漢籍電子文獻資料庫數位化工作流程簡介

製作日期：2005/11/30

計畫單位：中央研究院歷史語言研究所漢籍工作室

計畫名稱：中央研究院漢籍全文資料庫

計畫簡介：

古籍是歷代流傳下來的文化瑰寶，因年代久遠，加上種種破壞與耗損，使得大多數的古籍難以完整保存，能夠保存下來的古籍自然更顯珍貴。因此整理及保存古籍的完整性是一項非常重要，且需長期投入的工作。

鑑於古籍數量龐大，加上善本取得不易，匯集古籍的工作十分困難。從搜集、編目到進行各種研究，都必須花費相當的人力與物力；而人力不足及人工作業疏失，有時難免造成缺誤。

古籍電子化後，透過電腦的處理及全球網際網路的優越性，這些資料可無限制地被使用者利用。再者，使用計算機進行處理，可以進行大量且連續的操作，將資料匯集起來，經過學者專家相互的比對參照，常能發現前人所未見的新資訊，所以古籍電子化對於研究工作是極爲重要的突破。

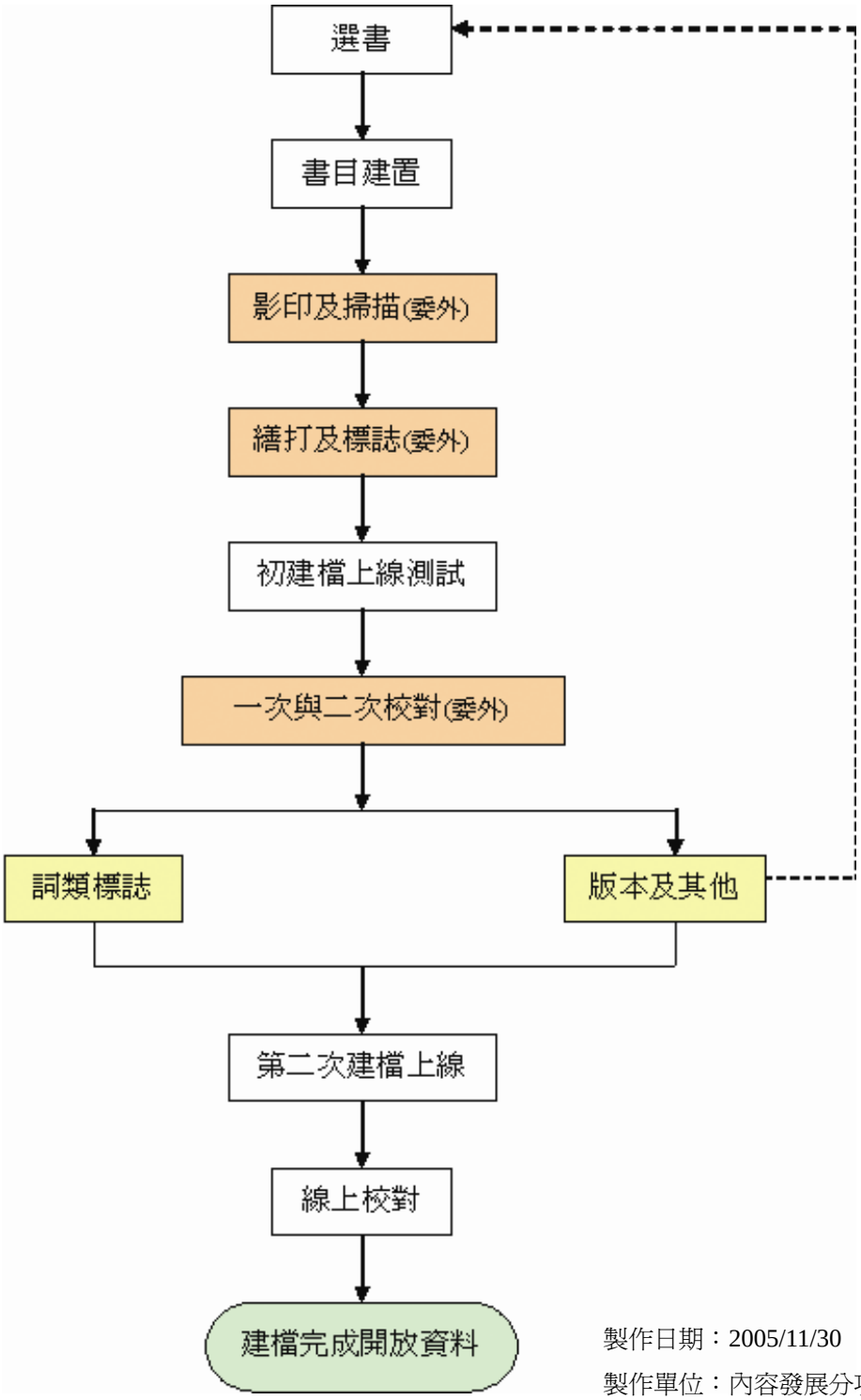
中央研究院歷史語言研究所（以下簡稱史語所）與中央研究院計算中心於 1990 年完成共同開發的二十五史資料庫，於 1995 年將 WWW 檢索程式上線命名爲「瀚典全文檢索系統」，1997 年瀚典改版至 1.3 版，但爲了因應電腦軟硬體不斷擴充與使用者需求，更爲精益求精，在現任主持人史語所袁國華副研究員的帶領下，已再次規劃改版事宜，以期能達成人文爲本、科技爲用的目標。

