※「數位時代的出土文獻」專輯※

互聯網時代出土文獻數據庫建設的思考與實踐

張再興 *

上個世紀九十年代，隨著電腦使用在中國的普及，古老的古文字研究也迎來了新的發展契機。從上個世紀末開始，我們即開始了先秦出土文獻的數位化研究，十多年來在這一領域進行了不懈的探索，解決了出土文獻數據庫建設中的一系列問題，開發了多種出土文獻數據庫。而互聯網時代的快速到來，又為出土文獻的數位化研究和應用提供了更為廣闊的發展空間。本文希望能夠介紹我們這些年對於出土文獻數據庫開發的一些想法和實踐。

一、徹底的數位化：出土文獻數據庫建設的必備基礎

我們在從事出土文獻數位化研究之初，就確立了數位化開發的徹底數位化原則。所謂「徹底」，是指對所有出土文獻的所有文字進行編碼，使之都能通過電腦進行顯示、檢索，從而完全摒棄傳統古文字在電腦寫作、編輯、出版等過程中常見的貼圖、替代（以編號替代、以方框替代等）、構件拼湊等方式。

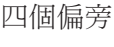
要實現徹底的數位化目標，首先需要對古文字進行造字處理，而造字首先涉及到字庫的電腦編碼問題。現有操作系統的漢字編碼空間是十分局限的。如 Windows 95 開始支持的 GBK 編碼只有二〇九〇二個漢字。雖然在 Windows 2000 中開始支持擴展 A 和擴展 B 的字形顯示，字形數量達到了六五五三一個，但是這兩個部分的字形在數據庫處理中仍有不少的缺點。

在這樣狹窄的空間裏無法同時容納所有的古文字類型。因此，我們只能採取字

* 張再興，華東師範大學中國文字研究與應用中心教授。

體的方式，建立各種類型的古文字字體。到目前為止，我們開發的古文字字體計有甲骨文、金文、秦文字（包括石刻和簡牘）、楚簡帛文、古璽印文、古陶文、侯馬盟書、貨幣文、《說文》小篆等近十種古文字字體，囊括了目前已見的所有古文字類型。上述字體部分已經通過軟體光碟的形式正式出版[1]，部分在我們網站上可以免費下載[2]。

但是字體形態的古文字字庫有著先天的弱點，難以將多種古文字字形統一編碼，從而使各種古文字數據庫的整合存在巨大的困難。隨著 Unicode 國際編碼體系對古文字編碼的展開，在海內外古文字學者的共同努力下，相信這一問題在不久的將來可以得到徹底的解決。

古文字材料在電腦處理過程中需要使用的古文字相關字形具有多個層次。為了確保徹底數位化原則能夠得到充分的貫徹，我們開發的各類古文字字體包含各個層次的字形。首先是古文字原形，依據原始字形拓片掃描造字，確保所造古文字字形的精確性。除了少量拓片模糊無法造字之外，一般不對字形拓片做技術處理。由於古文字字形的手寫特徵，使得字形的異寫普遍存在。因此，同一個字往往擁有眾多的字形。對這些字形全部進行造字既無必要，也無編碼空間的實際可能。因而要根據字形的區別性原則有選擇地造字[3]。其次是古文字的隸定字形。古文字原形字的準確隸寫是古文字釋讀的基礎，也是古文字數位化處理的重要方面。我們根據構件對應的原則，對能夠隸定的古文字字形進行比較嚴格意義的隸寫字形造字。由於古文字研究中常常涉及到古文字偏旁的結構分析，因此，古文字字庫還需要包含另外一個層次，即偏旁構件的字形。我們開發的金文字庫即包含了金文中拆分出的五三四個偏旁，如：  等。

古文字字體的建立只解決了字形在電腦中的顯示問題，更為重要的是如何讓用

1　已經正式出版的字體有金文字體，見臧克和、劉志基、張再興等著：《商周金文數字化處理系統》（南寧：廣西教育出版社、廣西金海灣電子音像出版社，2003 年）；戰國楚文字字體，見臧克和、劉志基、張再興等著：《戰國楚文字數字化處理系統》（上海：上海教育出版社，2003 年）；《說文》小篆字體，見臧克和、王平等著：《說文解字全文檢索系統》（廣州：南方日報出版社，2003 年）。

