺字內碼有限問題：這是指缺字太多以致於無法全部放入擴充區中，當然也就無法處理瀏覽及輸入問題。例如 Unicode 只提供 E000-F8FF 擴充區段。但對於 CNS11643 所規定的 48,027 個繁體字來說自然不夠使用。

瀏覽問題：指無法正確顯示文件中的缺字。

B1. 跨系統文件瀏覽：這是指無法正確顯示外來文件中的缺字。這是所有缺字問題的起源。例如甲電腦中的文件含有缺字 BIG5：FA40“塋”，當我們將這份文件移至乙電腦中，則會因為乙電腦對於 BIG5：FA40 的解釋不同而有錯誤的顯示。

B2. 跨程式文件瀏覽：這是指在各式應用程式間交換文件時發生的缺字顯示問題。例如“塋”對於 BIG5 內碼來說是一個缺字，在 BIG5 網頁中使用例外編碼“堃”來表示之，該網頁使用瀏覽器觀看時能正常顯示“塋”，但是用 notepad.exe 等本文編輯器來開啟時則會造成瀏覽問題。

輸入問題：無法自然使用輸入法輸入缺字。

I1. 輸入法相容問題：一般使用者都有其慣用的輸入法及輸入法程式，我們覺得解決缺字的輸入問題應該要讓使用者可以繼續使用其原本慣用的輸入法，而不須重新學一個輸入法或是更換一個輸入法程式。若非不得已需要更換輸入法程式也要堅持不需要重新學一套新的輸入法。

I2. 缺字檢索問題：一般人對於缺字是相當陌生的，即使可以使用現有輸入法來定義其輸入碼，但因為缺字的少用性使得使用者很難根據外形來猜測其注音輸入碼、拼音輸入碼或倉頡輸入碼。所以可能需要設計一個新的、方便的、直覺的輸入法(檢索系統)來擴充既有的各式輸入法。

以上兩者看起來有些矛盾，但是各有其道理，在製作上並不衝突。

傳播問題：指無法自然地傳播系統內使用過的缺字或內含缺字的文件。

D1. 缺字傳播問題：缺字應該由專家認真的造一次，然後輕易地傳播給所有人使用。而傳播的動作應該讓使用者在不知不覺中完成。

D2. 文件傳播問題：網頁或是應用程式的文件規格實在太多。在給定一份文件，如何在傳播前在文件中內嵌缺字資訊或是指示缺字資訊所在地，進而在傳播後可以取得缺字資訊？(此問題同跨系統文件瀏覽問題)

查詢問題：指無法以缺字碼當成關鍵字來對文件資料庫進行全文檢索。

Q1. 跨系統文件全文檢索問題：對網路上文件來說，會因為各個電腦的缺字內碼不協調而造成同字不同碼或同碼不同字，進而影響文件的全文檢索。

Q2. 系統內部文件全文檢索問題：對於內部文件來說，如果有同字不同碼或是同碼不同字問題則仍會發生查詢問題。

造字問題：大多發生在典藏過程。

M1. 重複造字問題：針對不同領域或不相干的環境，重複造字在所難免且問題不大，但是如果一個典藏過程中重複訂定缺字則浪費缺字碼且造成查詢問題。所以應該執行統一造字以避免重複造字。

M2. 造字完整性問題：造字工作對於一般人來說，是一份繁瑣的工作，一般人遇到造字工作時大都只製造該字的一種字型，零種或一種輸入法。其他寫作者將無法使用其慣用的輸入法來輸入該缺字或是顯示其想要的其他種類字型。故造字需要統一由為數不多的專家認真地處理，然後分享給其他人一起使用。

M3. 統一造字問題：為了在某計畫內執行統一造字，首先遇到的是哪些缺字要訂定，所以我們需要在典藏過程中需要一個紀錄或描述缺字的方法，所以引發了描述的問題。

描述問題：典藏過程中如何紀錄描述缺字。

R1. 缺字描述速度問題：在典藏過程中，我們不希望每遇到一個缺字就須造一個新字，或是需要去查詢是否已經造過，否則整個典藏流程就被中斷了。所以我們需要一個描述缺字且其描述方法簡單。