早期的檢索系統是在 UNIX 作業系統下開發的，歷經多次修訂，目前重新使用 JAVA 程式開發系統。資料庫乃以保存原書的文字與排版爲基礎，由層級（hierarchical）的目錄來對應書本的章、節、段落等結構，讓使用者得以據其調閱正文，或訂定檢索的範圍。

爲因應資料庫改版，舊系統的資料需重新校對，標誌需要修改；同時新資料的電子化工作也必須持續進行。

正在建構中尚未完成之書籍約有兩億二百零二萬字。其中已完成校對的書籍有《宋人傳記資料索隱》等二十種，約五千五百三十九萬字，於 2005 年上線。校對中的書籍，有《明代律例彙編》等二十五種，約三千零六十四萬字。另外待校對書籍，有《文苑英華》等二十六種，約一億四千四百二十一萬字，正陸續建構中。

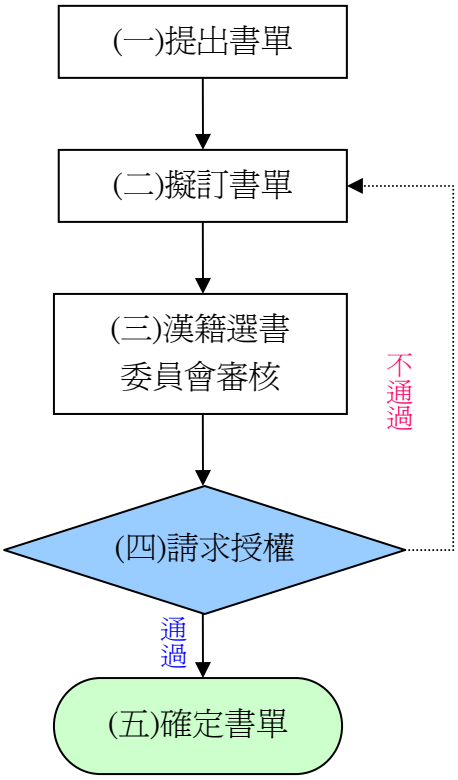
漢籍古文數位化工作流程圖



程序　　終端　　運行方向

數位化工作流程說明

一、選書



製作單位：中研院史語所漢籍工作室

（一）提出書單：

書目清單來源有二：

1. 史語所研究人員於選書委員會時提出書目。
2. 主持人根據漢籍全文資料庫完整性所提出的書目。

（二）擬訂書單：

整理後條列出書目清單（表 1）。

表 1、漢籍全文資料庫待輸入書單

	漢籍全文資料庫待輸入書單	
	書　　名	版本
1	七才子詩選	
2	九卿議定物料價值 四卷	(清)工部編，清乾隆元年(1736)刊本
3	二十五史外人物總傳要籍集成	董治安主編；濟南：齊魯書社，2000
4	入告初編 一卷，二編 一卷，三編 一卷	(清)張惟赤撰，清順治(1644-1661)末刊本
5	入幕須知 五種	(清)張廷驤輯，清光緒十八年(1892)浙江書局刊本
6	八旗人口冊	不著編人，清光緒間(1875-1908)排印本
7	十科策略 十卷	(清)劉文安著
8	三流道里表	(清)唐紹祖等纂修，清乾隆年間武英殿刊本
9	三流道里表	不著編人，清同治十一年(1872)江蘇書局重刊本
10	三流道里表 不分卷	(清)查克順等纂，清乾隆四十九年(1784)刊本
11	三省礦防考 二卷	(明)劉應元撰，明隆慶元年(1567)刊本
12	三朝聖諭錄 三卷	(明)楊士奇輯錄，明鈔本；漢籍資料庫有建置《國朝典故》
13	三賢政書 三種	(清)吳元炳輯，清光緒五年(1879)序刊本
14	上諭內閣 一百五十九卷	(清)允祿等輯，清刊本
15	上諭合律鄉約全書 一卷	(清)聖祖諭，(清)陳秉直解，清康熙間(1662-1722)刊本
16	于山奏牘 八卷	(清)于成龍著，清康熙年間(1662-1722)刊本
17	于清端公政書 八卷，外集一卷，續集一卷	(清)于成龍撰，(清)蔡方炳編次，清乾隆二十六年(1761)刊本
18	于肅愍公奏議 十卷	(明)于謙撰，明嘉靖二十年(1541)杭州重刊本
19	于肅愍公集	
20	大明九卿事例案例 不分卷	不著編人，明鈔本
21	大明令 不分卷	(明)太祖撰，鈔本
22	大明律 三十卷	胡瓊集解，胡效才增附，日本蓬左文庫藏明嘉靖刊本(本院圖書館查無此書)

