數位圖書館／博物館的索引典建置系統

Thesaurus Construction System for Digital Library/Museum

陳光華
Kuang-hua Chen
國立台灣大學圖書資訊學系副教授
Associate Professor, Department of Library and Information Science, National Taiwan University

【摘要】

就 WWW 資訊檢索應用而言，資源必須經過有系統地組織與整理，才能得到滿意的檢索結果。控制詞彙的角色在前述的背景下，扮演越來越關鍵的角色。控制詞彙通常是透過索引典或是主題表體現，提供詞彙之間的等同關係、階層關係、與關連關係。藉由索引典，使用者可以瞭解資訊系統的知識結構，可以選擇適切的檢索詞彙，以獲致比較滿意的檢索結果。然而，索引典的建構與管理卻是一項困難工作，若沒有設計良好的系統，則無法有效地發揮索引典的長處。目前沒有資訊檢索系統或是數位圖書館系統完全整合索引典子系統，使得檢索效能始終不高。本文探討索引典的建置，發展索引典系統的功能模式，建立索引典建置與管理之模式，並實作雛型系統。

【Abstract】

From the viewpoint of WWW-based IR researches, the metadata and resource description framework become much more important than ever before. Therefore, the application of the controlled vocabularies increasingly plays an important role. The recent literature also report that controlled vocabularies significantly improve IR effectiveness. Thesauri and Subject Headings consist of controlled vocabularies and represent the knowledge structures in various relations in details. However, the construction and management of thesauri is not an easy task and the integration of thesaurus subsystem into IR systems or Digital Library/Museum is not considered in the current IR systems or DL projects. The paper will investigate the applications of thesauri and subject heading, propose a model to handle the thesauri or the subject headings in a systematic way, and implement the prototype of a practical system.

關鍵詞：索引典、數位圖書館／博物館
Keyword: Thesaurus, Digital Library/Museum

壹、序論

網際網路的時代使得吾人可以輕易地取得資訊，然而資訊的品質則是令人質疑的重要問題，爲了提升資訊的品質，各種網路的服務應運而生，包括原生型服務、

加值型服務、訊息性服務。（註 1）近年來，數位圖書館研究逐漸受到大家的重視，其目的是希望提供綜合型的服務，讓使用者可以透過數位圖書館系統滿足資訊的需求。因而，一個嚴謹的數位圖書館系統，文獻資料在典藏之前，必須經過分類編目的過程，編目人員通常進行兩種不同的分析方式：一爲實體分析（Physical Analysis），處理的是文獻的作者（Author）與題名（Title）等資料；另一爲內容分析（Content Analysis），處理的是文獻的分類以及爲該文獻設定若干標題（有時必須複分）。文獻的作者與題名是沒有任何爭議的，然而文獻的分類與標題則是必須經由編目人員一番思考，才能妥善處理的。爲了達到較爲一致的處理方式，圖書館文獻資料的分類編目是採權威控制的方式進行的，也就是說分類有既定的分類法，如國會圖書分類法（LCC）、杜威十進位分類法（DCC）、中文圖書分類法；標題則有既定的標題表，如國會標題表（LCSH）、醫學標題表（MeSH）、中文圖書標題表。圖書館編目館員根據前述權威控制的方式進行文獻資料的編目，而每一筆文獻資料的編目資料（是一種 Metadata，也就是詮釋資料）就成爲該文獻資料可能的檢索點（Access Point），至於所使用的詞彙則被稱爲索引詞彙。透過詮釋資料，可以檢索各種類型的資料，如文字、圖片、音訊、視訊，因而，詮釋資料的著錄益形重要。圖書資訊學界對於資料的著錄、組織與整理，已有長久而完整的作法，如何將之運用於資訊檢索系統或是數位圖書館，將是重要的研究課題。

索引詞彙事實上扮演兩個角色，其一是讓資料著錄者選擇適當的詞彙，以表達所處理的文獻資料的主題；其二是讓資料檢索者下達適當的詞彙，以檢索經適度處理的文獻資料。索引典或是標題表承載者控制詞彙，因此必須與索引者以及檢索者有效地互動，才能取得令人滿意的檢索結果。然而，除了製作精良的光碟資料庫之外，目前網際網路上的檢索系統使用索引典的情形並不多見，本文將探討索引典實際應用的情形，分析索引典系統應具有的功能與規格，並實際發展一套雛型索引典建置系統。這套索引典建置系統，不僅可以定義詞彙關係，輸入控制詞彙，驗證詞彙關係，瀏覽控制詞彙，查詢檢索詞彙，還可以整合於資料著錄系統與資訊檢索系統，有效提升著錄與檢索的一致性。

本文的結構如下：第貳節將探討索引典的相關研究；第參節將說明索引典建置系統的研究方法；第肆節探討索引典的使用需求；第伍節說明本研究使用的軟體工具或技術；第陸節則說明索引典建置工具程式的使用流程與各項功能畫面；第柒節則是簡短的結論。

貳、文獻探討

網際網路的發展使得吾人對於資源取得的管道與以往有極大的不同，同時各種資源型式的比重也隨之不同。數位圖書館的研究在前述的背景之下受到各國政府、學術、商業等機構的重視，紛紛投入大量的人力物力。由於數位圖書館是科際性（Interdisciplinary）的研究，各領域的專家紛紛由不同的角度審視數位圖書館，例如有些學者認爲數位圖書館是資訊檢索系統的延伸，而有些學者認爲數位

圖書館是實體圖書館虛擬化的延伸，更有些學者認爲數位圖書館就只是網路化的光碟資料庫。對於將數位圖書館視爲是實體圖書館延伸的研究者而言，經過多次的調查與研究，得知使用者使用 OPAC 或 WebPAC 的檢索模式。例如，根據美國圖書館學會的調查，公共圖書館的讀者多數使用標題檢索，學術圖書館與專門圖書館的學者多數使用題名或著者檢索；而耶魯大學的研究顯示，56%是題名或著者檢索，而 33%是主題檢索（包括標題與分類號）。（註 2）然而，隨著圖書館作業逐漸的自動化，線上查詢也成爲檢索作業的主流，讀者檢索的方式也有所改變，根據台灣大學黃慕萱教授 1996 年所做的調查，有 82.5%的檢索是屬於主題檢索。（註 3）由上述的數據顯示，主題檢索是資訊檢索的主要方式。數位圖書館系統透過網際網路提供各類型的資訊服務，對於資訊檢索的功能而言，主題檢索的重要性將益形重要，因爲使用者更加地多元化，無法預期使用者具有何種背景知識，因而，數位圖書館的研究者與建置者必須強化主題檢索的功能，提供額外的服務子系統，如索引典子系統，有效協助使用者主題檢索的需求。