R2. 缺字描述式重複問題：我們用來描述缺字的方法必須唯一或是可對應到唯一的缺字。

在繼續探討缺字問題之前，我們得了解缺字問題的使用者及其特性，我們歸納出六個角色。

缺字問題的六個角色

瀏覽者：這是一批人數眾多的角色，是上述瀏覽問題的主角，其對缺字問題的感受，是我們最要考量的。

寫作者(輸入者)：這是一批寫文章的人，人數較瀏覽者少，其面對缺字問題，最需要的是自然地使用其慣用的輸入法輸入系統中的缺字，故我們又稱之為輸入者，是輸入問題的主角。

傳播者：是指將文件或缺字資源放在網路上供人瀏覽的人。

查詢者：不管是跨系統或是系統內部的文件查詢，搜尋引擎的設計都是查詢問題的主角。

造字者：這是一批專家，各自對其領域內的缺字相當熟悉，人數更少，對造字者來說，如何將自己造的字分享給別人使用是其面對的最重要問題，故造字者是傳播問題的主角。

典藏者(描述者)：是指在典藏過程中，資料繕打的輸入者，因其常常遇到缺字，固其性質與普通寫作者不同，需要一套快速而方便的缺字紀錄方式。

以上六種角色並非完全獨立，常常是瀏覽者兼寫作者、寫作者兼傳播者、寫作者兼查詢者、典藏者兼造字者、造字者兼傳播者。就人數多寡來看，瀏覽者眾、寫作傳播者次之、典藏造字者少。對瀏覽者來說，缺字是要能正確顯示。對寫作者來說，任何缺字要能任意的輸入。對典藏造字者來說，缺字是要能統一不重複的訂定，對傳播者來說，缺字要能輕易地傳播的。理想中的缺字解決方案，除典藏造字者外，瀏覽者、寫作者、查詢者、傳播者對缺字問題都該盡量是不知不覺的，也就是要他們使用原本的習慣來對待缺字與非缺字。

3. 先前技術

在缺字問題的解決方法裡，我們整理出以下五個現存的解決辦法：

T1.全字內碼法：此方法是將目前所有出現過的各國文字盡量地蒐集，將之統一編碼，制定統一標準，供系統使用，Unicode System[7]就是屬於這種方法。對於繁體字來說，最好就是能將 CNS11643 全字庫[1]整個納入內碼使用。但是全字內碼法通常需要非常長的時間討論，審查才能公佈使用。此方法可以大大地減少缺字問題的發生頻率，但是與時俱增的、特殊領域的或是極少使用的缺字問題仍然存在。而且當缺字問題發生後，此方法對各類缺字問題，則未提供解答。

T2.擴充字集法：目前的缺字內碼數有限，所以可以設計一組一組的擴充字集來重複使用內

碼擴充區，當寫作者與瀏覽者使用相同字集時，則可正常瀏覽。此方法受限於缺字內碼數有限的影響，並無法將所有的字放入系統中。

T3.造字統一法：在擴充字集法中，字集是由字集廠商制定的，對於一個典藏或是報業機關來說，使用廠商制定好的字集，往往無法達到該機關的特殊需求，於是為了機關內的文件交流，得要將機關內的造字工作統一。滿天星公司的字霸軟體[9][9]就有結合 CNS11643[1]提供這樣的解決工具來統一機關內的造字工作、規定機關內缺字與內碼的對應。

T4.部件組字法：中央研究院的謝清俊教授，發展了一套部件組字法[8]，是將中文字用一個個的部件組合起來，例如“林”就可視為兩個“木”並排。這樣的部件組字法，可以很容易地組出很多我們遇到的缺字，我們可以用這個方法來描述缺字的長相。其實這樣的一套部件組字法，也是一套中文的輸入法。使用部件組字法，在瀏覽問題上，可以外掛 plugin 程式來解決網頁瀏覽問題，但對於系統內部眾多應用程式來說，則無法順利瀏覽，而且寫作必須學習這套部件組字法。