2　見華東師範大學中國文字研究與應用中心網站：http://www.wenzi.cn/tiyao090104/Index.asp。

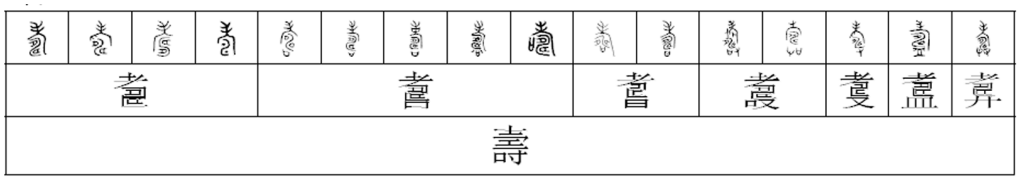
3　有關選擇原則，詳見張再興：〈古文字字庫建設的幾個問題〉，《中文信息學報》，2003 年第 6 期。

戶檢索和輸入古文字。為此我們開發了相應的古文字字形輸入編碼[4]，並編製了古文字輸入法軟體，已經正式出版的有金文輸入法和戰國楚文字輸入法。

除了標準輸入法以外，我們還編製了基於偏旁檢索的古文字字形檢索文檔，以供查找各類古文字字形[5]。

另外，各層次的字形之間也必須建立起對應關係，才能更好地使用這些字形。如原形字與相應的隸古定字形之間的對應，原形、隸古定與楷書字頭之間的對應等，從而建立起字形關係樹。特別是對於一些異體字形眾多的字而言，這種對應顯得十分必要。如表一顯示了金文中「壽」字的眾多原形及其對應的隸古定形體。基於這樣的對應關係，我們開發了便捷的楷書與古文字原形字之間的對應轉換系統，可以在 Word 操作系統中實現兩者之間的互相轉換[6]。

表一



二、適應出土文獻特徵的釋文資料庫

處於古文字階段的先秦出土文獻，字形多變，同時，出土文獻又是一種手寫的文字系統。要讓這種文字系統被電腦接受，首先要進行全文隸寫釋文，建立具備全文檢索功能的出土文獻釋文資料庫。

出土文獻字形的釋讀一般可以分為三個層次：第一個層次即是按照原始文字結構對文字進行隸定。古文字的隸定一向有寬嚴之別；為了最大可能地反映古文

4　劉志基：〈簡說「古文字三級字符全拼編碼檢字系統」〉，《辭書研究》，2002 年第 1 期。

5　依據《說文》部首編排的金文字形庫檢索文檔網址：www.wenzi.cn/jinwen/font/jinwenbushou.htm。

6　見《商周金文數字化處理系統》光碟、《戰國楚文字數字化處理系統》光碟。

字的原貌，我們將採用嚴式隸定，即以構件為最小隸定單位，除了構件內部的線條差異之外，文字構件的增減替換、構件方位的變化等都可作為區別標記。如我們對金文中各種結構的「鑄」字的隸定：（[illegible]）、（盄）、（[illegible]）、（[illegible]）、（[illegible]）、（[illegible]）、（鉊）、（[illegible]）、（[illegible]）、（[illegible]）、（[illegible]）、（[illegible]）、（[illegible]）、（[illegible]）、（[illegible]）、（[illegible]）、（盥）、（盅）、（[illegible]）、（鑄）。這樣，不僅可以儘量重現文字結構的原貌，同時也便於進行異體字形的各項研究，如斷代、分域等研究。當然，要完全做到這一點很難，因為金文書寫的不規範性，使構件形體、方位等有時很難認定。第二層次是在字形準確隸定的基礎上確定這個字形是後世的何字，具體來講是在《說文解字》或後世字書中找到對應的字。這個字可以稱為「考釋字」。第三個層次是部分字在出土文獻中按照本字讀不通，需要破讀。這個破讀後的字可以稱為「破讀字」。破讀的情況在出土文獻中很常見。

因此，出土文獻釋文庫中的釋文其實應該包含三個層次的釋文：隸古定釋文、釋字釋文、破讀釋文。當這三個層次的釋文整合在一起時，需要有必要的符號區分其中各個層次的文字。由於各個層次的字形具有關聯對應，所以釋文的全文檢索可以在隸古定字、釋字、破讀字各個層次中進行。