1998 年國家科學委員會有鑑於數位圖書館／博物館的研究越來越重要，因此推動「數位博物館先導計畫」。台灣大學執行其中的「淡水河溯源」與「資源組織檢索規範」二個先導計畫，其中「淡水河溯源」屬於主題型計畫，而「資源組織檢索規範」屬於支援型計畫。本人於「資源組織檢索規範」計畫負責資訊檢索服務系統的設計，主要的研究工作著重於設計資訊檢索與詮釋資料著錄的服務系統，並建構一個使用淡新檔案索引典的系統，包括詞彙瀏覽、詞彙檢索、以及詞彙對映等三部分的功能。詞彙瀏覽之機制可以讓使用者藉由瀏覽的方式，得知系統索引典的架構，瞭解詞彙之間的關係。首頁可顯示該索引典第一層的詞彙，藉由點選的方式可以瀏覽第二層的詞彙。詞彙檢索之機制讓使用者檢索索引典，提供二種檢索方式：檢索索引典記錄及檢索索引典詞彙。使用者於「檢索詞彙」之查詢文字框輸入欲查詢之詞彙，可以選擇完全比對或是近似比對二種模式，系統即送回檢索結果。詞彙對映之機制允許使用者使用系統索引典沒有收錄的詞彙，當使用者使用的檢索詞彙不是系統使用的索引詞彙時，本對映機制會將其自動轉換索引詞彙。例如，詞彙「學生」將被轉換爲索引典使用的「貢生」、「功名」、「文童」、「武童」等詞彙，轉換的模式有大規模的擴展與小規模的擴展二種，大擴展的轉換比較不精確，而小擴展比較精確。（註 4）

紐西蘭學者 Alastair Smith 曾經檢視 11 個數位圖書館，仔細比較各數位圖書館提供的檢索功能，多數的系統提供布林運算功能，但僅有五個提供控制詞彙的功能，而且僅有二個系統提供相關詞彙的功能。（註 5）由這些情況顯示，目前的數位圖書館系統的建置，僅著重於將典藏品數位化，提供主題式的數位收藏環境，設計良好的展示網頁，結合網際網路上既有的搜尋引擎或是分類目錄，然而，並沒有實質上提升檢索服務的層次。對於習慣於以檢索方式取得資源的使用者而言，如何強化數位圖書館系統的檢索功能是最爲關切的課題。實體圖書館對於文獻資料的主題有一套系統化的處理方式，無論是索引典或是標題表代表的是系統內部的知識結構，而個別的詞彙則代表特定的知識或概念。因此，若能使用自動、

半自動或人工的方式，賦予文獻資料適當的敘述詞（Descriptor，亦即索引典或標題表使用的詞彙），而檢索系統又能有效地運用，則數位圖書館服務的層次必定能夠提升。

在索引典使用於資訊檢索系統或是數位圖書館／博物館系統之前，都必須實際建構索引典，例如黃慕萱教授建構了「淡新檔案索引典」，本人才能建構淡新檔案索引典瀏覽與檢索系統，並將之與資訊檢索系統整合，而後將之運用於查詢問句的擴展。索引典的建構有自動與人工的方式。輔仁大學曾元顯教授嘗試建構統計式共現索引典（Co-occurrence Thesaurus），評估其效能，並討論其可能的應用。（註 6）所謂的共現索引典與圖書館界所稱的索引典並不完全相同，圖書館界使用的索引典，詞彙間有明確的等同關係（Equivalent Relation）、階層關係（Hierarchical Relation）、或是關連關係（Associative Relation），但是運用統計式的方法，自動建構的共現索引典，只能說詞彙間有關係，但是無法自動辨別是哪種關係。雖然如此，運用共現索引典於資訊檢索系統，仍然有很大的幫助，因爲它能夠擴展使用者的查詢問句，提高求全率（Recall）。

人工建構索引典，則能夠明確地界定詞彙的關係，雖然成本很高，但是其使用的價值相對也高。人工建構索引典，也有其他的問題必須克服，例如，在索引店建置過程中，廣狹義詞可能形成一個循環，造成詞彙關係的謬誤，詞彙的選擇與修正等等，因此必須要有一個索引典建置工具，協助建置人員。MultiTes 索引典管理與建構工具是一個蠻有名的工具程式，提供使用者自訂詞彙關係、遵循 ANSI/NISO 標準關係、支援多語言、支援多使用者、提供各種索引典格式輸出。（註 7）也有一些索引典工具程式僅僅是索引典視覺化程式，它提供很方便的方式，讓使用者瀏覽索引典、檢索索引典、或是列印索引典，如 THESshow（註 8），Thesaurus Viewer（註 9）。這些工具以 MultiTes 最爲完整，但是，它卻忽略了索引典建構過程可能發生的錯誤，如詞彙關係的驗證。本文將描述我們發展中的索引典建置工具程式，探討各項功能及實作的方式。

有關索引典的建置、評估、與應用，可以參考註 10 所列的參考文獻。

參、研究方法

本研究將檢視索引典的效益，探討索引典的建置，發展索引典系統的功能模式，建立索引典建置與管理之工具，研究索引典系統與檢索系統的整合架構。具體之研究方法如下所示。

1. 文獻探討法
   檢視索引典與標題表的相關論文，探討索引典的評估，以及索引典相關的應用。
2. 系統分析法
   調查線上索引典的使用需求以及現有之瀏覽模式，研究並評估與檢索系統整合之模式。

3. 系統實作法

基本上我們認爲數位圖書館是由索引典子系統、詮釋資料著錄子系統、資訊檢索子系統、資訊典藏子系統等四個核心子系統，以及其他的子系統（如人機介面子系統）整合而成。索引典子系統必須提供控制詞彙給資料著錄子系統使用，以確定詮釋資料著錄之一致性。索引典系統必須提供各資訊檢索子系統使用的控制詞彙，以增進資料檢索的有效性。索引典系統必須提供系統領域知識架構，以加強使用者對數位圖書館典藏品的瞭解。在前述的考量之下，並配合當前電腦的使用環境以及網際網路以成形的標準，索引典系統的功能與規格如下。

- 執行於 Windows 作業系統，透過網路與其他異質性系統互動。
- 檔案儲存格式：以 MS Access 檔案格式爲內部檔案的儲存格式。
- 檔案瀏覽型式：以首頁展現階層架構的第一層詞彙，每一詞彙皆可點選，以展現其下之第二層詞彙，同時展現詞彙的完整路徑，讓使用者藉由瀏覽方式瞭解系統的知識架構。
- 詞彙關係的定義：除了階層關係、等同關係、關聯關係之外，建置者可自訂詞彙關係。
- 詞彙的輸入與編輯：輸入詞彙，依據定義的詞彙關係，建構輸入詞彙關係。
- 詞彙關係的驗證：檢查詞彙之間的關係是否相互衝突。
- 多語詞彙的處理：因應網際網路的應用需求及其跨越國界的特性，必須處理多種語言的詞彙。

肆、索引典需求描述

一、詞彙描述的需求

使用者在詞彙的描述上，其名稱可能包含一種以上的語言，在使用上，不論是用英文，或是中文或是其他種類語言爲載體來描述的一個詞彙名稱，必須具有完全相同的意義與地位，也就是說對詞彙本身而言，語言只是一種標籤，不具備定義詞彙的能力，詞彙本身擁有自己的性質，不受到語言名稱的影響。

