附件一

行政院國家科學委員會補助專題研究計畫 ■ 成果報告 □期中進度報告

（計畫名稱）

臺灣大學典藏數位化計畫-

子計畫七：數位典藏資訊技術研發計畫

計畫類別：□ 個別型計畫　■ 整合型計畫

計畫編號：NSC91－2422－H－002－525

執行期間：91 年 1 月 1 日至 91 年 12 月 31 日

計畫主持人：林一鵬教授

共同主持人：陳銘憲教授

計畫參與人員：雲晴煌、鄧維光、歐建志

成果報告類型(依經費核定清單規定繳交)：□精簡報告　■完整報告

本成果報告包括以下應繳交之附件：

□赴國外出差或研習心得報告一份

□赴大陸地區出差或研習心得報告一份

□出席國際學術會議心得報告及發表之論文各一份

□國際合作研究計畫國外研究報告書一份

處理方式：除產學合作研究計畫、提升產業技術及人才培育研究計畫、列管計畫及下列情形者外，得立即公開查詢

□涉及專利或其他智慧財產權，□一年□二年後可公開查詢

執行單位：國立台灣大學計算機及資訊網路中心

中　華　民　國　九十二 年　四　月　九　日

壹、 計畫概述

1. 計畫時程：

全程執行期限：自民國 91 年 1 月 1 日起至民國 95 年 12 月 31 日止

第一年執行時程：自民國 91 年 1 月 1 日起至 91 年 12 月 31 日止

2. 參與人員：姓名、職稱、任務

類　　別	姓　名	服務機構/系所	職稱	在本研究計畫內擔任之具體工作性質、項目及範圍
主持人	林一鵬	台大資訊工程系	教授	整體計畫之規劃與研究
共同主持人	陳銘憲	台大電機工程系	教授	協助整體計畫之規劃與研究
博士班研究生兼任助理	雲晴煌	台大電機工程系	兼任助理	協助數位典藏資訊技術研究與開發
博士班研究生兼任助理	鄧維光	台大電機工程系	兼任助理	協助數位典藏資料庫與環境之建置
博士班研究生兼任助理	歐建志	台大電機工程系	兼任助理	協助數位典藏資訊儲存放置之技術研究與開發

3. 計畫目標：

本子計畫的主要目標是研發適宜整合多媒體型態資料的數位化技術、協助開發其他子類別對其典藏網路化時所需求的相關電腦技術，並建構以 Web 為基礎的平台以利於資料的交換與展示，用以提供此總計畫所涵蓋之子計畫其執行典藏數位化之所需，希望藉此計畫之執行，使珍貴的典藏資料能在電腦新科技的應用下，達到保存與供大眾使用之雙重功能。

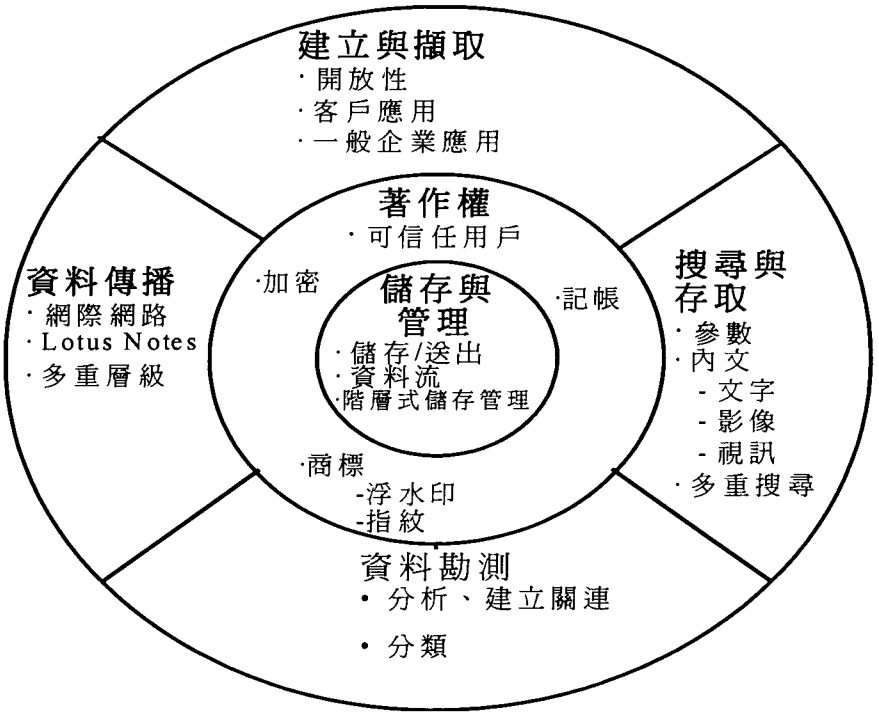
1. 就其他六個子計畫中的典藏資料作分析，歸類出所有的資料型態，設計適當的資料表示結構與轉換技術
2. 研究和發展數位典藏系統的建立與擷取技術
3. 進行研究並發展存放多媒體的資料庫，使資料庫能夠處理文字、圖片、聲音和影像資料
4. 開發多媒體資料之呈現、資訊勘測系統之建構以及資料管理之技術
5. 撰寫網頁、架設網站，與本校其他六個子計畫的數位畫成果進行整合

