省諮議會公報議事錄數位化標準採用建議書

數位典藏國家型科技計畫 後設資料工作小組 2004/08/19

省諮議欲數位化其公報議事錄。後設資料工作小組根據省諮議會所提供的公報與議事錄原件，分析內容與格式，提供數位化的建議。本報告的目的，即在爲省諮議會提供系統建置的參考。

一、公報議事錄架構分析與研究

分析是針對省諮議會所要數位化的物件進行。分析的對象與素材如下：

- 台灣省議會公報－第六十九卷合訂本（上、下）二冊
- 台灣省議會議事錄－第八屆成立大會、第二次臨時大會、第七次大會，共三冊

採行的分析方法：

- 整體結構分析－依據文件目次與實際組成方式，分析文件組成的方式。
- 標準適用性分析－查看文件的屬性與內容，依整體的效率，判斷建議採用的標準。

（一）整體結構分析

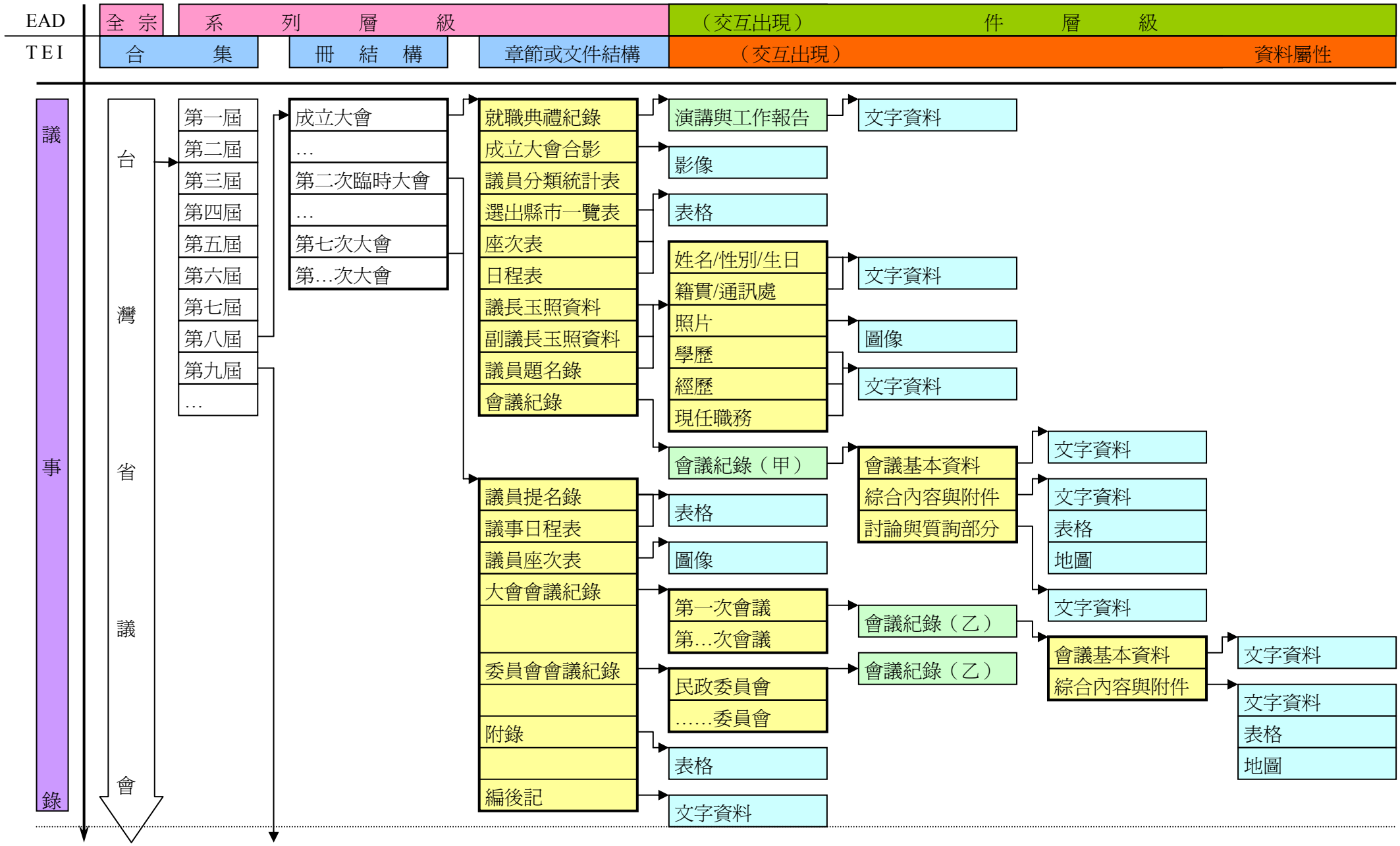
依照文件的結構的方式，歸納出以下兩個表：

- 議事錄結構表
- 公報結構表

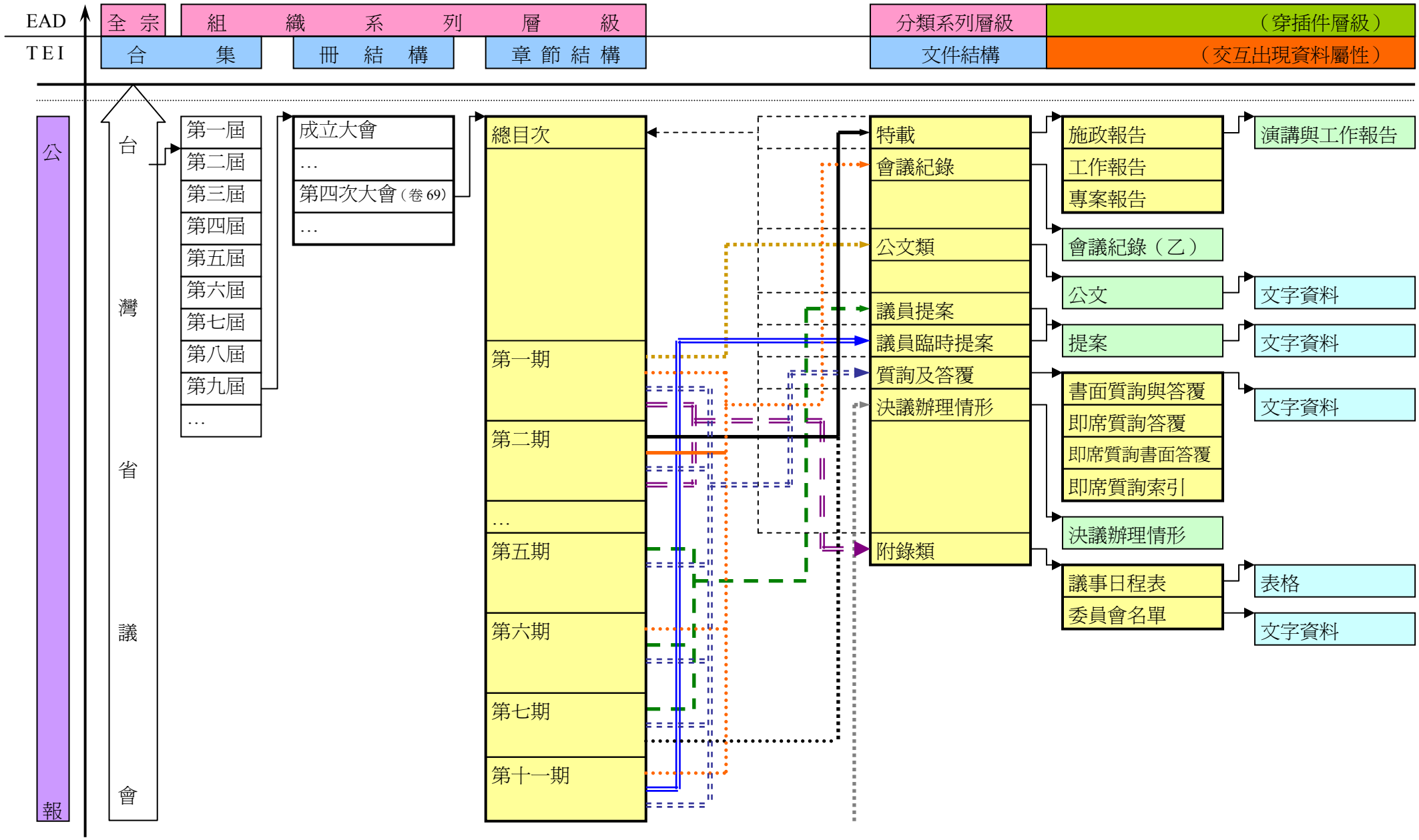
考量到公報與議事錄的性質不同，但可能要形成一個整合資料庫。先借用檔案的概念，再導入文件格式分析的概念。兩表呈現下列的分析結構爲：

- 所有的議事錄視爲一個「全宗」，公報視爲另一個「全宗」。
- 各屆、期與文件本身的目次結構，類比至宗卷等層級。
- 內文常採用的文件格式。
- 文件格式內的資料屬性。