只要字形足夠清晰，第一個層次的隸定引起不同看法的情況相對而言比較少。但是考釋字、破讀字的意見分歧經常很大。顯然，出土文獻的釋文庫應當能夠反映這種分歧。

出土文獻的研究離不開文字拓片的認讀。各種文字類型的拓片圖版及其相關考古學屬性是出土文獻的最基本特徵。有關出土文獻的數據庫也必須完整反映這種特徵。例如，青銅器銘文，除了拓片之外，每個青銅器還有出土地點、器形類別、器組關係、時代、現藏地點等屬性。金文研究歷史悠久，數百年來，金文拓片還有眾多的著錄屬性。這些屬性無論是對於青銅器研究，還是對於金文研究，都是必不可缺的，都必須在釋文資料庫中加以反映[7]。

釋文庫的釋文還應具有動態的全文檢索功能，並能自動編纂各種形式的逐字引得[8]。

7　詳參張再興：〈簡論「金文語料庫」的設計〉，《辭書研究》，2002 年第 3 期。

8　詳參張再興：〈論古文獻引得的計算機自動編纂〉，《辭書研究》，2002 年第 5 期。

目前，我們已經完成了大部分先秦出土文獻的釋文庫，內容包含殷商甲骨文（包括《甲骨文合集》、《花園莊東地甲骨》、《小屯南地甲骨》）、周原甲骨文、商周金文、楚簡（包括包山楚簡、郭店楚簡、上海博物館藏戰國楚竹書、曾侯乙墓竹簡、九店楚簡、長沙子彈庫楚簡、望山楚簡、信陽楚簡等）、秦簡（包括睡虎地秦簡、龍崗秦簡、周家臺秦簡、放馬灘秦簡）、古璽印文、古貨幣文、石刻文、古陶文、侯馬盟書等。其中，商周金文、戰國楚文字我們已出版了數據庫光碟，《花園莊東地甲骨》在網上發布了檢索系統[9]。新發表的嶽麓秦簡等的釋文庫建設工作也正在開展。

三、面向語言文字研究深度開發的語料庫

面向語言文字研究的出土文獻數據庫需要能夠滿足文字、詞彙、語法各方面研究的需要。為此，我們在出土文獻釋文資料庫的基礎上進行進一步的標注，開發了面向文字研究的字形屬性庫、面向詞彙研究的詞義屬性庫，以及面向語法研究的語法屬性庫。

（一）字形屬性庫

字形屬性庫提供文字學研究所需的出土文獻字形相關屬性。具體包括以下內容：

1. 單個字形的原始拓片。根據已發表出土文獻拓片剪裁而成。

2. 該字形的隸定、釋字以及破讀。

3. 字形的結構分析。對字形進行文字學意義上有理據的立體結構分析，標記各層次的構件及構件所在方位。例如甲骨文  字，隸定作  ，其結構可以分析為：庶 s < 火 d 石 t > 众 x < 人 z 人 m 人 y >。其中，< > 表示層次，構件後面的字母代表其在字形中所處的方位。

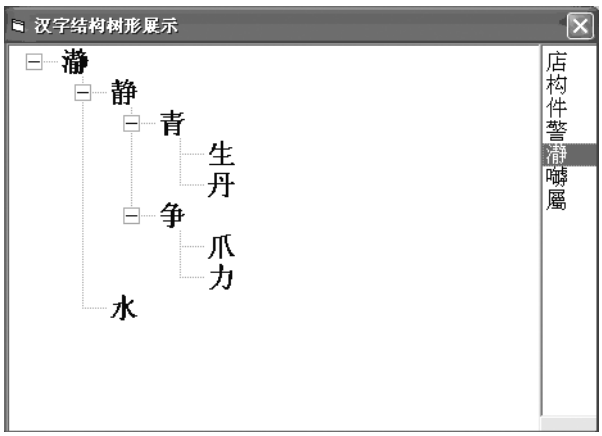
對這樣的標記進行電腦自動分析後，可以自動構建文字系統的構件屬性庫[10]，該庫包含構件的名稱、構件所在上級構件名、構件所在層次、是否底層構件、構件方

9　網址：http://www.wenzi.cn/huadong/index.htm。

10　詳參張再興：〈漢字結構統計分析系統〉，《中國文字研究》第三輯（南寧：廣西教育出版社，2002 年）。

位等屬性。該屬性庫可以用作構件的統計分析研究，也可呈現圖一所示字形的樹形結構。

圖一



構件的功能標記也是字形結構分析的重要內容。某個構件在字中表意、表聲，或者只是一個綴加的構件，直接影響到文字的釋讀。傳統「六書」分析也需要在對構件進行功能分析的基礎上進行。

4. 字形的各種關聯與篩選。如金文字形屬性庫與青銅器斷代庫的結合，可以排列某個字形或某種構件的標準器字形。再如幾種古文字字形屬性庫的關聯，可以對比各類出土文獻的字形。