（四）計畫內容：

1. 資料建立與擷取
2. 資料儲存與管理之設計規劃
3. 資料存取系統之開發
4. 規劃整合計畫之技術交流網站
5. 整合計畫之技術交流網站實作
6. 資料儲存與管理相關技術的課程規劃
7. 資料儲存與管理相關技術的課程教學

（五）執行方法與過程：

為能夠有系統地將各種多媒體型態的資料數位化、整理與分類，並且建立一有效率的多媒體資料庫系統來存放。進而依據此資料庫，建立以 Internet 中 Web 為基礎之多媒體資料交換與展示平台，以提供友善的介面作為未來學術研究、學術資料查詢、及大眾教育之目的。



圖一、計畫各部分的功能概略圖

本計畫有六大主要研究項目，如圖一所示：

(1) 建立與擷取 (Create & Capture)

(2) 儲存與管理 (Storage & Management)

(3) 搜尋與存取 (Search&Access)

(4) 資料的傳播 (Distribution)

(5) 資料的安全性與著作權管理 (Security & Rights Management)

(6) 資料勘測 (Data Mining)

在以下內容，說明每一研究項目之研究方法與進行步驟：

1. 建立與擷取 (Create & Capture)：

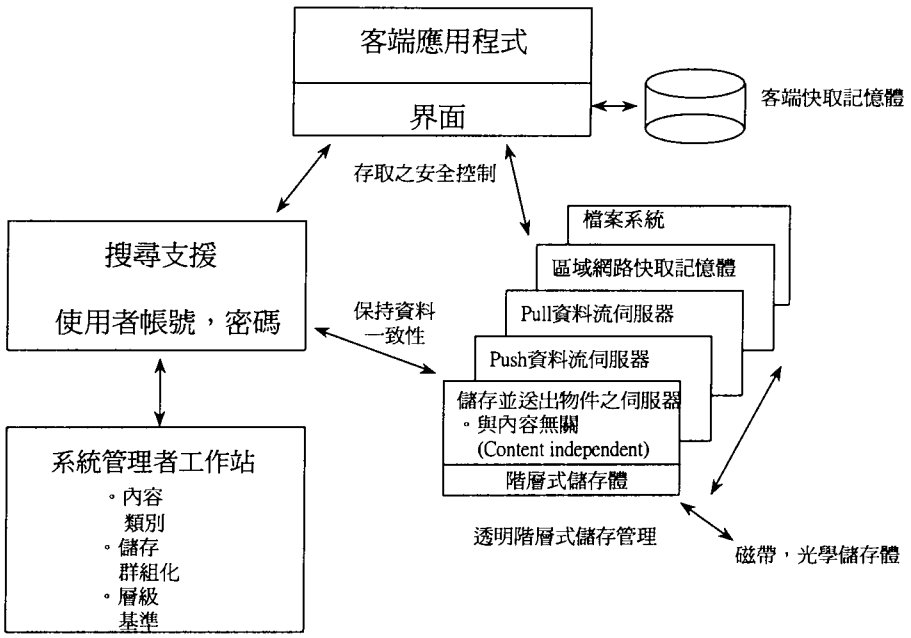
建構數位典藏的第一個步驟，就是要建置的初期內容並把已存在的資料作統一格式的轉移。用現今的電子軟體和硬體技術已能把現存部分的資訊變成數位格式的型態。目前有兩種常用的機制能把不同來源的數位化內容納入數位圖書館中：一種是利用目前存在的工具，把資料轉成數位化。另一種則是把原本不同格式的舊有數位化資料轉入到新規劃之數位圖書館的資料庫中。在此，本子計畫著重在軟體工具的開發與使用者介面的設計，以加速數位典藏的建置工作並可縮短操作人員的訓練時間。