（三）漢籍選書委員會審核：

交由漢籍選書委員會審核並排定建置順序。

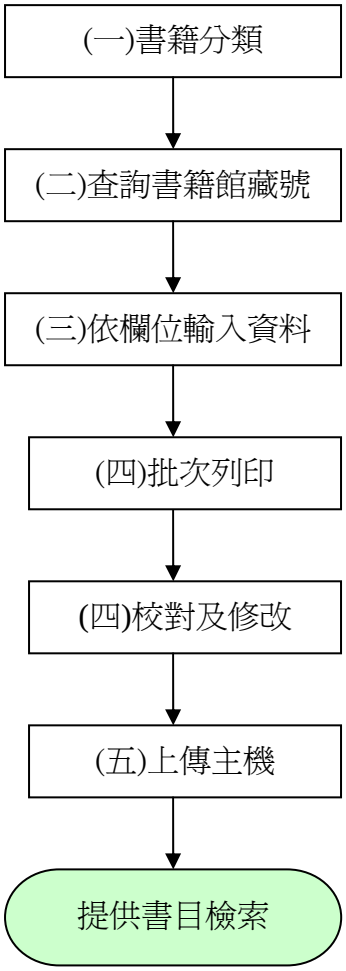
（四）請求授權：

請求著作權授權，如果無法取得授權必須重新擬訂書單，或者更換可以取得授權之版本。

（五）確定書單：

確定數位化書目清單。

二、書目建置



製作單位：中研院史語所漢籍工作室

（一）書籍分類：

依四庫圖書分類法之「經」、「史」、「子」、「集」，並增列「叢書」、「其他」二種，分門別類（圖一)。

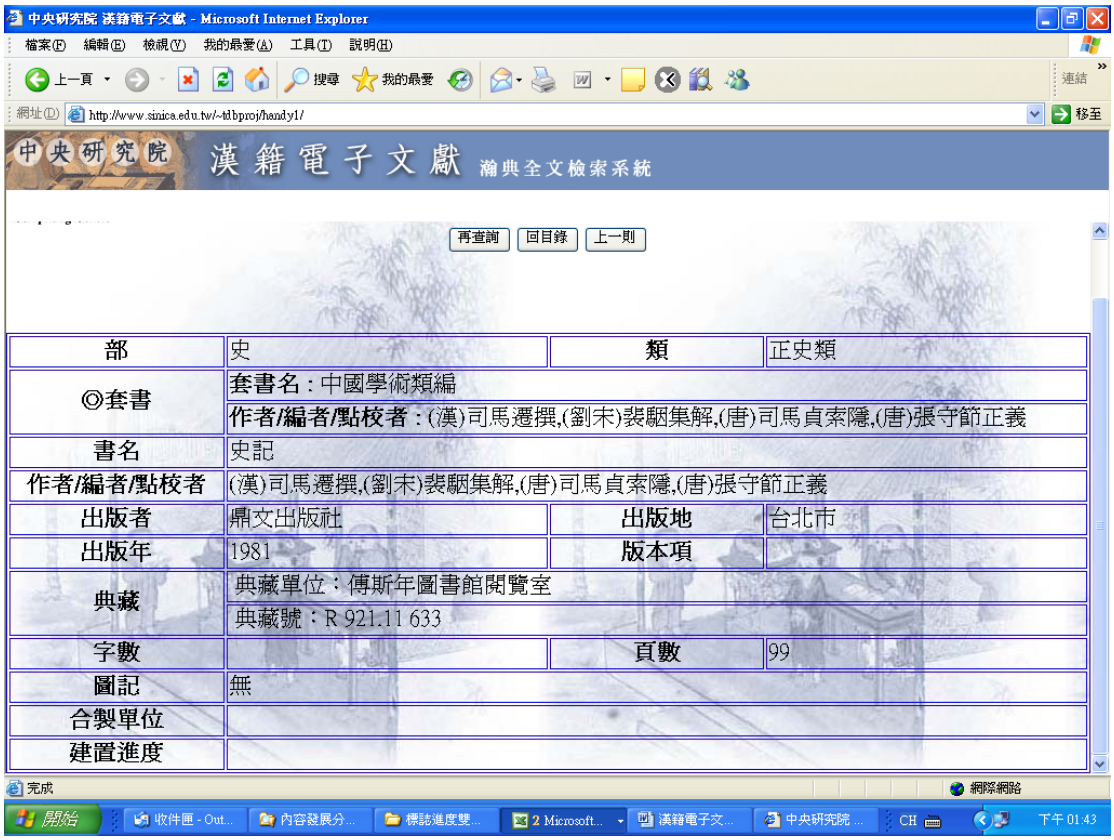
圖一、漢籍電子文獻資料庫目錄

（二）查詢書籍館藏號：

目前漢籍主要書籍爲中央研究院各圖書館館藏，其他爲史語所漢籍工作室購買，另一部份爲研究人員提供。可從中央研究院各圖書館查詢預定數位化之書籍的館藏號，提供讀者原書典藏處。