5. 古音標記。採用王力先生的古音系統，關聯漢字古音屬性庫，並參考形聲字的聲符關係等確定詞的語音屬性。以後將逐步增加其他古音系統，以資參照。

（二）詞義屬性庫

出土文獻詞義屬性庫提供詞彙的相關數據。具體包括以下內容：

1. 詞形。上古漢語雖然以單音詞為主體，但是複音詞也已經大量出現。而兩周時期正是複音詞形成的重要時期。因此，尚未結合成詞的固定結構也先作為一個成分切分，標記出其獨特屬性。這既可以避免複音詞切分標準的爭論以及複音詞切分的遺漏，又可以為複音詞發展過程的研究提供重要的材料。

2. 詞的語法結構。複音詞兩個語素之間的語法結構關係是詞彙研究的重要內

容。對複音詞要自動測量詞的長度，並標記並列、偏正、修飾等語法結構屬性。

3. 詞義。根據已有考釋研究成果，在上下文語言環境中歸納出的詞義，也就是「義項」。用簡明的現代漢語來說明。

4. 詞義的意義類別。詞義的意義類別標記將同一類的意義彙集到一起，對準確理解詞義和研究同義、反義、類義等詞義關係提供更為便捷的途徑，也為出土文獻的專題研究提供詞彙基礎。

5. 搭配。從位置上看包括前搭配和後搭配，從功能上看包括語法搭配和語義搭配等。

（三）語法屬性庫

語法屬性庫在全面標記詞的詞性、句法功能的基礎上，提供相應的語法研究檢索及統計。具體包含以下內容：

1. 詞性。

2. 句法功能。

3. 句子類型。

目前，我們已經完成有關字形屬性庫的構建，詞義屬性庫和語法屬性庫的建設也已開始進行；特別是詞義屬性庫，已經完成金文、楚簡、秦簡的詞義、詞性等標記。由於這兩個屬性庫的工作量和工作難度巨大，這一工作仍將是我們未來多年的主要工作目標。

四、網路數據庫的獨特優勢

互聯網的普及讓古文字輕鬆地走出了學者的個人書齋、個人電腦，讓各層次的大眾參與其間。因此，互聯網已成為出土文獻數據庫的最好發布載體。

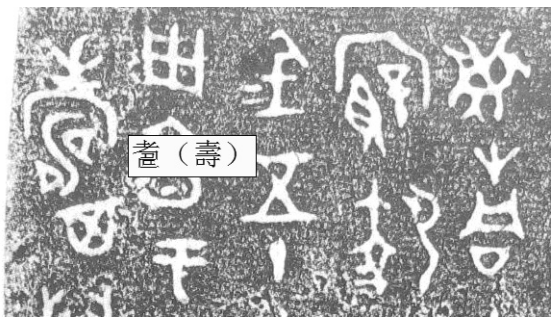
（一）普及與研究的兼顧

出土文獻資料並非只有古文字等專業領域的專家在使用，即使普通大眾也常有學習、查閱古文字字形等出土文獻資料的需要。因此，除了針對語言文字研究的專業數據庫外，還需要面向各層次用戶提供各種檢索資料。

例如，面向普通用戶的金文閱讀需要，我們在金文網路數據庫中開發了金文拓

片釋文即點即現的功能，當游標放到金文拓片的某個字形上，即可顯示該字形的相應釋讀楷書[11]。圖二顯示游標放到拓片字形上時，顯示其釋文為「耆（壽）」。

圖二



再如，針對普通用戶的古文字字形查閱要求，我們開發了常用古文字字形網上檢索系統[12]。輸入楷書，即可檢索相應的各種古文字字形。

（二）多層次多方位的關聯對應

互聯網時代數據之間的自由跳轉，徹底改變了傳統工具書時代的資料檢索方式。基於網路高度集成的出土文獻數據庫，同樣也為數據庫中各類資料的整合和關聯提供了極其便利的條件。

1. 原始文獻與各家釋讀的關聯：

如前所述，古文字的釋讀常存分歧。因此，原始文獻與各種釋讀研究文獻的高度集成關聯，是我們對出土文獻語料庫的基本要求。這種關聯既要求顧及到原始文獻各層次的語篇單位，如意義完整的語段、天然的載體單位（如青銅器的器單位、簡帛的簡單位等），也要求顧及到字詞的單位。而字詞單位的關聯既要關聯當前字形的釋讀研究，又要關聯該字出現在其他文獻中的釋讀研究情況。