T5.字型資源法：瀏覽問題是缺字問題的起源，所以也最早被思考，進而解決的。此方法是將文字字型以資源方式看待，在文件中指出缺字字型資源所在或是內嵌缺字字型資源然後利用自己的描繪技術將字型正常顯示，值得注意的是她們都未將字型嵌入統中以供其他應用程式使用。Microsoft 的 Typography[2]、Bitstream 的 Portable Font Resource[4]、威鋒數位的 Dyn@File[10]都是此類的例子。

4. 可攜式字集資源架構

根據對缺字問題的了解及應現存五類解決方法的啟發，我們提出可攜式文字資源架構[11]及可攜式字集資源架構來解決缺字問題，本文主要介紹可攜式字集資源架構。

在開始介紹可攜式字資源架構之前，我們先使用下面五段來介紹五個詞：“缺字資源”、

“內碼字集資源”、“文字識別碼”、“缺字完備文件”及“將缺字本體資訊嵌入作業系統”。

缺字資源：為了解決缺字問題，我們當然應該先討論一下文字及缺字。何謂一個文字，是一個字碼，還是一個字型？在內碼系統中，我們固然可以以一個字碼(例如 BIG5：B56F)約定俗成地代表眾人心中對某一個文字(例如:發)的概念，可是我們現在處理的是缺字，也就是失去了內碼系統約定俗成的功能。所以在這個情況下，何謂文字？我的想法就是該文字的基本資訊，包括該文字的各式字型、各式輸入法、各式發音、各式用法，各式語意、辭彙，各式內碼及與其他文字的各式關係，如近似形、近似音、近義或同義等文字關係。我們總稱這些文字基本資訊為該文字的本體資訊。所以我們認為利用本體資訊來定義一個文字是一個合理的方法，特別是用來定義缺字。為了解決缺字問題，我們提出“可攜式文字資源”觀念，此方法是將一個或多個文字的本體資訊結合成為網路資源，放在網路上供人使用，我們稱內含文字本體資訊的網路資源為文字資源，如果該資源內含缺字，我們亦稱之為缺字資源，這裡我們特別強調，我們所謂的文字資源中的文字不一定要都是缺字，我們也可以定義一些內碼已經定義過的非擴充區碼，所以即使整套內碼中的所有字碼都重新定義過，這樣扔就是一份我們所謂的文字資源。

內碼字集資源：可攜式文字資源包含一群文字的本體資訊，若其中的文字擁有內碼的設定(例如：BIG5:FA40)，我們特別稱之為可攜式內碼字集資源。有了內碼字集資源的概念，我們就可以設計一整套的內碼字集或是內碼擴充字集(例如:BIG5)，並且指定各字碼的各式字型、各式輸入法，各式基本資訊等，然後放在網路上，且利用網路上的位址為該內碼字集資源命名。所以說可攜式內碼字集資源亦是一種可攜式文字資源。

文字識別碼：在內碼系統中，文字是以內碼為識別碼。對於缺字來說，同一台電腦系統中，擴充區中的缺字內碼仍可做為缺字的識別

碼，但對於跨系統間，缺字內碼就無法做為缺字的識別碼。所以為了在跨系統間能夠正常辨識缺字，我們需要一種文字識別碼使其缺字識別碼在網路上是可辨識的。例如使用 CNS11643 交換碼，或是使用缺字資源的 URL(Uniform resource Locator)。文字識別碼是文字識別的基本單元，我們將識別碼分為兩種，一種是本機文字識別碼，意即在本機系統中識別文字的方式，例如擴充區字碼；另一個是網路文字識別碼，意即在網路中識別文字的方式，例如 CNS11643 碼或是缺字資源的 URL。

缺字完備文件：文件係指使用數位記載文字序列的數位檔案。例如網頁或是其他各式文書應用程式所使用之檔案。內含缺字的文件，我們稱之為缺字文件。為了解決缺字問題，在文件中使用缺字時，必須明顯地或隱含地，但明確地，指出或內嵌所需的缺字資訊，我們稱這種文件為缺字完備文件。但是因為各應用程式有自己的檔案格式，所以並不是每一種文件皆可以(不經由原應用程式廠商)輕易修改成缺字完備文件。例如對於網頁 HTML 格式來說，我們可以增加一個標記來指示缺字資源所在 <pws src=abc.pws>，所以網頁 HTML 格式可以輕易地被修改成為缺字完備文件。