表一：議事錄結構表


EAD
TEI
全宗
系列層級
（交互出現）
件層級
合集
冊結構
章節或文件結構
（交互出現）
資料屬性
議事錄
台灣省議會
第一屆
第二屆
第三屆
第四屆
第五屆
第六屆
第七屆
第八屆
第九屆
…
成立大會
…
第二次臨時大會
…
第七次大會
第…次大會
就職典禮紀錄
成立大會合影
議員分類統計表
選出縣市一覽表
座次表
日程表
議長玉照資料
副議長玉照資料
議員題名錄
會議紀錄
演講與工作報告
文字資料
影像
表格
姓名/性別/生日
籍貫/通訊處
照片
學歷
經歷
現任職務
文字資料
圖像
文字資料
會議紀錄（甲）
會議基本資料
綜合內容與附件
討論與質詢部分
文字資料
文字資料
表格
地圖
文字資料
議員提名錄
議事日程表
議員座次表
大會會議紀錄
委員會會議紀錄
附錄
編後記
表格
圖像
第一次會議
第…次會議
民政委員會
……委員會
表格
文字資料
會議紀錄（乙）
會議紀錄（乙）
會議基本資料
綜合內容與附件
文字資料
文字資料
表格
地圖

表二：公報結構表


EAD
TEI
全宗
合集
組織系列層級
冊結構
章節結構
分類系列層級
文件結構
（穿插件層級）
（交互出現資料屬性）
公報
台灣省議會
第一屆
第二屆
第三屆
第四屆
第五屆
第六屆
第七屆
第八屆
第九屆
…
成立大會
…
第四次大會（卷 69）
…
總目次
第一期
第二期
…
第五期
第六期
第七期
第十一期
特載
會議紀錄
公文類
議員提案
議員臨時提案
質詢及答覆
決議辦理情形
附錄類
施政報告
工作報告
專案報告
會議紀錄（乙）
公文
提案
書面質詢與答覆
即席質詢答覆
即席質詢書面答覆
即席質詢索引
決議辦理情形
議事日程表
委員會名單
演講與工作報告
文字資料
文字資料
文字資料
表格
文字資料

（二）標準適用性分析

1. EAD 適用性的考量

由上面的兩個表，可以看到，公報與議事錄的結構原則如下：

- 先有屆別，各屆又包含不同的大會。
- 以每次大會爲一個單位，將文件合訂成冊。
- 文件依發生的順序依次排列。
- 文件的內容安排有固定的邏輯順序。
- 由於文件屬性的不同，公報與議事錄又各自擁有不同的結構：

	公報	議事錄
製作方式	合訂	合刊
收錄文件	各期公報	各次大會會議紀錄 委員會會議紀錄
索引參照	索引彙編	合刊文件目錄

由上述各點觀之，公報與議事錄具有層級的概念，可以互檔案的概念互相參照。如此則可採用 EAD 來爲描述數位化的全文影像。

2. TEI 適用性的考量

以上的結構但若強調文件分析，則可以發現，公報與議事錄中的文件，不外乎下列數種固定的格式：

- 一般會議紀錄
- 含討論與即席質詢紀錄的會議紀錄（附即席質詢索引）
- 演說與工作報告
- 議員提案
- 公文
- 書面質詢與答覆
- 議事日程表
- 其他（委員會名單等）

而不論何種文件格式，內容不外乎是文字與圖像。表格本來應該算在文字的範圍，但是由於表格是一種特殊文字的格式。還是將他切分出來，成爲一類。故得三種基本類型：

- 文字：文件的主要表現方式，貫穿整體文件。

- 表格：格式化的文字內容，例如議事日程表、各種統計資料及概算表。
- 圖像：如地圖，以及無法規格化的文字，例如：座次表。

公報與議事錄的內容，主要是文字類型，而所包含的文件大都具有一定的規格。如果以「文件」的方式來處理，亦可採用 TEI 做爲全文標誌的標準。

3. EAD 串連 TEI 的考量

使用 EAD 與 TEI 作分析，分別得到不同的結構。但若將兩個結構排比在一起，以整體觀之，則可發現下列可以相互對應的情況：

	EAD	TEI
數位化物件總體	全宗	合集
組織模式	系列層級	章節與文件結構

- 以 EAD 標誌的觀點觀之，不論其所切分的方式爲何，當分析到最小的層級（即件層級）時，文件不再分割。
- 以 TEI 的標誌觀點觀之，任何文件不論其層級爲何，最後都是資料屬性，即文字、表格、及圖像的結合。