因此，針對不同類型的釋讀文獻，具體的關聯方式應有所不同。如對於金文

---

[11] 2002 年夏含夷教授來訪問時，曾談起過在拓片上即時顯示隸定楷書的問題。2009 年 10 月，在芝加哥大學東亞系召開的一次學術研討會上，鳳儀誠先生也提出這種希望。本節所述金文網路數據庫網址：www.wenzi.cn/jinwen/jinwen.htm。

[12] 網址：http://www.wenzi.cn/guwenzizixingjiansuo/guwenzizixingjiansuo.htm。

研究來說，著錄類、青銅器斷代類研究文獻宜對應到器。文字編類、字典詞典類、字詞專題研究類文獻則以不限定具體出處的字關聯比較合適。文獻通釋類、集釋類文獻可以對應到具體出處的字詞。而詁林類文獻的情況比較複雜，有些能對應到具體器中的字，有些則只能對應到字。

為了便於檢索，各家集釋用簡明提要的方式展示，同時連結原始的考釋文獻[13]。在金文網路數據庫中，我們將各家考釋關聯到相關字的注釋下顯示。

2. 個體、樣本與全體的關聯：

某個青銅器或某支簡中的某個字形只是出土文獻中這個字的眾多字形中的一員。在字形上，它與其他字形的不同可能只是書寫的不同，也可能存在結構的不同。在時代上可能存在前後變化關係。各種文字編著重於選擇羅列代表性的字形。在出土文獻數據庫中，則可以把具體字形、代表字形、全體字形三者關聯起來。我們的金文網路數據庫在點開金文拓片某個字形的連結後，分時代顯示金文字形的字形彙編。其中第二欄顯示《金文編》已收字形，第三欄顯示《金文編》未收的其餘全部字形（見圖三）。

圖三

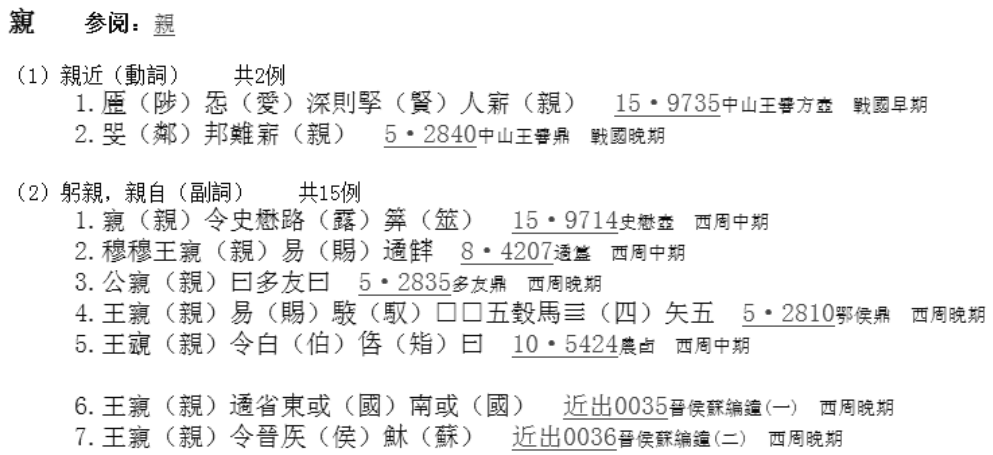
3. 個體字詞與整體語篇的關聯：

出土文獻的語篇由一個個的字詞組成。個體字詞與整體語篇之間的雙向關聯能為使用者提供巨大的便利。而這正是傳統紙質工具書所無法企及的。我們在金文網路數據庫中提供幾種用於關聯的超連結：(1) 字形彙編：拓片中的字可以連結到該

[13] 由於版權方面的限制，原始文獻在網路上公開展示的只能是其中沒有版權問題的部分。

字的字形彙編，每個字形出處都可連結到各自的器銘界面。(2) 文例彙編：器銘各字釋義下的字可以連結到文例彙編界面，此界面是傳統詞典和引得的結合，以各字的義項為單位，窮盡列舉各義項的所有語言環境（見圖四）。通過文例的出處也可跳轉到各自的器銘界面。