將缺字本體資訊嵌入作業系統：是指將缺字本體資訊中的各式字型、各式輸入碼安裝至作業系統中使得其他應用程式不需要做任何的變更就可以透過作業系統來顯示缺字，可以透過作業系統控制的輸入法來輸入缺字。“將缺字嵌入作業系統”，除了指將字型及輸入碼安裝至作業系統中以外，也指將缺字的其他本體資訊放至系統的某個位置供其他應用程式使用，但此時我們並不保證應用程式可以不做任何變更就可以使用缺字的其他本體資訊。

所謂可攜式字集資源架構，其方法如下：

當電腦在瀏覽缺字完備文件時，軟體會自動取得該文件所指示的缺字本體資訊，並且將缺字本體資訊嵌入作業系統中，使得其他應用程式

也可同時順利使用這些缺字，並且能夠利用這些缺字來製作缺字完備文件，放到網路上流通。

因在缺字碼動態指定技術[11][11]中，某缺字的缺字碼並不是在各個電腦中皆相同，故在系統當機後會造成文件中的缺字解釋完全遺失，故進而發展可攜式字集資源架構--缺字碼靜態指定技術。在靜態指定技術類中，最好的當然是全字內碼法，但全字內碼法只能減少但不能解決缺字問題，而前全字內碼法需要長時間的審查，所以我們仿照網際網路上的資源概念，建設文字資源網路並且依藉大部分應用程式支援各式字型的特性來發展靜態指定技術。

我們發現現今大部分的文書應用程式皆有支援字型的設定，所以我們加入擴充字集法的概念，發展管理字集的邏輯架構。首先，我們使用網路上常用的 Uniform Resource Locator 來定義字集的名稱，並且更重要的，我們使用字集的名稱來定義字集所連結的字型，如此當支援字型的應用程式使用字型時，我們將可從字型名稱中取得字集名字，進而能夠辨識該文字。所以在這裡，可攜式內碼字集資源在 windows 作業系統中，實作上就是以內碼系統為中心，包含一些字型檔(.ttf)，一些輸入法擴充碼(.tbl)，及其他我們自定的本體資訊。所以靜態缺字碼指定技術是以字集名稱加上內碼擴充碼來作為文字的識別碼。

例如中國時報網站中的各式文件中含有缺字，中國時報可以設計一套 miss 字集，名稱為 miss@www.chinatimes.com，別名為“中國時報缺字集”，而在該字集中定義各式字型如“標楷體_miss@www.chinatimes.com”，“細明體_miss@www.chinatimes.com”，或是自定的“棋盤體_miss@www.chinatimes.com”，並且在其有支援字型的應用程式文件中使用之。那麼瀏覽者只須安裝 miss@www.chinatimes.com 字集(中國時報缺字集)即可解決瀏覽問題。基於這樣的想法，靜態缺字碼指定技術在程式寫作方面，最主要是發展自動化可攜式內碼字集資源的工具。所以接下來我們依序為各位介

紹本技術在瀏覽、輸入、傳播、查詢、造字、紀錄等問題上的解決方法。

瀏覽程序：

```
http://www.chinatimes.com/news.htm
<html><head></head><body>行政院長游錫
<font face="標楷體_miss@www.chinatimes.com">
FA40</font>的
<font face="細明體_miss@www.chinatimes.com">
FA40</font>是個 BIG5 的缺字。
</body></html>
```

以上方文件為例，“堃”是一個 BIG5：FA40 的缺字，所以在寫作上就已經將它的字型特別選取為“標楷體_miss@www.chinatimes.com”，當使用者開啟這份網頁時，我們安裝在瀏覽器上的 plugin 會從網頁文章中抽取字集名稱，並且從名稱中確認字集資源在網路上的位址，然後自動從位址下載字集資源檔並安裝取用其中的各式字型，(例如標楷體_miss@www.chinatimes.com)，然後依正常程序即可正確瀏覽該文件。至於應用程式的文件來說(如 Word 應用程式的 doc 文件)，我們也不需要任何的變動，因為我們假設此分外來文件是由瀏覽器下載而得，則該瀏覽器必定瀏覽過該網站，自然安裝過該網站的字型，所以也可以正確瀏覽其他應用程式的文件。雖然如此，我們依舊盡我們所能為各式應用程式發展辨識字集及自動下載安裝字集的 plugin 或是 macro，例如，我們可為 Microsoft Office 軟體寫 Visual Basic for Application 程式，讓 Office 文件在開啟時即可辨識文件中缺字所使用的字集名稱，並自動下載安裝瀏覽之。