類比資料包括書籍、手抄原稿、圖片、影片、影像和聲音唱片等，這些資料較易損毀且某些是較貴重的紀錄媒體。這些類比資料的不易保存是資料需數位化之重要誘因。這些類比資料可透過掃瞄技術把其數位化。本計畫已使用光學字元辨識 (OCR)、聲音辨識系統，影像壓縮系統等技術，把典藏的資料數位化。其中有關書籍與手稿部分可用光學字元辨識和聲音辨識系統技術把資料轉成文字檔。而圖片、影片、影像與聲音可用壓縮 (compression) 技術數位化。其目的是使資料能夠容易的在網際網路上讀取。

2. 儲存與管理 (Storage & Management)：

在各類的資料數位化之後，需要有穩定性好和可靠度高的媒體來儲存與管理這些資料。目前的儲存媒體種類有：(1) 磁性儲存體 (Magnetic storage)：如磁碟、磁帶等；(2) 磁碟容錯系統(Redundant

Array of Inexpensive Disks：RAID)：如磁碟陣列（disk array)，RAID0，RAID1 和 RAID5 等；(3) 光學儲存體（Optical Storage System)：有高密度、低成本的優點。如 WORM(Write Once Read Many)、可複寫式光碟片 (Erasable optical disk)、數位影像碟片(Digital Video Disk）和光碟櫃（Optical Jukeboxes）等。



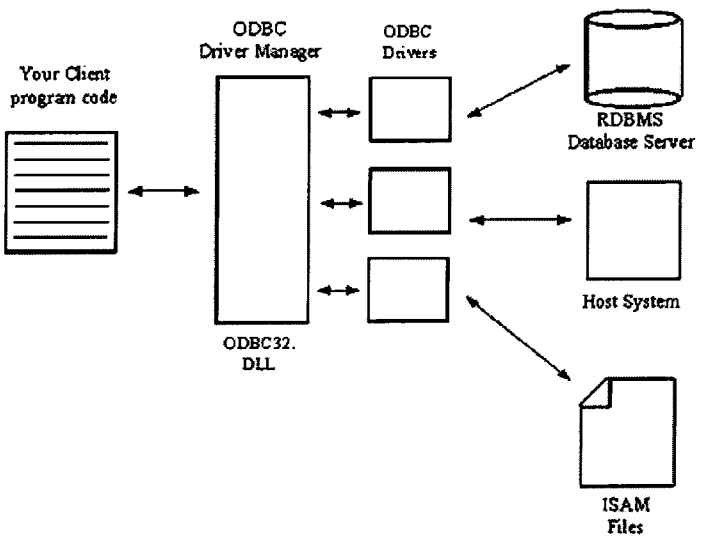
圖二：儲存體管理

圖二為儲存體管理架構圖。以典藏資料數位化後的龐大資料量來看，本計畫會以光學儲存系統（Optical Storage System）為主體。光學檔案系統（Optical File System；OFS） 是用來描述資料在光學媒體（optical media）中儲存格式的資訊。此外，OFS 也是在光學媒體中管理極大數量資料的儲存系統。它可被用在線上儲存和管理數十億筆文件資料的檔案系統。

此外，為了能夠在網路上讓多個使用者快速的存取資料。因此使用階層式儲存管理（Hierarchical Storage Management；HSM)。HSM 是一種多階層的儲存結構。最上層的儲存體速度最快並且與網路相連，但成本也是最高。越往下層，速度越慢，成本越低。檔案移動在

這些不同的儲存階層中。當檔案被存取時，它會移到最上階層。當使用的空間到達高標時，最不常使用的檔案被移往下一層。直到低標到達為止。當儲存體使用落在低標時，最常使用的檔案會帶回到較高層的儲存體。檔案的來回移動會增加使用命中率，因而減少反應時間。值的注意的是，在此，本子計畫使用應用資訊勘測技術來分析各個階層式儲存中的資料，快取具高度關連性的資料，進一步提高資料擷取的速度。