在使用上，使用者要能夠自行設定詞彙各種語言的名稱，而各個詞彙在軟體內具備有唯一性（Unique），也就是說，由不同語言所描述的同一詞彙，必須擁有屬於自己的 Identity，不與其他不同的詞彙混淆。在這樣的定義之下，兩個不同的詞彙可以允許具備同一種語言的名稱相同（在某些語言中對某些性質的描述較其他語言短少），而藉以辨別兩個詞彙不同的依據爲軟體內的 Identity。

二、詞彙關係定義上的需求

詞彙與詞彙之間存在關係爲索引典建立的依據，而利用詞彙的關係藉以達到語言中介與轉換，更是索引典最終的目標。就使用者而言，詞彙間的關係究竟具備怎麼樣的性質，才能完備的描述詞彙的架構，同時讓使用者在建立時的容易認知，與資訊系統在解讀時的嚴謹邏輯。在本文將『關係』做了幾項邏輯上的描述，

以方便往後軟體設計上的邏輯根據，和使用者使用軟體建置索引典的根據。在關係的描述上，將『關係』分爲不同的幾種性質：

1. 階層關係

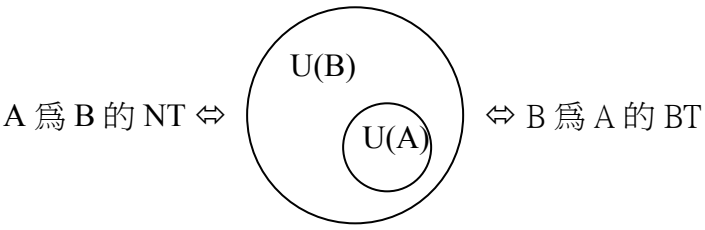
在關係上描述從屬、父子、階層關係，其關係意義爲“Has-A”的關係；而此種關係在定義上成對出現，並且相互對應。同時具備樹狀的結構；但是詞彙間在具備此類關係時，可以爲多對多的對應關係。以邏輯的方式描述可以爲此種關係定義下列的邏輯觀念：假設有 A、B 兩個不同的詞彙；有兩關係 NT 與 BT，爲成對的階層關係， BT 代表上位關係； NT 代表下位關係。A、B 兩詞彙具備這樣的關係，在描述上，我們用下列的方式描述：

A 爲 B 的 NT ⇔ B 爲 A 的 BT

或是 A “Has-A” BT B ⇔ B “Has-A” NT A

或是 A→NT→B ⇔ B→BT→A

從另外一個角度來看，若將 A、B 兩詞的描述範圍視爲兩個集合 U(A)、U(B)，兩者同時還具備有從屬或者是包含的關係。



舉例來說：若 A 詞彙是鳥類而 B 詞彙是麻雀，可以瞭解 A 爲 B 的 BT；若 A 詞彙是貓而 B 詞彙是寵物，因爲兩者的描述範圍無法完全爲對方涵蓋，所以二者無法構成上位或下位的關係。

2. 雙向關連關係

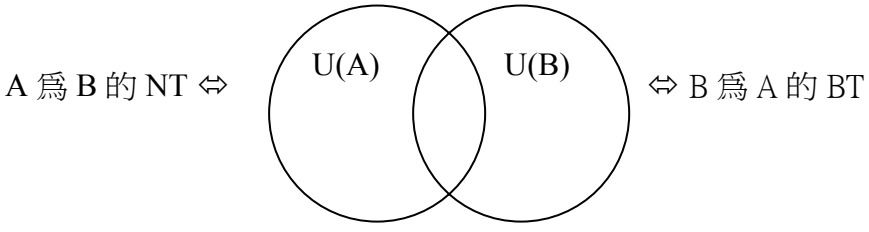
此種關係描述詞彙具備某些延伸相同的意義，可以互相參照，由於關連性是建立在兩個詞互相關連的基礎上，必須要同時存在，這也是一種“Has-A”的關係。假設有 A、B 兩個不同的詞彙，有雙向關連關係 RT，可用下列的方式描述：

A 爲 B 的 RT ⇔ B 爲 A 的 RT

或是 A “Has-A” RT B ⇔ B “Has-A” RT A

或是 A→RT→B ⇔ B→RT→A

從詞彙描述範圍的界定來看，可以用下列方式表示：

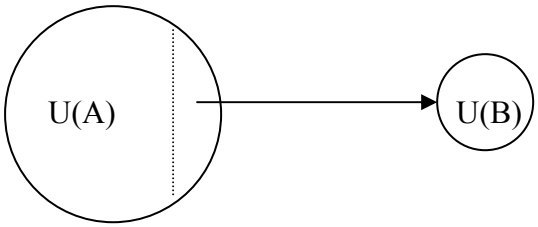


3. 單向聯想關係

在系統中此爲唯一的單向關係，在條件上最爲容易滿足，在定義上，只需要單一詞彙的單方面定義就能存在，另一詞彙並不需要另外定義兩詞的關係。其關係依然爲“Has-A”的關係。假設有 A、B 兩個不同的詞彙，有單向關連關係 PT，可用下列的方式描述：

B 爲 A 的 PT ⇔ A “Has-A” PT B ⇔ A→PT→B

⇔



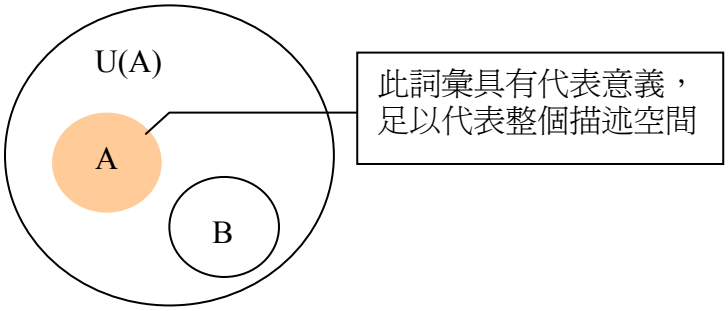
4. 等同關係

顧名思義，具有這種關係的二個詞彙，其意義完全一樣，但是還是必須定義其中之一爲敘述詞，敘述詞爲權威控制的詞彙，另一爲款目詞（Entry Term）。假設 A、B 二詞彙在語意上完全相同，有兩種關係 USE 與 UF（Use For）用以表述這種關係。

A 詞彙 UF B 詞彙 ⇔ B 詞彙 USE A 詞彙

其關係意義爲“Is-A”的關係（A UF B ⇔ B USE A）

以詞彙描述範圍來界定，可得到下列的邏輯：



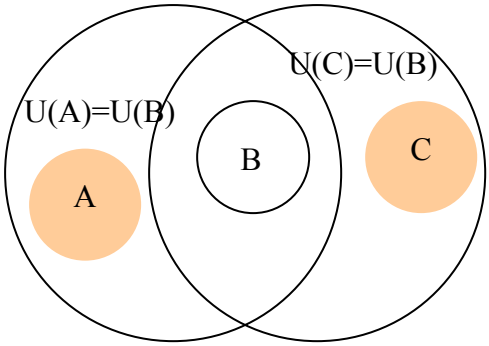
而由於詞彙控制上的需要，這兩種關係同時具備有其他邏輯上的限制，以確定控制詞彙使用上的合理性，以上述的假設定義，具備有下列的邏輯。