圖四



4. 字與字的關聯：

先秦文字系統的劇烈發展，使得當時文字系統中字與字的關係十分複雜糾結，許多字在當時文字系統中的關係與後世並不一致。在許多關係難明的情況下，關聯相關字的材料自然可以提供更多的研究便捷。如《說文》將同訓為「至也」的窺、親分成兩字，《金文編》沿此分立字頭。但是在金文中「窺」都用作「親」，表示親自。再如，「命」、「令」後世分化成兩字，但在金文中的使用也很複雜。我們在各字的字形彙編、文例彙編中提供相關字的連結，以便參閱（見圖三）。

5. 構件與整字的關聯：

從整字可以獲知相應的構件組成，從任何一個構件可以獲知此構件所組成的所有字。

（三）傳統文本與現代網路的結合

網路數據庫跟傳統文本相結合，是我們在出土文獻數據庫建設中的又一嘗試。

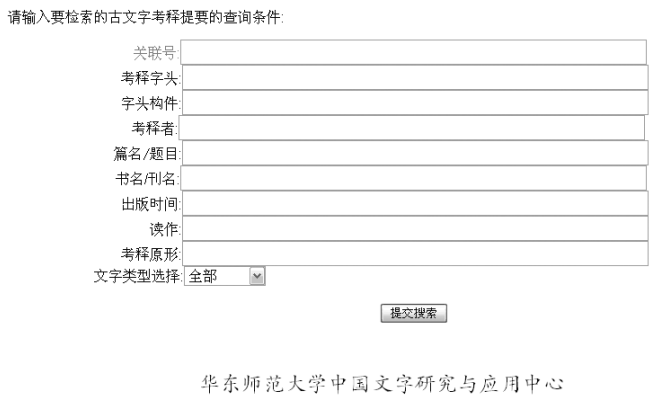
網路數據庫的超強檢索能力及其海量的儲存能力能夠有效彌補傳統文本工具書的一些先天不足。

比如詁林類工具書的編排，可以採用考釋結果作為字頭，也就是不管是什麼字形，考釋者認定是什麼字就放到什麼字下。也可以採用所釋原形，也就是考釋對象作為字頭。無論採用哪種方式，都有弊病。前者所考釋的字頭下涉及的字形五花八門，後者又看不到考釋為某個字的究竟有哪些形體。而這兩種方式在傳統紙質工具書中又不能同時進行。也不能根據考釋對象排一遍，再根據考釋結果又排一遍，這樣很不經濟，最多只能用增加檢索來解決。

我們通過電腦網路的方式比較好地解決了這個問題。我們出版的《古文字考釋提要總覽》[14] 以提要的形式彙集了前人對古文字考釋的研究成果。這部書涉及的古文字類型也很多，包括甲骨文、金文、簡帛、璽印、石刻等。因此在正文編排上也是採用考釋結果的編排方式，也就是考釋者考釋成什麼字就放到什麼字頭下。這在文本編輯方式下是比較便捷的處理方式，而為了彌補這種編排方式的缺陷，我們又開發了與這部書配套的網路檢索數據庫 [15]，可以根據原字形、字形出處、考釋者等進行檢索。檢索界面見圖五。

圖五

《古文字考释提要总览》关联检索（测试版）



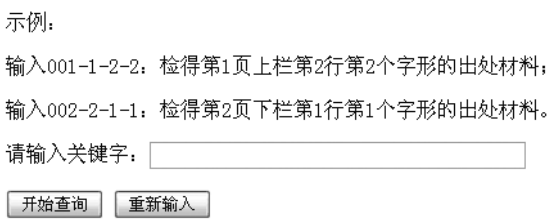
[14] 第一冊、第二冊已由上海人民出版社於 2008 年、2010 年出版。

[15] 網址：http://www.wenzi.cn/tiyao090104/Index.asp。

再比如說傳統古文字文字編，在一個字下列出代表字形，以供參考。但是缺乏相關上下文，對於形體的釋讀以及意義的考察不能不說是很大的缺陷。前人為彌補這種缺陷也做出過種種努力。如《金文大字典》在每個字形下摘抄文句，《古璽文編》與《古璽彙編》配套使用，《秦文字類編》下編是文句摘抄等。但是這種方式的不方便依然是顯而易見的。

我們出版的《中國異體字大系——篆書編》[16] 同樣採用了網路配套形式。選錄的每個古文字字形所處的原始文獻都可以通過網路檢索得到[17]。檢索界面見圖六。

圖六



[16] 劉志基、張再興主編：《中國異體字大系——篆書編》（上海：上海書畫出版社，2007 年）。
[17] 網址：http://www.wenzi.cn/pages/guanlianshuxi/YT_Index.htm。