另一方面，本子計畫已對網際網路和資料庫系統做整合。我們已建立一套多媒體資料庫系統，使用者可透過 Web 瀏覽器來取得資料。因此，多媒體資料庫會提供兩種介面：ODBC（Open Database Connectivity）和 JDBC（Java Database Connectivity）。ODBC 是一個標準，而且是一個開放式資料庫存取的應用程式介面（API）。根據這個標準，即可存取資料庫中的資料。圖三為 ODBC 架構圖。

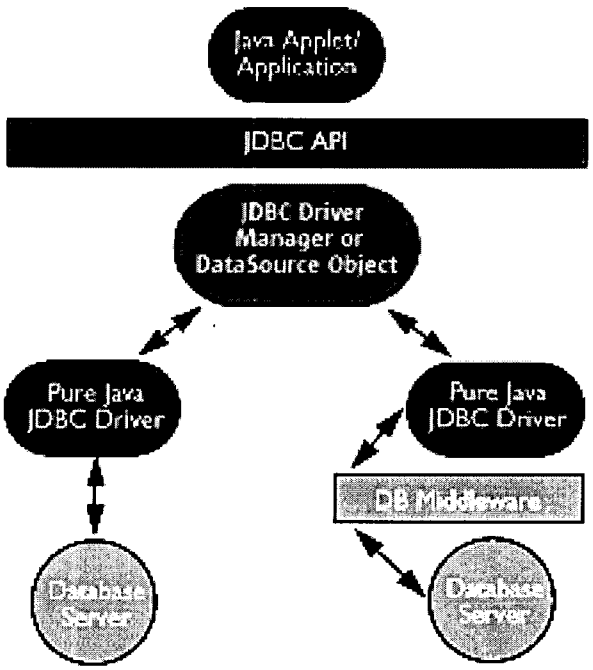


圖三、ODBC 架構圖

另外，JDBC（Java Database Connectivity）的技術可使網頁存取資料庫的資料。它是 Java language 的一組 API。而 Java 是近來在 Web 十分重要的語言，能使 Web 和使用者進一步的互動。所有的

Java 程式在執行前都透過編譯器編譯成一個虛擬機器（virtual machine）的機器碼。這些機器碼稱為 Java bytecode，Java bytecode 透過一個解譯器（interpreter）便能被執行。Java 的跨平台特性，使得 Java 非常適合在多平台的網路環境使用。JDBC (Java Database Connectivity) 的發展，使得 Java 程式開發者透過 JDBC 套件，可以容易的發展出資料庫系統的應用軟體。

JDBC 類別程式主要提供的功能有三：(1) 連結到任何資料庫、(2) 執行標準的 SQL 查詢指令、(3) 處理所得到的查詢結果。其目的是要讓發展者把精神花在資料庫中各種資料結構的設計工作，而不是用在 Java 和不同資料庫連結的各種細節。總而言之，Java 和其他語言最大的不同在於它的 Platform-Independent。因此，Java 在提供資料庫解決方案時，也要有一個開放性且是 Platform-Independent 的解決方案。也就是說同一個程式在不需要重新編譯的情況下，就可以任意換置在後面的資料庫系統。而達到『寫一次、編譯一次、到處執行』的特性。我們認為，Java 的特色對於整合多個子計畫所發展出的平台可很大的助益，因為數位化典藏的資料量是非常的龐大，資料極可能分別存放在多處並利用高速網路串聯起來，這樣的環境之下，發展一個具有分散式能力的軟體平台是很空困難的。



圖四：JDBC 之架構圖

圖四為 JDBC 之架構圖。在 JDBC 的架構下，DBMS 有獨立介面，如此可達到一般的 SQL 資料庫存取的架構和透過單一介面可使用不同資料源等目的。此外，發展者使用 JDBC 只需寫一次資料庫介面，程式可存取任一個 data source 且並不需要改變程式。

本計畫已利用上述的儲存管理系統來建構多媒體資料庫，並提供 ODBC 與 JDBC 之介面建立網頁，方便使用者透過網際網路搜尋與存取本資料庫系統。

3. 搜尋與存取（Search & Access）：

數位典藏中最主要的功能就是資訊探索（Information Discovery）。目前的資訊探索模式有三種：

◆ 瀏覽（Browsing）：使用者要讀取其有興趣的文件，其連結（Link）關係或網頁（page）關係並沒有任何依據來做決定。因此，也稱為無結構之漫遊（unstructured tour）。