輸入程序：既然靜態缺字識別碼指定技術以字集名稱加擴充碼為全域性文字識別碼，而以擴充碼為區域識別碼，所以在輸入法上也有不同的做法，一種是輸入法程式輸出“字型名稱加擴充碼”(注意這裏確實是“字型名稱加擴充碼”而非“字集名稱加擴充碼”)來同時告知應用程式該字內碼及選用的字型，只是這個方法目前不可行，因為各式應用程式皆未支援，另外

一種是讓使用者自己明確地或是隱含地選定所使用的字集，在有支援 plugin 或是 VBA 等外掛程式的軟體，我們會利用該技術來鑑定目前游標所在字元的字型來告訴輸入法將輸入的文字的字集，然後輸入法就使用該字集的輸入法擴充碼來檢索文字，然後輸出該文字的內碼(不含字集資訊)。在這裡我們如果從“字集加上擴充碼”當成文字識別碼的角度來看，系統並不會出現同碼不同字的問題，但是如果從“內碼”當成文字識別碼的傳統角度上來看，則本系統發生罕見的系統內同碼不同字問題，所幸這個問題在檢索問題以外的瀏覽、輸入、傳播等問題上並不嚴重。

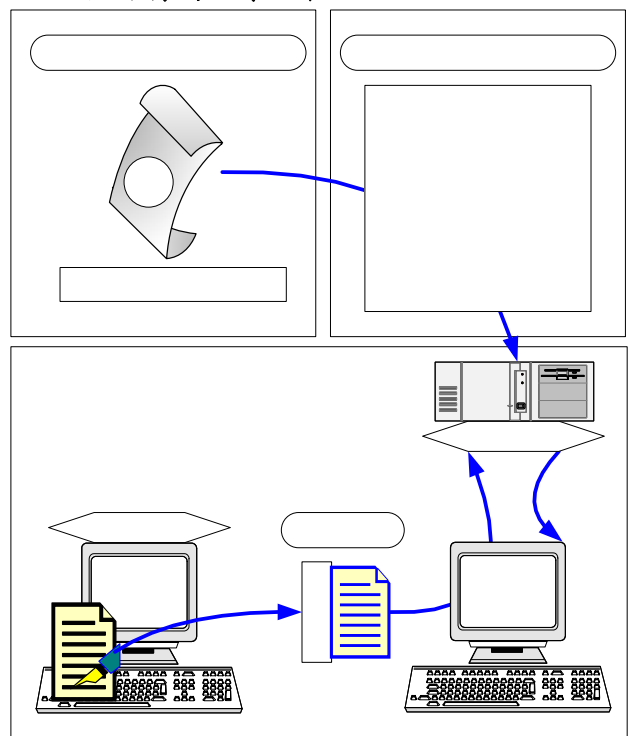
傳播程序：在傳播程序中，我們假設文件大部分的文件皆來自於瀏覽器網路下載，所以在瀏覽文件之前應該已經下載安裝過使用字集，所以應該可以正常瀏覽。雖然如此，我們依舊將傳播過程分成兩大類，第一大類是可寫巨集指令或可附加模組的應用程式，或公開文件格式者，例如在 Microsoft Office 上可以使用 Visual Basic for Application 來寫巨集指令，有例如在 Microsoft Internet Explorer 或 Netscape Navigator 上可以附加 plugin 程式。因為這一類程式可以利用技術修改其行為，所以我們在接到該應用程式文件時，我們可以掃描該文件取出字型名稱，並依字型名稱中的機器名上網去取得該文字本體資訊。而第二類則是我們無法方便地更動應用程式的行為，此時將包裝其文件格式，我們將文件傳播時需合併缺字資源包裝成一份檔案 pod(portable document)檔，在傳播後，我們的軟體會將 pod 檔解開或下載文字本體資訊並嵌入系統中即可。