EAD 與 TEI 之間，有異有同，應可兼取，截長補短：

- 以 EAD 擅長處理的層級觀念，標誌「全宗／合集」與「系列層級／章節與文件結構」的部分。
- 以 TEI 專長的電子全文，表現「件／文字型資料屬性」。

系統的開發雖然需要採行標準，但是，也並未限制只能採單一的方案。若能兼取其長，可以表現更完全的整體架構與文件特性。兩者的結構可以並行，可以採分段或是平行建置的方式進行。不同的工具可提供不同的服務，使用者也可以就其需求而多一種使用選擇。

二、範例研討

（一）EAD 範例研討

1. 總督府公文類纂（https://sotokufu.sinica.edu.tw/sotokufu/query.php）

- 資料來源／國史館台灣文獻館；系統開發：中研院計算中心。
- 提供欄位查詢，檢索結果爲條目式的資料庫記錄。
- 若有影像檔，可連結到文件的影像。

2. Bergmann, Wilhelm. Diaries and Transcripts, 1969-1972.
（http://roger.ucsd.edu/search/t?ucsd+mss+272）

- 加州大學聖地牙哥分校圖書館的數位典藏。
- 提供 EAD、HTML、Catalog Record 三種瀏覽模式。
- EAD 爲 SGML 檔，需要另裝 Reader，但檔案可下載。
- HTML 與 Catalog Record 可線上閱讀。
- 可看到檔案層級，與各層級中的實體文件的歸檔方式。
- 完全沒有提供影像以供連結。
- 檔案最重要的是全宗的概念，全部的檔案視爲一個整體。雖然在描述數位的物件，但是，未數位化者仍可描述。

3. University of the Pacific
（http://findaid.oac.cdlib.org/findaid/ark:/13030/tf496nb393）

- 加州線上檔案（Online Archive of California）的數位藏品。
- 不同於 UCSD 的界面設計。
- 未提供 EAD 原檔，已經轉成 HTML 格式。
- 影像部分獨立出來，可直接選取。
- 打開 Container List，可以看到檔案收整整理的方式。
- 選擇 Entire EAD，可以看到件層次以上的所有描述。
- 有數位影像者可以連結。提供四種選擇：影像與描述資料、預覽小圖、中解析度圖、及高品質大圖。

4. American President Lines Records
（http://findaid.oac.cdlib.org/findaid/ark:/13030/tf4j49n761）

- 加州線上檔案（Online Archive of California）的數位藏品。
- 加州線上檔案的系統已整合。大致的功能與「University of the Pacific」部分相同，不再贅述。
- 打開 Collection and Series Description，可以看到檔案收整整理的方式。
- 全宗的紀錄很長，顯示 EAD 使用的尺度具有很大的彈性。

（二）TEI 範例研討

1. 中研院漢籍全文資料庫（http://www.sinica.edu.tw/ftms-bin/ftmsw3）

- 爲台灣第一套全文資料庫，內容包含整部二十五史、整部阮刻十三經、超過兩千萬字的臺灣史料、一千萬字的大正藏以及其他典籍，合計字數一億三千四百萬字並以每年至少一千萬字的速率持續成長。

- 擁有驚人的檢索效率，大幅節省研究人員皓首窮經的時間。
- 使用的系統爲中研院計算中心開發，文件的標誌規格爲中研院內部自行研發。
- 原先只提供 telnet UNIX 的檢索，目前已提供 WWW 網頁界面功能。

2. 台灣網路上的公報系統

線上公報的範例

爲了進一步了解公報議事錄系統所應具備的功能，到網路上找尋其他公報系統以爲參考。除了需要加入會員才能查詢者（專利公報），目前線上可供一般民眾查詢與瀏覽的公報系統有下列八個系統：教育部公報系統、國家圖書館政府公報電子全文、國家圖書館政府公報全文影象查詢系統、多公報系統、立法院公報、立法院公報影像、立法院議事系統、立法院質詢系統。

經過初步的使用測試評估，將這些系統的基本功能做了分析，結果如表四所示。並歸納下列三項系統設計時應的參考的功能：

目錄瀏覽

- 依原文件的卷期數表列，綱目清析，可做爲檔案分層的依據
- 必須層層翻閱，如不知道確實的卷期數，浪費時間精力。

欄位查詢

- 欄位輸入的內容越多，可提到檢索結果的完整性。
- 立法院所使用的欄位非常的多，製作上需要較高的內容分析專業需求。
- 由於公報資料所牽涉的內容非常廣泛，若欄位輸入的內容僅限於某些基本的描述，過於精簡，則檢索的效率會大打折扣