（三）依欄位輸入資料：

輸入類別、序號、書名、作者、出版者、出版地、出版年、典藏單位……等各項欄位（圖二）。



圖二、漢籍電子文獻資料庫書目欄位

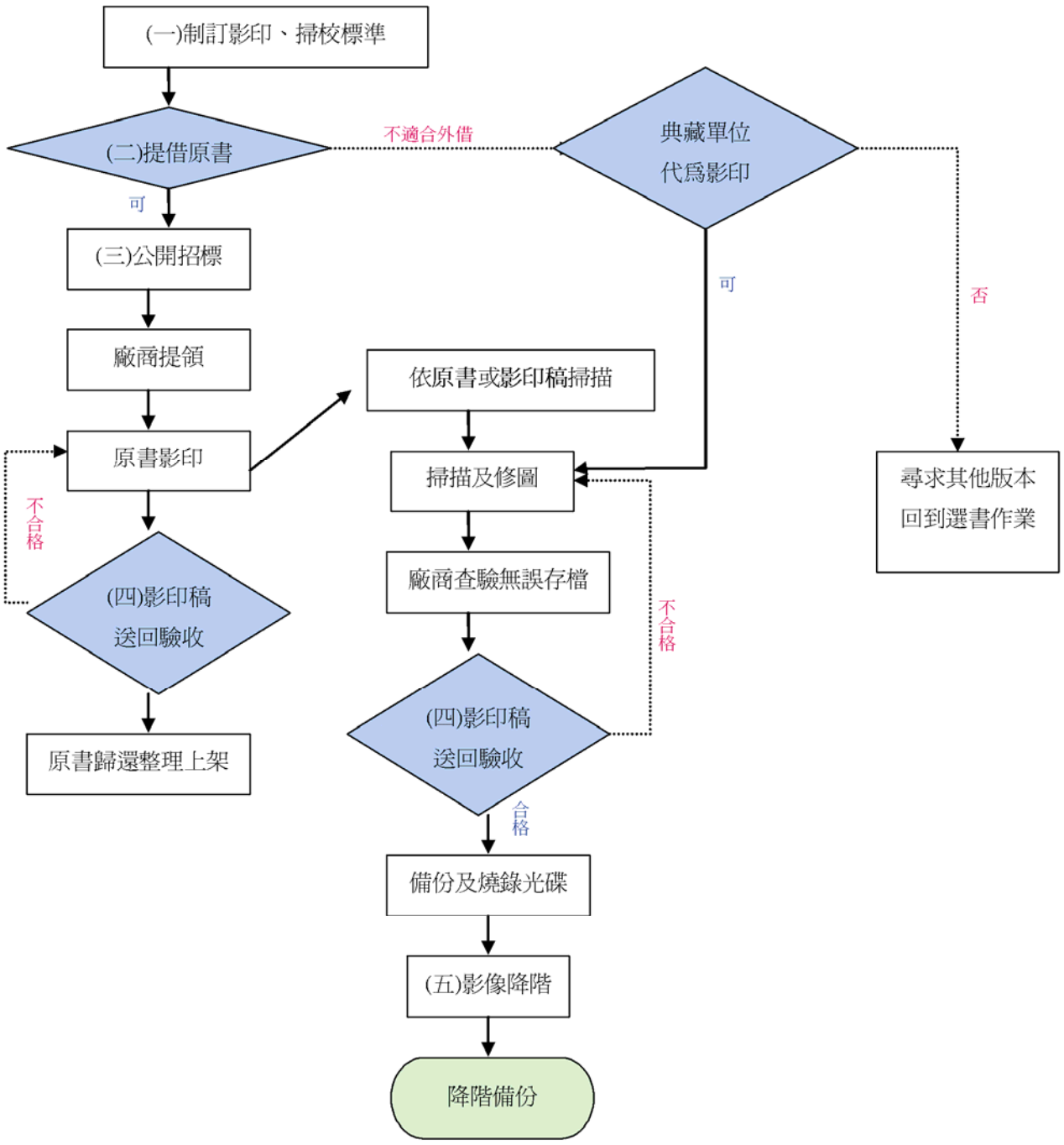
（四）批次列印、校對及修改：

每一年度書目資料輸入完成後必須列印樣稿，以便校對及修改。

（五）上傳主機：

核對資料無誤後，完成上線程序，以提供書目檢索。

三、影印及掃描（委外製作）



製作單位：中研院史語所漢籍工作室

（一）制訂影印、掃校標準：

1. 影印稿爲提供輸入廠商繕打之用，因此依照字體大小及清晰度決定比例大小。

2. 掃描圖檔則依照〈傅斯年圖書館全彩影像掃描及校驗相關作業標準〉，將原書 1：1 之比例，以頁爲單位，規格爲 300dpi、全彩（黑白）、TIFF 格式存檔。

（二）提借原書：

1. 根據書單，提調所選定之書籍以提供得標廠商影印及掃描。目前書籍的來源有三：一是漢籍提供經費購書，二是研究人員提供，三是院內圖書館館藏之書籍。

2. 如遇不得外借之書籍，依照各館藏單位之規定辦理仍不能處理者，提回至「選書」作業重新選擇版本。

（三）公開招標：

依照中央研究院招標規定施行。

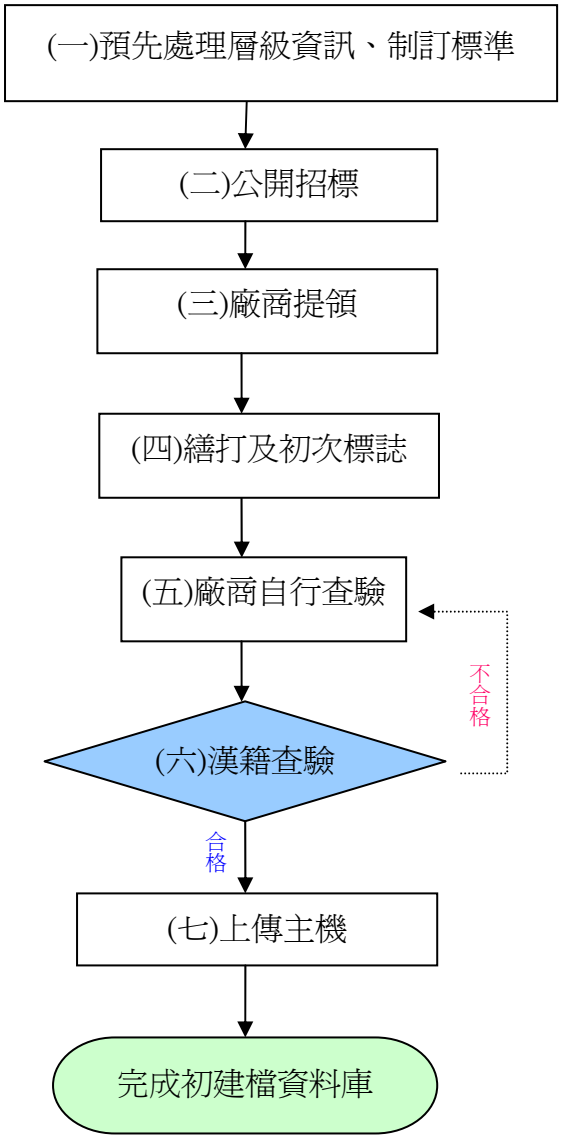
（四）驗收：

影印完成後廠商須將影印稿及原書送回本單位，驗收合格後即可。掃描完成後廠商須將檔案送回史語所漢籍工作室，驗收合格後進行影像降階工作。

（五）影像降階：