A UF B => A 對 B 存在 1 對多的映射關係

也就是說 A：[U(A)=U(B)=…] ＝ 1：N

若是存在詞彙 C：U(C)≠U(B) => A NOT UF C

一個詞彙群不能同時擁有兩個不同的敘述詞，也就是說兩個敘述詞所代表的詞群不能有交集的詞彙，也就是以下的狀況不可能出現：



三、詞彙、關係建立的便利性

使用者在使用索引典工具程式時所期待的介面（Interface）型式，怎麼樣的編排方式會讓使用者感到最親切，有最少的學習困難，同時也達到最大的功能，以下就幾個分類來做討論。

1. 視窗化介面（Visualization）

現階段程式的使用，由於個人電腦的普及，程式的使用，大多採用單機作業的方式設計，再加上微軟視窗作業系統在全球造成的熱潮，使得程式視窗化變成必然的趨勢，如此能降低每個使用者的學習曲線，在較短的時間內學習會使用程式的能力。在這邊，將所製作軟體的執行背景作一個簡短的描述，程式必須能在微軟的視窗作業系統下運作正常，並且所有的動作都是由視窗化的介面清楚呈現。

2. 重複（Redundancy）詞彙或者關係的處理

對於索引典建置上，使用者期待軟體能在建置時提供一些功能以方便建置和往後的除錯。在除錯方面，首要的就是詞彙重複的問題，索引典建置軟體，必須要能夠對於使用者已經建立的詞彙在再次鍵入時提出警告，以免往後索引典內有重複的資料；在詞彙關係的建置上，也必須具備相同的功能。

3. 階段化（Serialization）製作

在索引典的製作上，不可能要求使用者一次做完，在建置是一定是分階段完成，在建置軟體就必須具備存檔和讀檔的功能。如此索引典檔案才能便於製作、儲存、備份、複製…等資料處理與利用。

4. 關係之邏輯檢查（Logical Validation）

以傳統的方式建構索引典，在詞彙關係的邏輯驗證上力有未逮（在關係很多很複雜時，容易產生邏輯錯誤而很難被察覺），在索引典建置軟體中，透過電腦軟體的輔助，必須要能避免這種狀況的發生，在使用者建置的同時提出警告；另外一方面，也需要在開啓索引典檔案前先做邏輯的檢查，以確保檔案是否被改變，造成邏輯上的謬誤。

四、檢索和展現的方式

1. 檢索的方式

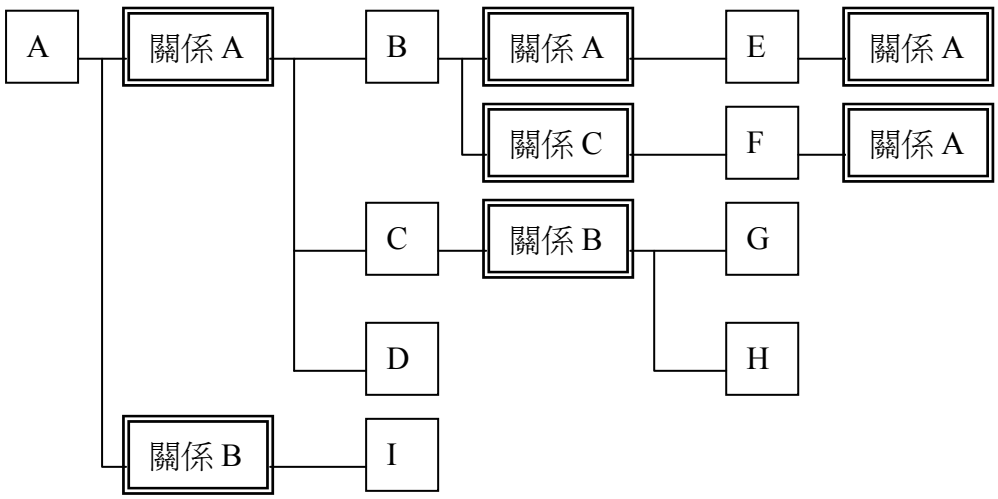
以索引典使用的方式來看，檢索的目的在瞭解某個特定詞彙和其他詞彙的關係，或者是該詞彙與其相關連的詞群之間的關係。因此在檢索上，可以分爲兩種方式：其一，直接找到欲展現的特定詞彙；其二，從其相關的詞彙找到所欲展現的特定詞彙。

在直接定位特定詞彙上，系統必須提供詞彙片段找尋，讓使用者能夠用片段的『字』找到欲查詢的特定詞彙。舉例來說：若使用者要查詢『非洲象』這個詞，可以透過輸入『非洲』或是『象』來進行找尋，並直接找到『非洲象』。

而從相關詞彙的檢索方式，就必須利用到系統展現的功能來尋找，可以透過相關連的詞彙，或是和該詞彙具備特定關係的詞彙，根據『關係』的線索來找尋。舉例來說：要查詢『酒精』，可以透過『乙醇』這個詞，經由關係的線索來尋找。

2. 展現的方式

索引典的展現關係到索引典建置和使用上的方便性，因此視覺化的作業在這個部分相當的重要。在這邊設定索引典是以下列樹狀而多層的方式呈現，在使用上，使用者可以針對特定的關係延伸，也可以就某特定詞彙橫向展開所有的關係詞彙，使用者不但可以藉由視覺化的介面來檢視詞彙，更可以同時透過樹狀結構的每一個節點，用『拖拉』的方式建立詞彙關係。主要展現的方式，如圖一所示的結構：



A：表示詞彙

關係 A：表示關係

圖一：索引典詞彙展現形式

伍、軟體工具的使用

本索引典建置工具程式，是藉由 Borland C++ Builder 5 來開發的介面，因此大部分的程式工具與套件，都是由該開發程式所提供。在程式的視窗化上有很多元件事必須的工具，由於程式的開發並沒有從 Microsoft 所提供的視窗元件函式庫（Win32 API）發展，在快速開發的考量下，選擇套裝軟體中已經模組化的元件使用。在本文就開發上所遇到的困難，說明其中較爲特殊的部分，以搭配在索引典建置程式使用上的方便性：

一、STL（Standard Template Library）

此爲 C++這種程式語言的一個函式庫，經由以往的程式設計師所設計的一些資料結構物件（Objects），讓設計者可以簡單的透過引用函式完成一些較爲複雜的資料結構建置，也可以省略一些繁複的除錯（Debug）步驟。在本系統中，設計上，需要有串鏈的資料結構（list）再加上，在 Borland C++ Builder 的開發環境，視窗元件所使用的字串（String）資料型態，跟在一般 C++程式所常態定義字串資料不完全相同（在 BCB 中用 AnsiString 作爲普遍使用的字串），導致再自行設計串鏈結構上遇上困難。因此採用 STL 的函式來引用 AnsiString 作爲串鏈，如此不但連結上較爲簡便，也省去爲串鏈資料除錯的動作。在系統中，不論是在展現上，或是邏輯檢定，都多次用到了 STL 的串鏈資料。