查詢程序：在查詢程序中，本方法造成了罕見的同碼不同字現象，但這並不嚴重，因為缺字本來就是一個少見的情況。我們一樣可以利用前後文的方法來進行文件的全文檢索。如果要使用到缺字本體資訊，則需要視搜尋引擎的設計而定，需要在傳遞內含缺字的查詢語句的時候連缺字本體資訊一起傳遞才能利用之。

造字程序：在同字不同碼方面，可攜式字集資源架構可謂是完全棄守，不過不要緊的是，該架構提供了無限的字碼空間，所以並不怕字碼浪費，只是會造成查詢的困難。至於在典藏過程中，我們仍然認為需要統一造字以將領域內缺字清楚整理。本架構並不進行造字工作，主要是造字者使用坊間造字軟體進行造字工作，然後使用本軟體將字型檔(.tte)及輸入法擴充檔(.tbl)包裝成可攜式字集資源(.pws)。

描述程序：本架構認為典藏過程中造字工作應該統一，但對紀錄描述缺字的技術，則完全沒有，我們認為謝清俊老師的部件組字法是最佳的缺字描述方式。

5. 一個可行的缺字流程



目前，缺字大都來自於研究領域與典藏機構，所以我們提出在典藏過程中可行的缺字流程。在典藏過程中，我們建議使用謝清俊老師的部件組字法來描述缺字，在所有文件都繕打完畢之後，整理各個缺字，進行統一造字製作可攜式字集資源，並產生一份完全內碼而沒有

部件組字式的文件庫，然後使用可攜式字集資源來解決瀏覽、輸入、傳播等問題。

6. 討論與結論

全字內碼法的優點在於大量減低缺字發生的頻率。擴充字集法的優點在於完全結合字集及字型概念，在作業系統中得到完全的支援，可被眾多應用程式使用。字型資源法的優點在於每項資源都有唯一的網路地址可供識別及指示資源所在。部件組字法的資源在於不需立即造字，即可馬上描述缺字長相，可描述無限多個缺字。統一造字法的優點在於可管理缺字，可易於專家造字使得造字可以完整。

可攜式字集資源法可謂是結合了擴充字集法及字型資源法的優點整合而成。在流程中搭配部件組字法及統一造字法，再加上原本的全字內碼法。所以說，這個本文所提的缺字處理流程可謂是集各方法之優點而成，且簡單易懂。

本文提出了可攜式字集資源架構用以解決缺字問題。文中探討並解決缺字問題的瀏覽、輸入、傳播、查詢、描述、造字等子問題。最後提出一個合併各家優點且簡單可行的缺字處理流程。

7. 誌謝

感謝徐代昕、林聖凱、蔡長恩、何浩洋、蔡雨利等人參與實作與討論。

8. 參考文獻

- [1] CNS11643.
<http://www.cns11643.gov.tw/web/index.jsp>
- [2] Microsoft Typography.
<http://www.microsoft.com/typography/default.asp>
- [3] Open Type Jamboree.
<http://www.microsoft.com/typography/jamboree/default.htm>
- [4] Portable Font Resource.
<http://www.bitstream.com/categories/developer/fontfusion/pfrs.html>
- [5] The Free Type Project.
<http://www.freetype.org>
- [6] True Type Topography.
<http://www.truefont.demon.co.uk>
- [7] Unicode. <http://www.unicode.org>
- [8] 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”，第二次兩岸古籍整理研究學術研討會，北京大學，北京，1998年5月11-13日。
http://www.sinica.edu.tw/~cdp/paper/1998/19980511_1.htm
- [9] 滿天星電腦股份有限公司，“字霸軟體”，
<http://www.astar.com.tw>
- [10] 威鋒數位 Dyn@File.
http://tw.dynacw.com/product_dynafire.htm
- [11] 謝育平、項潔，“可攜式文字資源--用以解決中文的缺字問題”，第三屆海峽兩岸科學技術學術研討會，中正大學，嘉義，台灣，2003/10/27-28。