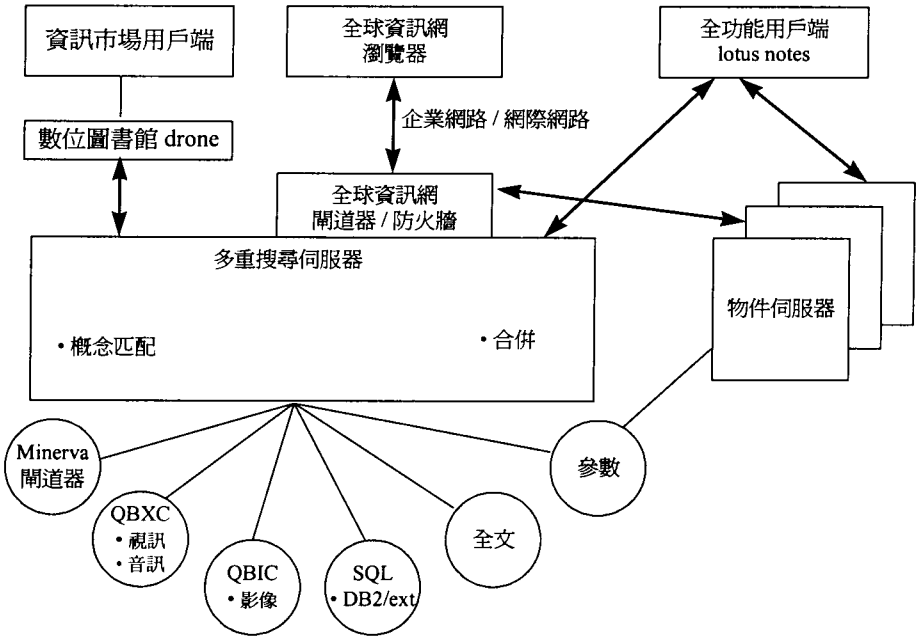
◆ 領航（Navigation）：由某機構開始組合編撰文件和物件變成連貫有組織的收集。使用者是運用策略或被引導沿著總體的資訊樹（Information tree）讀取有興趣的文件。此方式稱為有結構之漫遊（Structured tour），例如：Yahoo。

◆ 搜尋（Search）：由使用者送出 query 到資訊來源處。接著呈現出結果的概要。例如：AltaVista、Excite 等。

其中，資訊來源處可能是全文索引模式（Global index model）和分散式搜尋模式（Distributed search model）。全文索引模式是利用自動代理程式(robot 或 agent)在全世界的全球資訊網（WWW）的超文字架構（Hypertext structure）來回移動，遞迴式取回所提到的網頁或是用多個不同追蹤策略的 robot 所取回的網頁，來建立獨立的索引（indexes）。客端傳送關鍵字（keyword）到伺服器（server）端，利用此索引來找尋資料，並回傳結果回到客端。

而分散式搜尋模式的來源則是中間媒介或是中間搜尋器。它們選擇和轉播查詢（query）到數個實際的來源處，並取回個別的結果，而且結合和整理這些結果變成一個單一的資料集再回傳到客端。

在伺服器端，搜尋資料的策略已不是找出文件有關鍵字出現的方式，而是內容搜尋（content-based search）方式。此方式是依照關鍵字的意義與其概念來搜尋文件。因此，有可能意義相同但無關鍵字出現的文件被找出。此外，影像型態的搜尋可使用影像內文查詢（Query By Image Content：QBIC） 技術。QBIC 是允許使用者依據物件的顏色、材質、外觀和大小，在影像資料庫中搜尋出符合條件物件的影像。