全文查詢

- 可以檢索到欄位查詢所檢索不到的資料。
- 若檢索的詞彙太過平凡，可能查到太多資料，影響精確性。

表四：線上公報系統的功能比較

單位	系統	電子全文	目錄瀏覽	查詢功能		影像	備註
				全文	欄位		
教育部	公報系統 http://www.edu.tw/EDU_WEB/Web/publicFun/Query_Paper.php?UNIT_NAME=教育部公報	◎	◎				
國家圖書館[1]	政府公報電子全文[2] http://readopac.ncl.edu.tw/cgi-bin/ncl10/m_ncl10	◎	◎	◎			
	政府公報全文影象查詢系統 http://readopac.ncl.edu.tw/cgi/gaz/readncl/gaz_login.cgi		◎		◎	◎	每頁 2 或 6 元[3]
跨單位[4]	多公報系統 http://www.gazettes.com.tw/	◎	◎	◎		◎	
立法院	立法院公報[5] http://lci.ly.gov.tw/	◎	◎	◎		◎	
	立法院公報影像 http://npl.ly.gov.tw/www/home.jsp?page_url=library/legislative_gazettes.jsp		◎			◎	
	立法院議事系統 http://lis.ly.gov.tw/ttscgi/ttsweb?@0:0:1:/disk1/lg/lgmeet:回首頁://npl.ly.gov.tw@@0.5156348785478084				◎	◎	
	立法院質詢系統 http://lis.ly.gov.tw/ttscgi/ttsweb?@0:0:1:/disk1/textbase/qr:回首頁://npl.ly.gov.tw@@0.5746262536493891				◎	◎	

[1] 國家圖書館的兩個系統各立獨立，未整合，全文部分的查詢結果無法直接連到影像檔，必須重新查詢。
[2] 全文由各政府單位線上轉入。
[3] 依使用系統的不同，收取的費用不一樣。必須使用讀圖器。
[4] 可查詢七種公報：行政院、新聞局、財政部、臺北市、高雄市、臺南縣、高雄縣。
[5] 全文查詢僅限於 89 卷 50 期以後的資料，之前仍僅提供目錄查閱。

3. 大正藏（http://ccbs.ntu.edu.tw/cbeta/result/index.htm）

- 採用 TEI 來標誌的全文資料庫。
- 運用超文件的功能，將正文與注釋做連結。
- 利用注釋中的校勘資料，可還原不同時代的大藏經版本（四十幾種）。
- 利用標誌資料，可爲引用做複製引用的來源註解[6]。
- 由於瀏覽器的限制，一般使用者使用的網路版仍需轉換成 HTML 格式。
- 單機版有讀經器，直接處理 XML 的文件，並提供書籤功能。
- 各種功能可做爲 TEI 標誌全文資料庫的參考。

4. Mark Twain: *How the Chimney-Sweep Got the Ear of the Emperor*. A Critical Edition with Facsimile （http://etext.virginia.edu/users/tousignant/chimney/）

- 維吉尼亞大學（University of Virginia Library）的電子文件中心（Electronic Texts Center）的數位典藏。
- 提供兩種閱讀模式。一種只有電子全文，另一種是電子全文超連結至不同版本比較。
- 主要的內容是電子全文，再輔以四個不同版本的影像，包含馬克吐溫手蹟、打字機、及兩個不同版的印刷品。
- 無全頁影像，圖檔都剪成小塊，在上頁框點選相關段落時，四個小圖檔並排出現於下頁框中。

（三）EAD 連結 TEI

1. The Electronic Text Center's Liberian Letters（http://etext.virginia.edu/subjects/liberia/）

- 維吉尼亞大學（University of Virginia Library）的電子文件中心（Electronic Texts Center）的數位典藏。館藏兩種不同的檔案「Samson Ceasar's letters to David S. Haselden and Henry F. Westfall, 1834-1835」與「Letters from the former slaves of Terrell, 1857-1866.」
- 整份檔案已由 EAD 描述其整體結構。
- 件層次的文件，各有一至四頁不等的信件。提供影像檔與電子全文。
- 點選影像的頁數，會出現該頁的全文影像。
- 點選電字全文的部分，會出現 TEI 的文件結構。再選到正文的部分，會顯示所有的文件內容及相對應的頁面的縮圖。點縮圖也可以打開大圖的全文影像。

[6] 選取欲複製經文，貼入文件，可自動註記爲『《妙法蓮華經》「如是我聞…」(CBETA, T09, no. 262, p. 1, c19-20) 』顯示該經文出自何經，在大正藏的卷、冊、頁、欄數。

三、EAD 與 TEI 的結構、數位化流程比較

（一）EAD（Encoded Archival Description[7]）簡介

- EAD 是由美國檔案學會檔案描述編碼格式工作小組所研發。
- 主要應用在強調全宗與層級觀念檔案類型。
- 標誌的目的是爲數位化影像做「內容描述」。
- 爲數位典藏家型科技計畫推薦爲檔案類數位化標準。