二、ADO（ActiveX Data Object）

爲微軟公司（Microsoft）開發出來新一代的資料庫連接中介程式（Middle Ware），它本身是一個可供其他程式或軟體使用來控制資料庫的程式，如此一來程式設計師便不需要同時具備各個資料庫架構程式（Database Management System Engine）的知識，進而撰寫特定資料庫的驅動程式，以運作資料庫。透過 ADO 程式設計師可以用較爲簡單的方式控制資料庫。ADO 提供了程式設計師對資料庫進行『連接』（Connection）、『查詢』（Query）、『新增』（Create）、『修改』（Update）、『刪除』（Delete）的功能，只需要透過 SQL 語言（下述）便能輕易的達到對資料庫的控制。

三、SQL 語言（Structured Query Language）

這是現今使用最爲重要的一種關連性資料操作語言（Relational Data Manipulation Language），爲美國國家標準局（American National Standards Institute）選定作爲操控關連式資料庫的語言（Relational Database），包括 DB2、SQL/DS、ORACLE、INGRES、SYBASE、SQL Server、dBASE、MS Access 等眾多資料庫軟體都選擇以 SQL 語言作爲資料庫操控語言。基於普遍性和方便性的考量，索引典建置系統，選擇以 SQL 作爲資料操控語言，對索引典檔案進行查詢、新增、修改、刪除等動作。

四、XML（Extensible Markup Language）

XML 是一種開放的，以文字爲基礎的標籤語言，它可以提供結構的以及語

意相關的資訊。這些詮釋資料（metadata）提供附加的意義和目錄給使用那些資料的應用程式，在它跨平台、跨語言的優勢下，使得 XML 成爲資料傳輸的新標準。在索引典建置系統中，利用 XML 的優點，作爲資料匯出的格式，以利使用者在別的領域可以使用所建置的索引典。

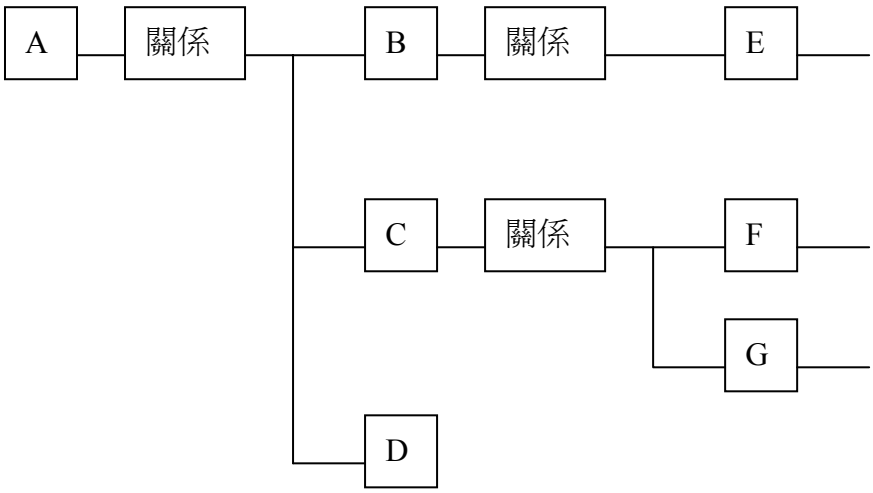
五、Data I/O（Input & Output）

在索引典建置軟體中，我們使用 C++在 IStream、Ostream 中所定義的開檔、讀檔、寫檔的機制，寫了複製檔案內容的函式（檔名和目錄可以自行指定）。在這樣的基礎上，對於索引典檔案的編輯控管能夠更爲接近使用者的想法，也就是在使用者選擇『存檔』之前，都有權力放棄之前編輯的內容，而索引典也能夠回復到剛開啓檔案時的狀態。

六、Validation Methodology

詞彙關係的驗證方式，所依循的邏輯主要來自於不同種類詞彙關係的定義，可以分爲階層性的驗證、對稱性的驗證、權威性驗證三個部分：

1. 階層性的驗證：驗證關係包括：BT、NT、UF、USE



圖二：詞彙關係之驗證

在 A 關係（例如 BT 或 NT）中，A 不能在出現在他以下的詞群中，也就是說 B、C、D、E、F、G…中不能再出現 A 詞彙。

2. 對稱性的驗證：驗證關係包括：（RT、RT）、（BT、NT）、（UF、USE）

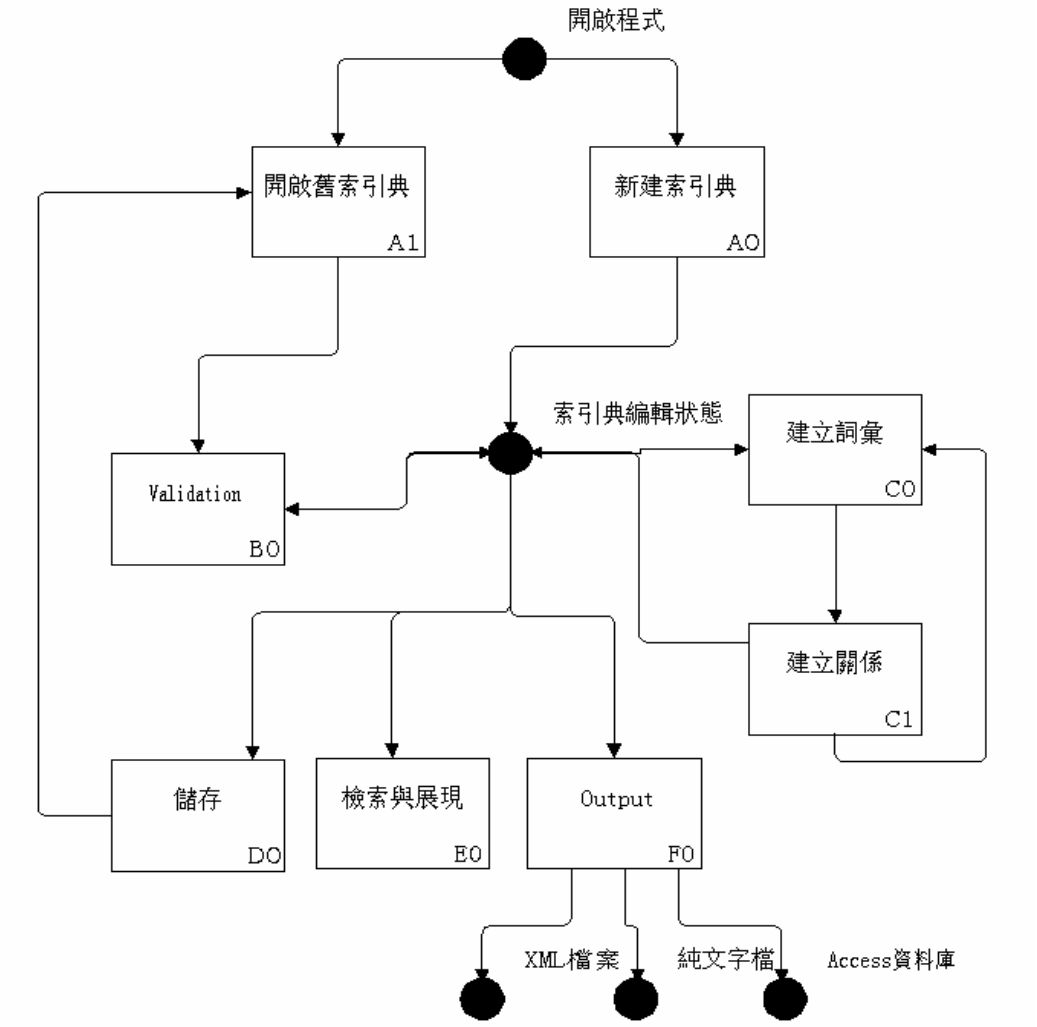
在兩兩一組的對稱性驗證中必須存在以下的組合：關係甲和關係乙爲對稱的一對關係，詞彙 A、B 存在：A-甲-B，則必同時存在 B-乙-A 的紀錄。