將圖檔依照〈傅斯年圖書館全彩影像掃描及校驗相關作業標準〉降階轉存，以利於後續作業，降階完成的檔案須另作備份。

四、繕打及初次標記（委外製作）5



製作單位：中研院史語所漢籍工作室

（一）預先處理層級資訊及制訂標準：

此爲發包委外工作，因此需依據各書籍體例及研究人員之要求，編寫檔案代碼及基本層級結構，針對各書列出輸入說明及規範、作業環境與限制等規定。

（二）公開招標：

依年度提撥經費，決定擬輸入字數並進行詢價，提出申請，經中央機關採購法規定，進行招標程序。

（三）廠商提領：

得標後，至工作室提領輸入文件，由漢籍工作室人員向廠商說明注意事項，提供加入 XML 語法標誌之程式及中央研究院之「缺字公用程式」，並示範操作，提供書面資料，讓廠商了解如何安裝與使用。

（四）繕打及初次標記：

1. 繕打：依據漢籍繕打原則（依書中原樣處理，內容中不清楚處，不作模糊的判斷與處理）進行文字輸入。

2. 初次標誌：使用 Keytext、EmEditor 等文書處理軟體做最初級的「層級」標誌。

（五）廠商自行查驗：

錯誤率達萬分之一以下方可送回。

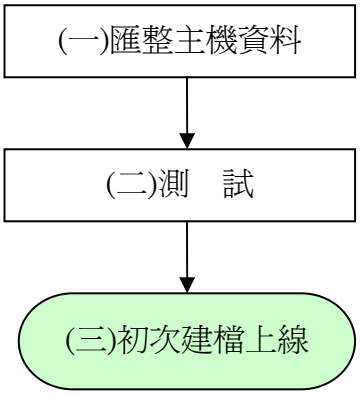
（六）漢籍查驗：

工作室同仁對送回檔案進行隨機取樣查驗，若達到合格要求(錯誤率達萬分之一以下)即上傳主機及燒製光碟進行備份，未達要求則退回廠商處進行修改，安排二次查驗。依採購法規定，所內或院方之驗收步驟，由相關單位派人至工作室對檔案進行查驗，合格即依規定處理，不合格則限期請廠商修改再重新進行驗收步驟。