（二）TEI（Text Encoding Initiative[8]）簡介

- 1987 年創位，由位於英國的 TEI Consortium 所研發，這個標準之前是由英國的計算與人文學會、計算語言學會、及文學與語言學計算學會等組織所贊助。目前的贊助單位已延申是美國人文研究基金與歐盟執行計畫等非英國家。
- 可應用在任何數位電子文件。
- 標誌的目的是爲電子文件加入關於格式架構與內容的註解。
- 爲數位典藏家型科技計畫推薦爲檔案類數位化標準。
- TEI 歷史優久，目前國內採用的單位有中華電子佛典的「大正藏」以及元智大學的「網路展書讀」計畫[9]。其他國外單位運用的類型有：文學、手稿、歷史文獻、語文資料。

兩者各有優劣，以下根據其適用特性、元素與文件結構、數位化流程兩項，表列如表三，以供參考。

[7] http://www.loc.gov/ead/ead.html
[8] http://www.tei-c.org/
[9] http://cls.admin.yzu.edu.tw/HOME.HTM

表三：EAD 與 TEI 標準的各種比較

標準	EAD	TEI	EAD 連結 TEI（含全文影像）	EAD 連結 TEI（不含全文影像）
資料庫類型	檔案影像資料庫	電子全文資料庫	檔案影像資料庫與 電子全文資料庫 兩者相互連結	檔案結構資料庫 電子全文資料庫 兩者相互連結
資料庫格式	欄位資料庫 影像資料庫	文字資料庫	欄位資料庫 影像資料庫 文字資料庫	欄位資料庫 文字資料庫
適用資料類型	圖像資料 文字資料	文字資料（圖像資料不能直接處理，但可以描述其內容，以連結的方式連到相關的影像。）	圖像資料 文字資料	文字資料
資料庫大小	欄位資料庫，空間較小 圖像資料庫，空間最大	文字檔，空間較小	欄位資料庫 圖像資料庫 全文資料庫 佔用空間最大；	欄位資料庫 全文資料庫 空間次小；
數位化方式	攝影或掃描	打字或 OCR	攝影或掃描 打字或 OCR	打字或 OCR
檔案驗證	顏色校正（可略）	文件校對（詳附件一）	顏色校正 文件挍對	文件校對
資料規格化	定義層級 定義資料欄位 影像內容分析	文件打字規則 文件格式分析（詳附件一）	定義層級 定義資料欄位 影像內容分析	定義層級 定義資料欄位 文件打字規則

			文件打字規則 文件格式分析	文件格式分析
內容建置/ 標誌	抽取影像內容輸入相關欄位	全冊文件架構標誌 文件內容標誌 標誌測試	抽取影像內容輸入相關欄位 文件架構標誌 文件內容標誌 標誌測試	相關架欄輸入欄位 單一文件架構標誌 文件內容標誌 標誌測試
專業需求	熟悉整體的結構 判斷影像中應擷取的資料內容	熟悉標誌 熟悉整體的結構 深入了解內容，判斷應標誌內容	熟悉整體的結構 判斷影像中應擷取的資料內容 熟悉標誌 深入了解內容，判斷應標誌內容	熟悉整體的結構 熟悉標誌 深入了解內容，判斷應標誌內容
系統開發	資料庫系統	各種工具程式： 檔案比對程式 看圖校字程式 缺字系統 標誌器 XML 標誌檢查器 XML 閱讀器 XML HTML 轉換器	資料庫系統 影像系統（含管理及秀圖系統） 各種工具程式： 檔案比對程式 看圖校字程式 缺字系統 標誌器 XML 標誌檢查器 XML 閱讀器 XML HTML 轉換器	資料庫系統 各種工具程式： 檔案比對程式 看圖校字程式 缺字系統 標誌器 XML 標誌檢查器 XML 閱讀器 XML HTML 轉換器
檢索方式	資料庫欄位比對	全文比對	資料庫欄位比對 全文比對	資料庫欄位比對 全文比對
檢索效率	精確度高	回溯率高	使用者可依需求選擇適合的工具	使用者可依需求選擇適合的工具

四、結論與建議

根據公報議事錄的類型與內容來分析，公報議事錄資料庫的建置，有多種可能性可以考量。我們建議：先考慮系統的全貌與功能，決定資料庫的類型，再進行系統的內容建置與功能開發。建議應就下列數點，進行更深度的考量：