3. 權威性的驗證：驗證關係包括：UF、USE

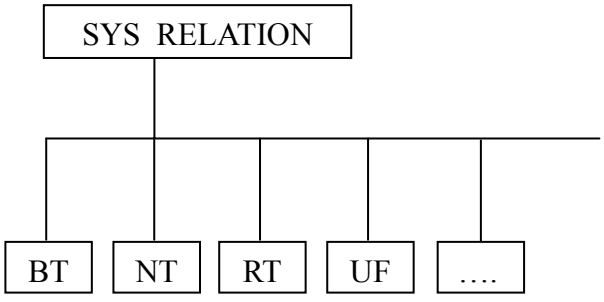
具備有一對敘述詞和款目詞的關係，而一個款目詞最多只能擁有一個敘述詞。

陸、軟體運作與使用方法

本節說明索引典建置軟體的運作和使用畫面，整個系統將同時就兩個方面分開說明，分別是使用者對系統介面所下的指令和系統程式對索引典檔案所做的處理。圖三爲整個索引典建置工具程式的使用流程圖。由於本工具程式的各項功能畫面非常容易瞭解，本節僅展示各項功能畫面，而並不多加解釋。

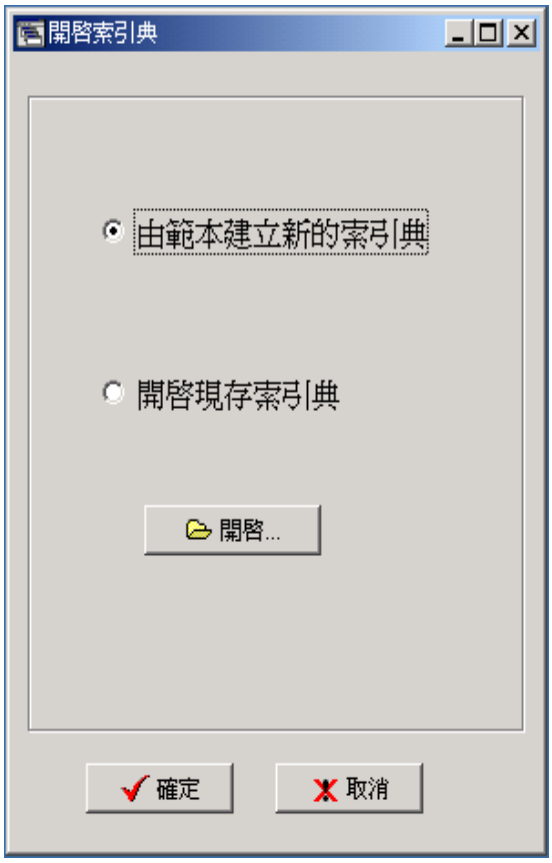


圖三：索引典建置工具程式使用流程圖

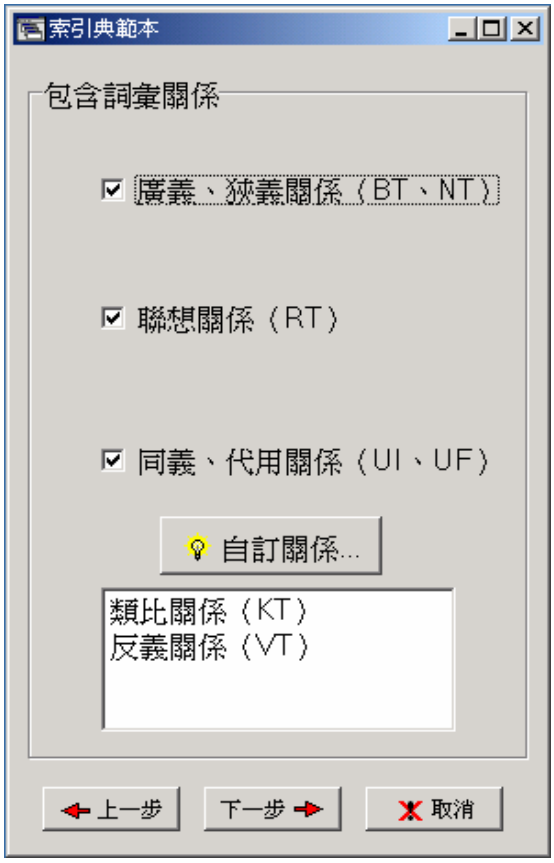


NUMBER	RELATION
1	BT
2	NT
3	RT
4	UF
…	…

圖四：資料庫詞彙關係架構圖

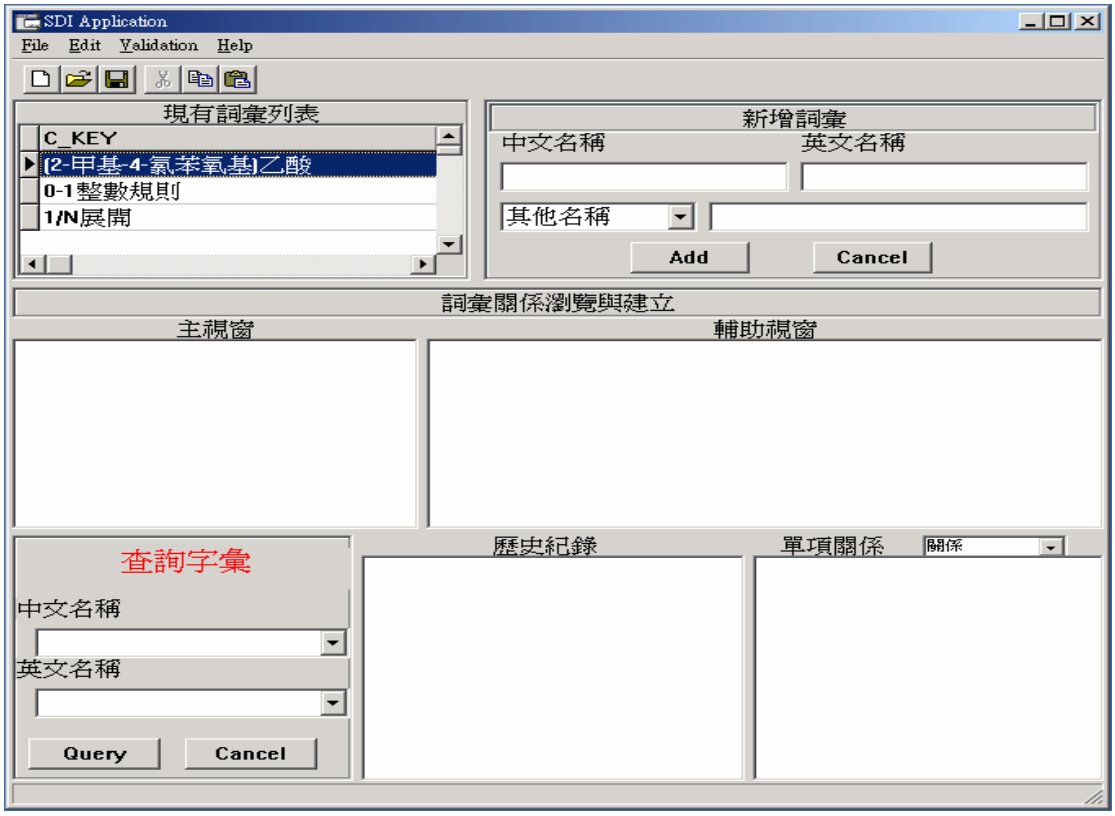


圖五：系統的啟始畫面

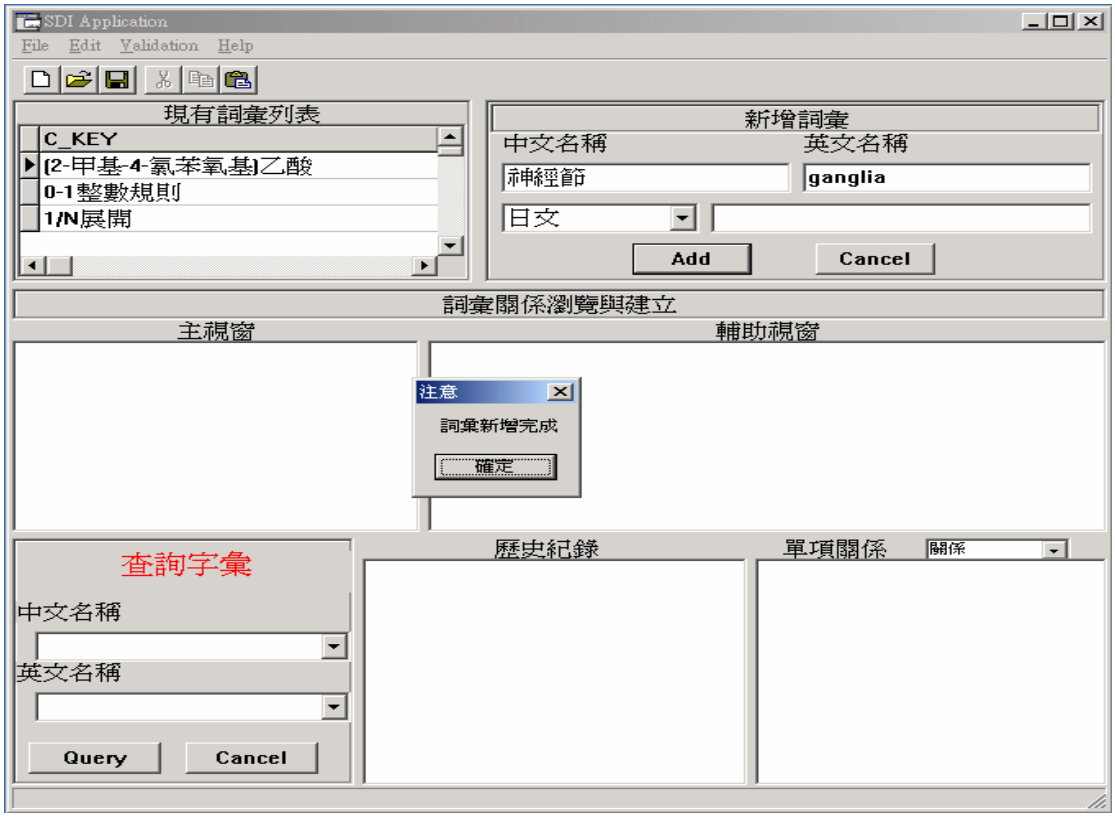