（七）建置初校稿資料庫：

將驗收完成之檔案上傳，建置初校稿資料庫，供所內同仁使用。

五、初次建檔



製作單位：中研院史語所漢籍工作室

（一）匯整主機資料：

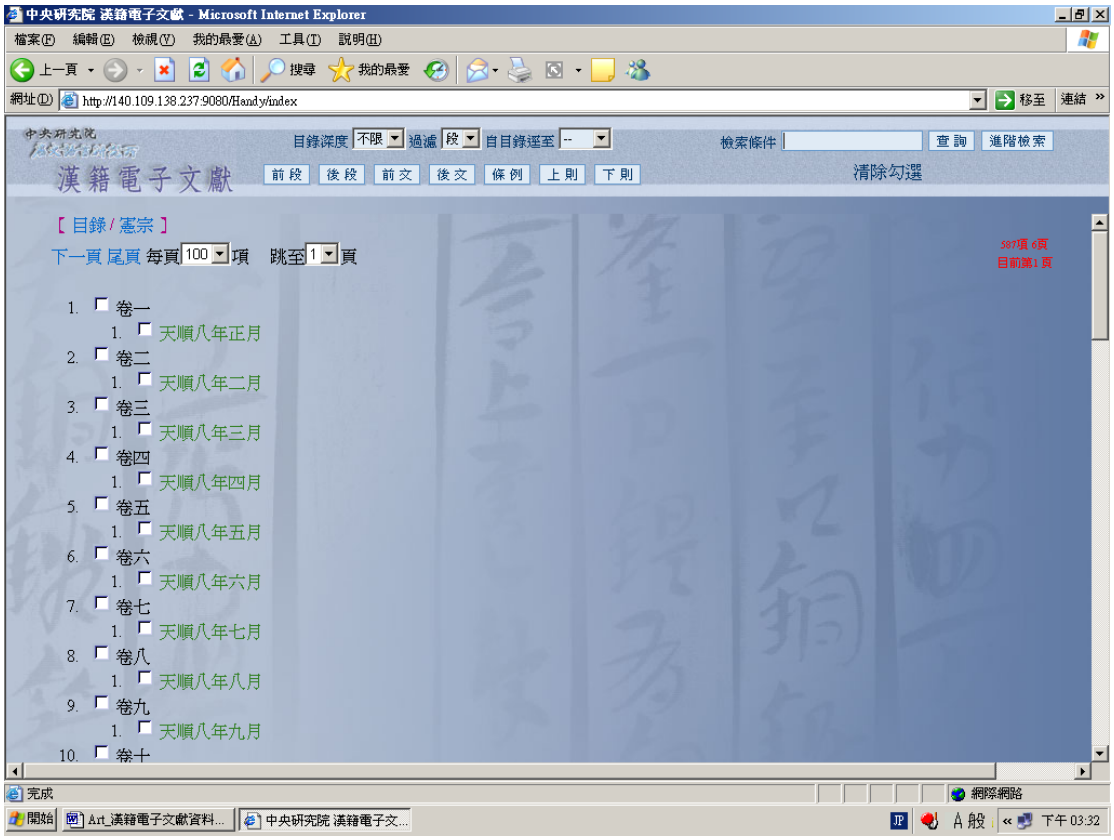
將繕打後的資料上傳主機，並且備份廠商送回的原始檔。

（二）測試：

進行資料庫上線測試。

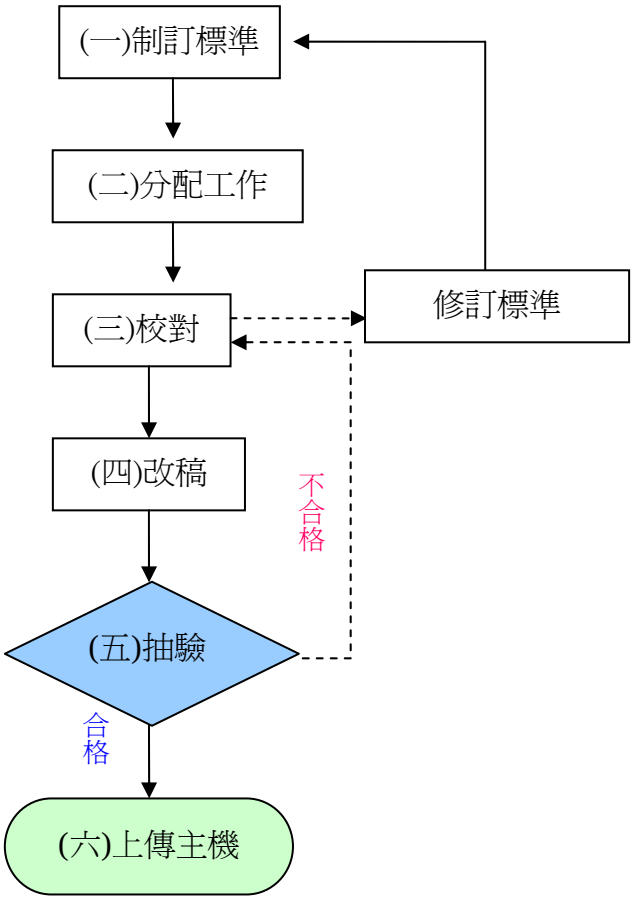
（三）初建檔資料庫上線：

測試無誤後初建檔完成，並開放資料庫（圖三）提供所內同仁使用。



圖三、漢籍電子文獻資料庫瀏覽畫面

六、一次及二次校對（委外製作）



製作單位：中研院史語所漢籍工作室

（一）制訂標準：

1. 制訂異體字、訛字、缺字、避諱字等挑字原則。
2. 制訂抽驗標準。

（二）分配工作：

1. 依總量平均分配。
2. 依難易分配給程度相當之工作人員。

（三）校對：

按照原書逐字校對，並針對每本書的差異之處將標準略加修改，以符合實際。

（四）改稿：

修正校對挑出之錯誤。

（五）抽驗：

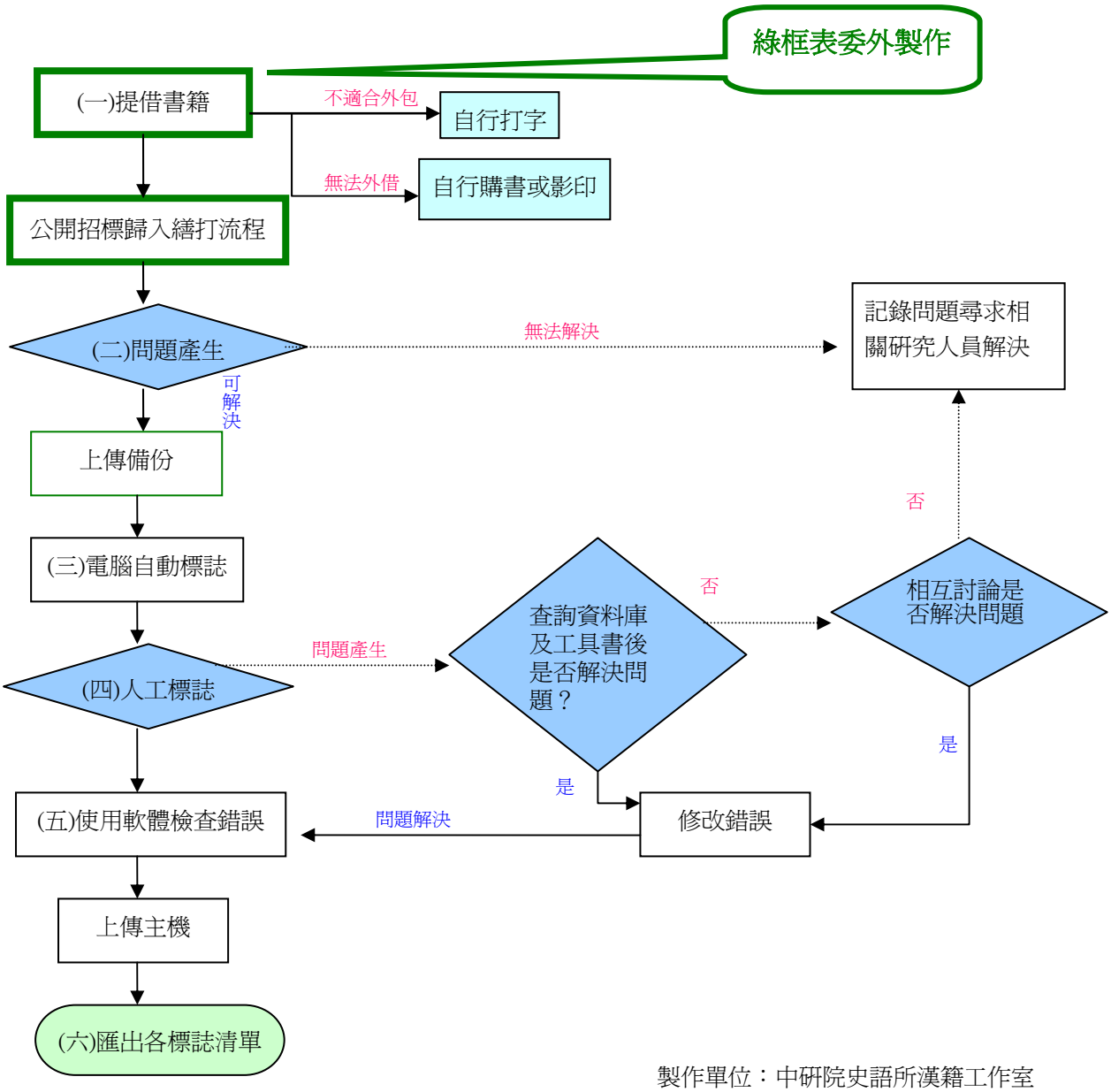