圖五：搜尋與存取

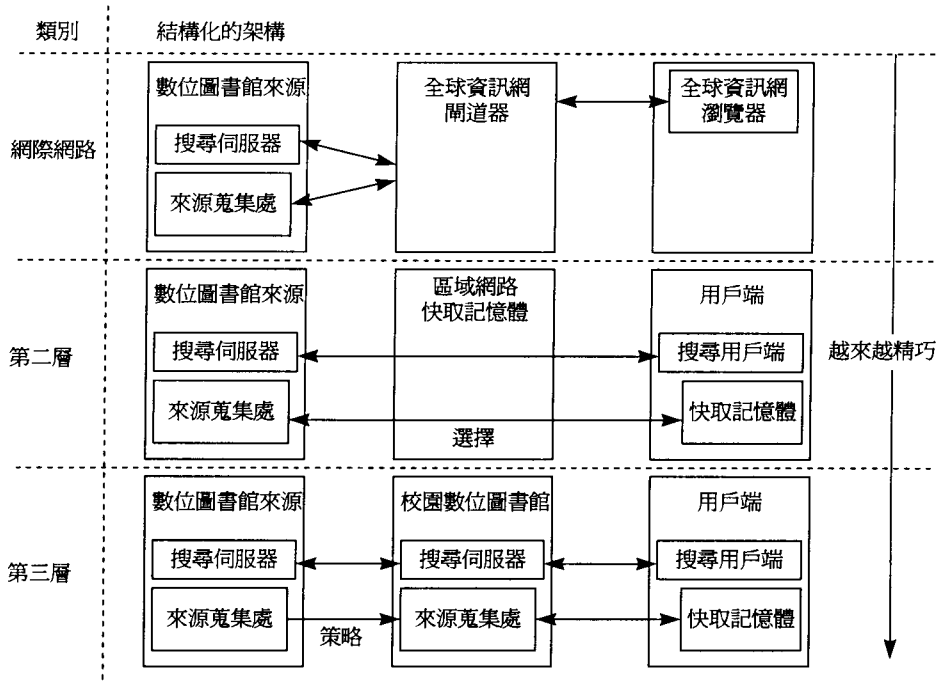
數位典藏除了提供各別資料型態的搜尋外，也提供了混合型態的搜尋，如圖五所示。例如：輸入某人的姓名，則會找出有關此人之文件、圖片、影像及聲音等資料。為了能夠達到混和式資料型態搜尋，已使用 metadata 的技術。而 metadata 是描述物件的特徵，如出版日期、作者名稱或是在書中的頁碼等等資料。因此，資料的搜尋是針對 metadata 內對物件特徵描述的內容進行搜尋，如此即可對不同型態資料進行查詢。

在本計畫中，我們已提供標題搜尋外，也提供全文檢索的搜尋功能。除了文字搜尋外，也進一步使用 metadata 的技術來搜尋多媒體型態的資料。

4. 資料的傳播 (Distribution)：

把資料數位化後，會把資料放在網路上供使用者查閱。因此借書的方式不像現在要親自到博物館建築物中借閱書籍。其行為模式與目前瀏覽網站相同，亦即在客端使用圖形介面的瀏覽器來查閱資料。圖