（一）EAD（全文影像資料庫）

- 架構類似於檔案的資料庫，較易上手。
- 系統架構與檔案的雷同，將來若要整合爲同一系統，可能會比較容易。
- 容許文件本身結構之外的分類架構。
- 加強控制檢索詞彙欄位的重要性，多輸入內容以提昇回溯率。
- 應建立人名權威權威檔、機構權威檔、主題詞彙表等，以提高檢索效率。
- 系統委託中心開發，目前已有三個系統，可以做爲成果的參考：
  中研院近史所近代外交經濟檔案：http://dipeco.sinica.edu.tw/
  國史館：http://dftt.drnh.gov.tw:8080/DAP/chinese/Database/Database.jsp
  台灣文獻館：總督府公文類纂：http://db.th.gov.tw/~textdb/test/sotokufu/

（一）TEI（電子全文資料庫）

- 公報與議事錄的資料類型，絕大多數爲文字資料，可考慮。
- 以檢索的功能而言，電子全文在檢索內容的表現較高，如能全文電子化，應可提供更好的服務。
- 全文資料庫標誌需要較專業的訓練，較不易上手。
- 需要另行安排訓練課程，上一些標誌相關的課程。
- 剛開始對標誌不是很熟悉的期間，進度可能會很緩慢。
- 目前尚無其他國家型計畫採用。

（三）EAD 連結 TEI（融合影像資料庫與電子全文資料庫）

- 可兼採全文資料庫回溯率高與欄位資料庫精確率高的長處。
- 可提供原件的影像。
- EAD 欄位輸入部分，系統架構與檔案的雷同，將來若要整合爲同一系統，可能會比較容易。
- 除了掃描影像，還需要全文標誌，需要最多的經費與人力。
- 或許可以規畫成機構內部的長期計畫。
- EAD 與 TEI 可以分開進行，各部分可以分開串連。
- 建議初期執行時，先進行影像資料庫的部分。待一至三次大會的影像資料庫建置完成後，再開行進行一至三次大會的電子全文文件的標誌。
- 建置影像資料庫的期間，可以將全文拿去送打或是 OCR，由於有比較充裕的時

間準備電子全文檔案。可以進行詳細的成本評估（檔案內容的正確性與校對的時間相關，不可不慎），選擇費用較低品質較高的方式進行。或是兩者都採用，再進行文件的比對。

- 目前尚無其他國家型計畫採用。

（三）EAD 連結 TEI（檔案件層級連結電子全文）

- 可兼採全文資料庫回溯率高與欄位資料庫精確率高的長處。
- EAD 欄位輸入部分，系統架構與檔案的雷同，將來若要整合爲同一系統，可能會比較容易。
- 不需掃描影像，但是仍要負擔欄位輸入與標誌的工作。
- 或許可以規畫成機構內部的長期計畫。
- EAD 與 TEI 可以分開進行，各部分可以分開串連。
- 建議初期執行時，先進行欄位輸入部分。待欄位輸入進度超過一半後，再開行進行電子全文文件的標誌。
- 欄位資料輸入期間，可以將全文拿去送打或是 OCR，由於有比較充裕的時間準備電子全文檔案。可以進行詳細的成本評估（檔案內容的正確性與校對的時間相關，不可不慎），選擇費用較低品質較高的方式進行。或是兩者都採用，再進行文件的比對。
- 目前尚無其他國家型計畫採用。

參考資料

1. 檔案編碼描述格式標準元素辭典（EAD Tag Library），中文版（稿）：
   http://www.sinica.edu.tw/~metadata/ead/ead_titlepage_3.htm
2. 文件編碼組織 後設資料標誌集 選錄版（TEI Lite），中文版（稿）：
   http://www.sinica.edu.tw/~metadata/standard/rarebook/TEI921224/index_c.htm
3. 「中央研究院漢籍電子文獻」相關訊息：
   http://www.sinica.edu.tw/~tdbproj/handy/ftmsw3.html
4. 謝清俊、林晰，中央研究院古籍全文資料庫的發展概要，1997 年 3 月：
   http://www.sinica.edu.tw/~tdbproj/handy/thesis.html
5. 中華電子佛典協會，CBETA 電子佛典集成，光碟版，2004 年 4 月。
6. 中華電子佛典協會，CBETA 電子佛典集成，網路版：http://ccbs.ntu.edu.tw/cbeta/
7. 「國家圖書館政府公報電子全文」簡介：
   http://readopac.ncl.edu.tw/cgi/ncl10/ncl10info?312c77566a7a4a537379
8. 「國家圖書館政府公報全文影像查詢系統」系統簡介：
   http://readopac.ncl.edu.tw/cgi/gaz/readncl/gaz_login.cgi