根據所制訂的標準，每本書抽出相同的字數，給非校對該書的工作人員抽驗。不合格者，組長詢問校對此書的人員之工作狀況，依情況列入績效考評，並且重新校對，如遇到書籍內容難度太高（如：以草書、行書繕寫，或字體不易判斷的手寫書稿等），必須尋找合適的校對人員、參考工具書或利用其他版本之書籍，重新校對。

（六）上傳主機：

完成的檔案上傳於主機備份。

七、詞類標誌

「詞類標記」工作分爲「詞類輸入」、「電腦斷詞」與「人工處理」等三個階段，除詞類輸入委外製作，另兩項皆由漢籍工作室執行。



製作單位：中研院史語所漢籍工作室

（一）提借書籍：

依資料庫所需，從中央研究院各圖書館提借相關詞類書籍與工具書。

（二）問題產生：

因繕打時必須對內容進行判斷，如遇到經由查詢工具書及相互討論還不可解

決之問題，需記錄並尋求研究人員之協助。

（三）電腦自動標誌：

先進行人名、地名等工具書或參考資料的繕打工作，再由資料庫進行電腦自動標誌。

（四）人工標誌：

檢查電腦無法處理的問題，目前先處理人名、地名、朝代、著作、職官、族名、年號七個部份。如遇有爭議的問題，先就資料庫查詢資料，無法解決再由小組相互討論;若尚有無法解決的問題,必須先記錄問題再尋求相關研究人員解決。