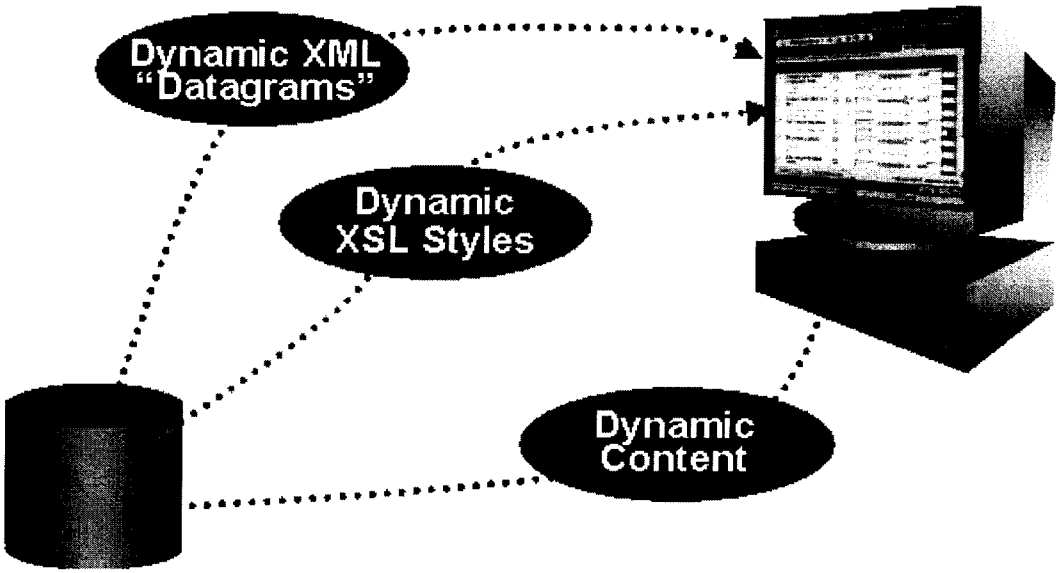
六為分散式存取架構圖。



圖六：分散式存取架構

由於本計畫的目的是要協助典藏資料數位化後放上網際網路上，使用者透過全球資訊網存取資料。然而，我們所要數位化的資料範圍相當的廣，有植物標本典藏數位化、動物博物館典藏數位化、昆蟲標本典藏數位化、台灣地質科學典藏數位化、人類學典藏數位化以及台大博物館典藏數位化等，各領域的資料類型與型態完全不同。因此，我們把解決使不同類型資料整合在一起，並且使用同一格式顯示。我們擬採用 XML（eXtensible Markup Language）來解決不同類型的資料呈現在網頁上的問題。可延伸性標誌語言（XML）是 SGML（Standard Generalized Markup Language）的簡化版，SGML 是用來形容和定義結構化的電子文件之標準。因此，XML 包含了 SGML 的特性，如使用"文件形態定義"（Document Type Definition, DTD）來指定文件的結構、轉換成為多種的輸出格式以及提供所有資料一致的顯示外觀，方便維護與管理。因此，我們可以使用 XML 的技術來達

成不同類型資料用相同格式來顯示的目的。



圖七、XML 示意圖

XML 是另一種在 WWW 上資料呈現的方法，它不像 HTML，是一種版面描述語言，是一種單一的標誌語言。而 XML 是一種 meta-language，可以被用來定義任何一種新的標誌語言。XML 把資料的型態與資料的呈現方式分開，這種資料導向的好處是同種資料可以再不同的地方以不以不同的方法呈現出來。因此，XML 使得網際網路上的資料呈現並不限制於使用者端條件的限制。可以保證資料不會有無法閱讀的問題。圖七為 XML 示意圖。XML 網頁包含三個部份：XML 宣告（XML Declaration）、文件形態定義（Document Type Definition ， DTD）和標示文件成品（Document Instance）。

5. 資料的安全性與著作權管理（Security & Rights Management）：

資料在網路上傳輸最擔心的是資料的完整性與安全性，因此對數位資料作保護是必要的措施。然而，容易監聽、攔截且無認證功能正是目前網際網路安全上的缺點，相關技術之研發乃極為重要：

◆加密（Encription）：對原始檔案進行加密的動作，到達client端在進行解密，這個目的是在保護資料在傳輸中的安全。

◆使用者認證（User authentication）：每個使用者都有其專屬的帳號和密碼。當使用者要閱讀需付費的資料，就會被記錄。在依據此記錄向使用者收費，必要時也可配合電子簽章與電子證明書的技術，使得使用者無可否認。

◆浮水印（Watermark）：在影像資料保護方面，為避免遭人使用影像處理軟體竄改原來的影像，可在每張圖片或是影像上加上專用的浮水印，增加竄改的困難度。另一種保護方式是在使用數位相機攝取影像後立即給每張照片專屬的關鍵（key）值，數位相機再進行編碼。編碼後的影像資料只有此關鍵（key）值能夠解開。因此，若試圖去修改圖片內容，則將無法再用原來的關鍵（key）值去解開，而達到保護的目的。

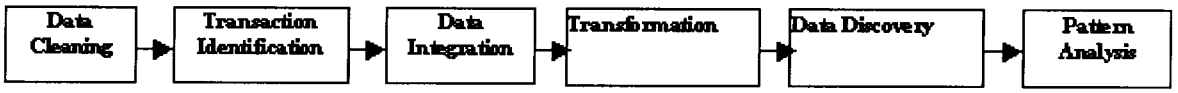
在數位化的世界裡，網路上的資料傳輸就是拷貝的行為。所以，保護的目的是除了需兼顧符合資料品質標準的完整性和資料公開所需之安全性外，也必需維護著作者之智慧財產權。因此，本計畫已研究這些技術來保障各子計畫系所的著作權。