附件一：

省諮議會公報議事錄全文資料庫的數位化流程

後設資料工作小組 2004/08/09

電子全文資料庫的資料建置，包含取得全文的電子文字檔、文件校對、內容標誌與標誌測試三個程序。

取得全文的電子文字檔

早期都用使用打字的方式，由於公報議事錄的內容，大多都是印刷字體，可以考慮採用 OCR 的方式，以節省時間與人力成本。全文輸入前，必須先研究文件的內容，詳究其格式，依照小組對所提供文件的初步研究，並參酌前人數位化的經驗，建議需考慮的面向如下：

- 資料分割的單位與檔案命名方式
  就所欲數位化的單位來分析文件，採行系統性的檔案命名方式。利用冊、卷、期、章、節等特性，爲文件與各章節命名。
- 規定標點符號與其他符號的打字方式。例如：
  國字的「零」：應使用全形的「０」，而非符號「○」
  標點符號：一律使用全形、橫式括弧與引號
  分數：¼打成 1/4
  帶數字符號：①打成（1）
- 是否依原文件的版面打字
  空行認定的方式爲何，何種狀況應留空行
  段落首行與縮排部分是否依樣空白，採用全形或半形空白
  條列式文件不可使用標號的功能，必須逐字繕打。
- 加入頁碼標誌
  每頁起始處，必須輸入頁碼。
- 圖象處理方式
  圖說文字是可標誌的部分，仍需以輸入在電子文件中。
  爲建立正確的連結，內文所含圖象必須採系統性的命名方式。
- 表格處理方式
  表格欲以文字或圖象的方式處理？若兩者皆要，則應訂定標準。
- 缺字處理方式
  使用何種缺字解決方案？（建議採用漢字構形資料庫，中研院技術，免費，可使用五萬多字形）
- 修訂部分處理方式

若原件經過貼字修改內容，是否處理？處理時應採取何種標記。

文件校對

中研院漢籍全文資料庫的標準號稱「三讀五校」。由各種經驗得知，校對工作曠時費日，而且高品質的校對人材可遇不可求。建議採行以下的流程：

- 程式比對一校：
  目前漢籍與中華電子佛典均採用此法。運用原理爲：分別打字，不同的人員犯同一個錯誤的可能性微乎其微。能速過濾出大部分的錯誤。如果採用 OCR 的方式來進行數位化的工作，則若不同的單位仍採用同一個辨識程式，誤辨的地方可能會相同。如欲採用程式一校，則應將採用不同的辨識軟體來進行。比對程式可以由本小組與計算中心溝通協助取得。
- 人工二校：
  一校完畢再以人工仔細校對一次。
- 如有需要可以進行三校。
- 缺字處理與記錄：
  建議採用漢字構形資料庫直接輸入，無法表達者記錄後回報相關單位處理。

內容標誌與測試

內容標誌可以分爲兩個部分：架構的標誌，內文的標誌。兩者都需要依據實際文本的研究，找出標誌的通則與規範，並選擇合適的標準進行標誌。標誌工作的要點如下：

- 標誌是「人」的工作，即便使用程式工具輔助，過程仍需要人的操作以掌握標誌的品質。
- 標誌是需要經驗的專業工作。必須同時了解文件的內容特性，並熟悉的選擇使用各種標誌。
- 爲了加速標誌工作的進行，系統開發者可能需要依標誌單位的需求，開發工具，至少必須有標誌程式與測試程式（網路上有很多免費的 Parser 可以使用）
- 標誌完成的文件必須通測試，驗證標誌有無語法錯誤。測試完成的文件，即可加入資料庫。

爲使工作順利進行，最好是使用電腦直接在文件上標誌，標誌完了可以立即測試標誌的正確性。建議採用的程順如下：

- 分析文件，了解整體的架構，提出架構標誌集及 DTD。
- 先進行架標的標誌與測試，建構資料庫的雛型。
- 分析文件，了解內文所包含的各項資料。再進一步分析出需要進一步標誌的內容部分，例如：人名、地名等詞彙的標誌。

- 就已完成的雛型資料庫，進行第二次標誌。

雖然要同時標誌架構與內文也是可行的策略，但採取二段式工作步驟的原因如下：

- 架構的標誌不需要檢視大量的內容，可以加速度文件的標誌速度，可以在短時間內取得較多的的已標誌文件，先構成一個資料庫。
- 新手在剛開始使用標誌時，通常需要一段適應期，標誌的速度才會變快，如果一開始就進行內容標誌，可能會影響進度。
- 內容標誌雖然可以加強檢索效率，但並非必需。如果經費及人力有限，可以分期慢慢完成。