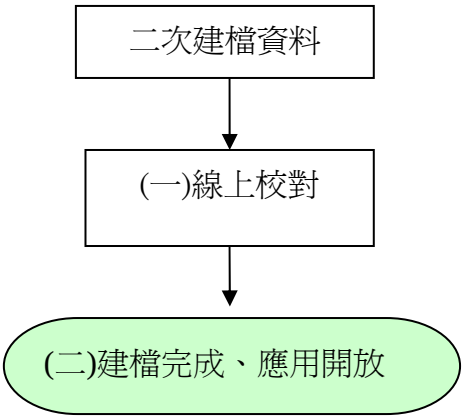
（五）使用軟體檢查錯誤：

使用標誌除錯軟體，可於加標誌的同時，檢查錯誤。

（六）匯出各標誌清單：

分別按人名、地名、朝代、著作、職官、族名、年號匯出名單以利後續處理。如：用《佩文韻府》標注《全宋詩》。

八、第二次建檔上線與校對



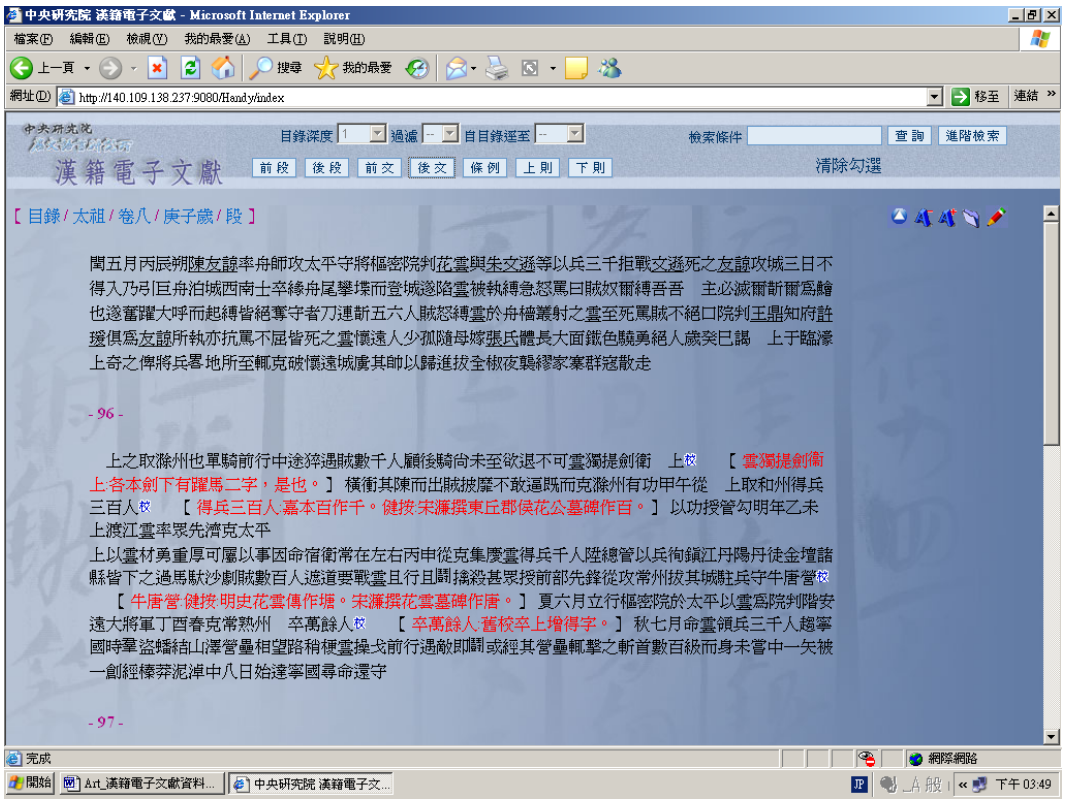
製作單位：中研院史語所漢籍工作室

（一）線上校對：

將二次建檔完成之檔案進行線上校對，以檢查及修正詞類標誌之錯誤。

（二）建檔完成：

開放資料庫給一般讀者使用（圖四）。



圖四、漢籍電子文獻資料庫畫面

※ 製作單位：數位典藏國家型科技計畫 內容發展分項計畫

中央研究院歷史語言研究所漢籍工作室

※ 文字撰寫：中央研究院歷史語言研究所漢籍工作室　助理李芳瑩

數位典藏國家型科技計畫 內容發展分項計畫

漢籍全文主題小組助理　謝筱琳

※ 圖片提供：中央研究院歷史語言研究所漢籍工作室　助理李芳瑩

※ 圖文編輯：數位典藏國家型科技計畫 內容發展分項計畫

漢籍全文主題小組助理　謝筱琳

※ 以上數位化工作流程，參考漢籍工作室於「2005 漢籍數位化合作建制研討會」發表之工作情形簡介。

致謝：

感謝中央研究院歷史語言研究所，漢籍工作室「漢籍電子文獻資料庫」之計畫主持人　袁國華老師、聯絡人李芳瑩小姐撥冗指導及提供實地拍攝與簡介編寫。並感謝漢籍電子文獻資料庫相關計畫人員之協助。