6. 資訊勘測（Data Mining）：

資訊勘測（data mining）意指由許多處理過後的資料去尋找新的資訊或學習新的知識。隨著網際網路的發展，在Web之環境下作資訊勘測（稱之為Web data mining）已成為日漸重要之課題。有鑑於此，我們已在本計劃中研究此方法對數位典藏資料的擷取可能帶來的好處，透過收集和記錄使用者運用一個以 Web 為主之系統的各種資訊，我們已應用資訊勘測（Data Mining）的機制來分析使用者透過瀏覽器來觀看數位圖書館的網頁的紀錄。藉此找出資料間的相關性。

並由此建立各領域的導覽路徑，幫助使用者能夠更快的取得資料與更有效率的學習新知。

使用者記錄資料被處理後所得到的資訊包括有：(1) Association（相關性），即何類之資料網頁常被一起使用；(2) Classification（分類），即何類之使用者常存取何類網頁及取用何類資料等；(3) Sequential Pattern（順序性），即使用者瀏覽資料網頁之順序性；(4) Traversal Patterns（路徑），即使用者在網路上瀏覽之路徑等。圖八為資訊勘測的分析步驟。



圖八、資訊勘測處理流程

此外，在本計畫中，我們也已會以資訊勘測技術來對所建立的數位典藏資料作完整的分析與分類，從龐大的典藏資料中“勘測”出不同文件或檔案資料之間的相關性，以增進『典藏數位化』的功能，讓使用者在對館藏作查詢後，能夠充分地掌握相關的資訊。

（六）關鍵字：

資訊勘測, 資訊探索, 階層式儲存管理, 全文索引模式, metadata, 可延伸性標誌語言, 安全性與著作權管理, 浮水印, 相關性規則, 分類, 順序性模式, 瀏覽路徑模式(Traversal Pattern)。

貳、執行成效

（一） 工作進度報告

1. 典藏內容類型與特色說明（本項資料分項計畫可免填）

此子計畫屬於技術開發與支援的計畫，故無典藏內容的產出。

2. 預定與實際執行甘梯圖

請參照附件一。

3. 執行困難說明

此子計畫執行上，並沒有遇到太大的困難。

4. 落後原因說明、因應對策、檢討與建議

此子計畫進度都依規劃之進度進行與完成。

（二） 經費執行運用與說明

此子計畫經費執行都依規劃之進度進行並完成採買設備與建制相關的入口平台網站。

4. 智慧財產權處理

此子計畫尚無重要的研發成果可以申請成為智慧財產與專利。

參、網站表列

肆、檢討與建議

i. 由於各子計畫有提出 Data Center 的建議與需求，用以在統一儲存與管理所數位化之資料，但建制與管理 Data Center 並沒有規劃在此計畫的工作內容中，所以有執行上的困難。

ii. 由於部份子計畫執行單位的資訊環境不佳，如電力供應不穩定、網路速度與頻寬不敷使用。因此，期以建置一 data center 可以提供電力與網路品質穩定與良好的資料儲存環境。

iii. 有鑒於各子計畫在執行資料數位化過程期間，遭遇到無法聘請到具有資訊相關人才來長期協助製作與維護各子計畫的數位化成果，而使得數位化的成果難以被妥善的維護與保存。因此，希

望透過 data center 的管理人員來協助資料數位化與維護數位化成果。

iv. 子計畫七的定位在如何建置一藉由資訊勘測技術來分析數位成果使用狀況的入口網站，同時也可提供相關技術與經驗給其他子計畫以協助資料數位化。因此，子計畫七非定位在協助其他子計畫在製作與維護數位化的資料。

91 年度預定與實際執行甘梯圖 預定進度 實際進度

工作項目 \ 月次		第1月	第2月	第3月	第4月	第5月	第6月	第7月	第8月	第9月	第10月	第11月	第12月
1	資料建立與擷取												
2	資料儲存與管理之設計規劃												
3	資料存取系統之開發												
4	規劃整合計畫之技術交流網站												
5	整合計畫之技術交流網站實作												
6	資料儲存與管理相關技術的課程規劃												
7	資料儲存與管理相關技術的課程教學												
8													
9													
10													
11													
12													
進度累計%	預定	8	16	25	33	41	50	58	66	73	80	90	100
	實際	8	16	25	33	41	50	58	66	73	80	90	100