圖六：定義詞彙關係的畫面

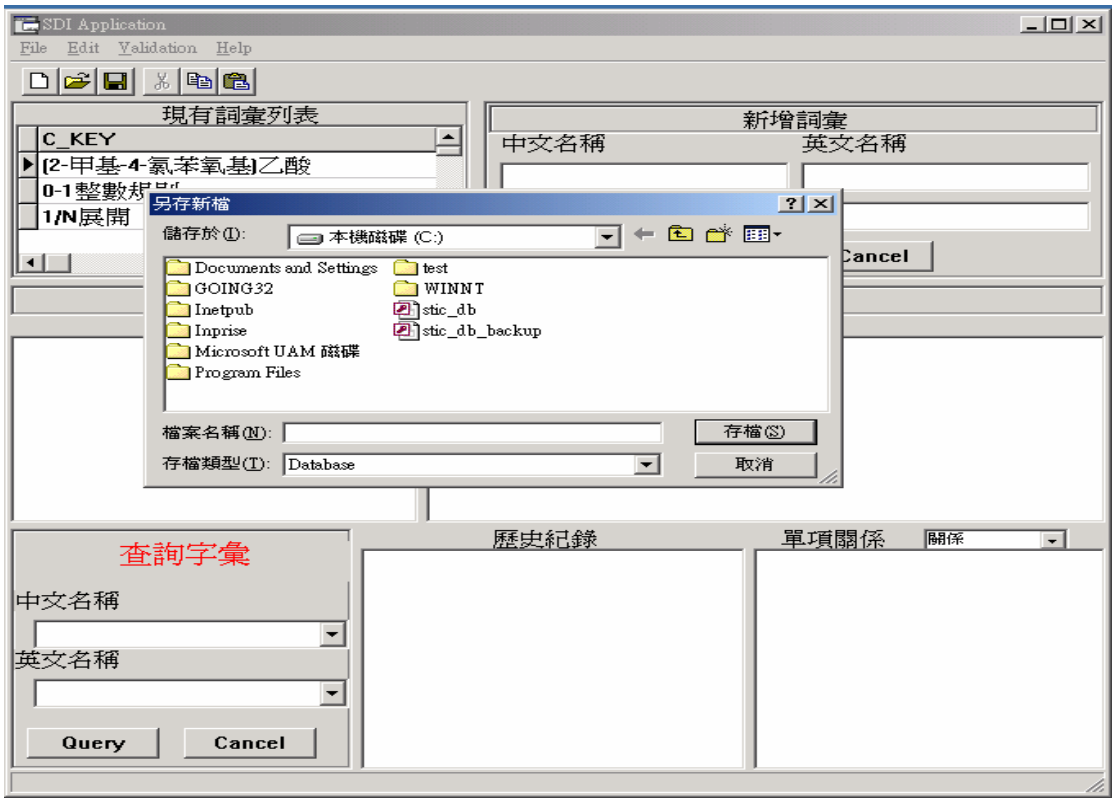
圖七：自訂詞彙關係（使用者自訂）



圖八：主功能畫面

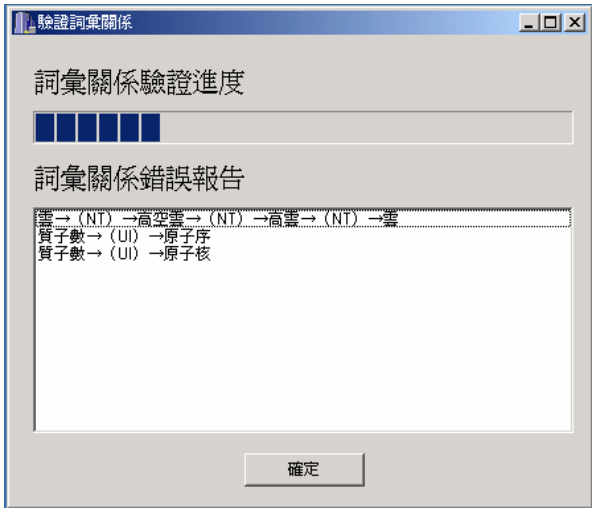


圖九：建立詞彙

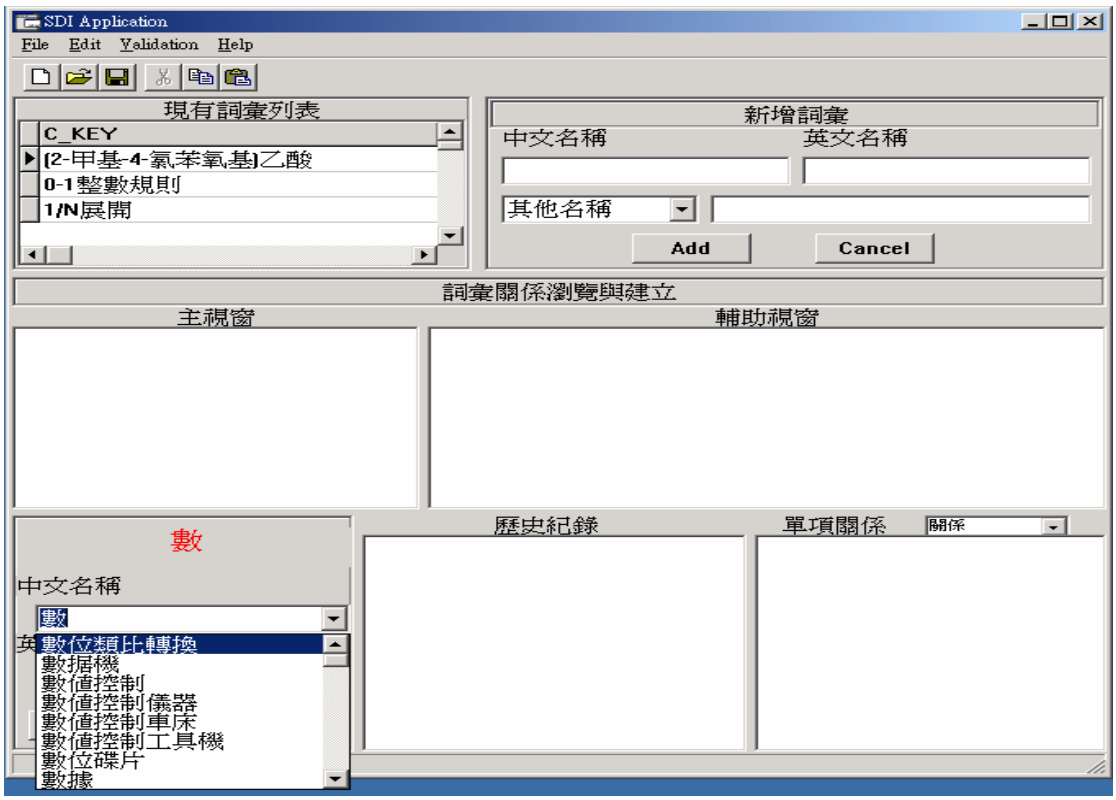


圖十：索引典儲存畫面

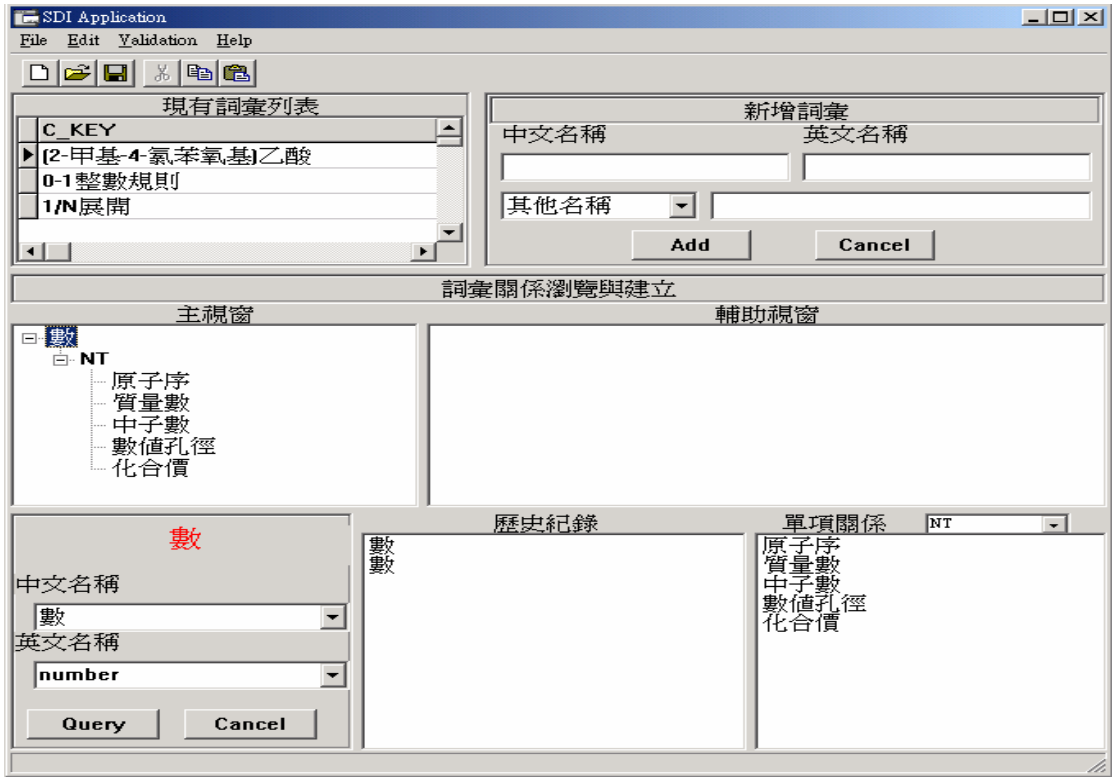
圖十一：驗證詞彙關係的邏輯



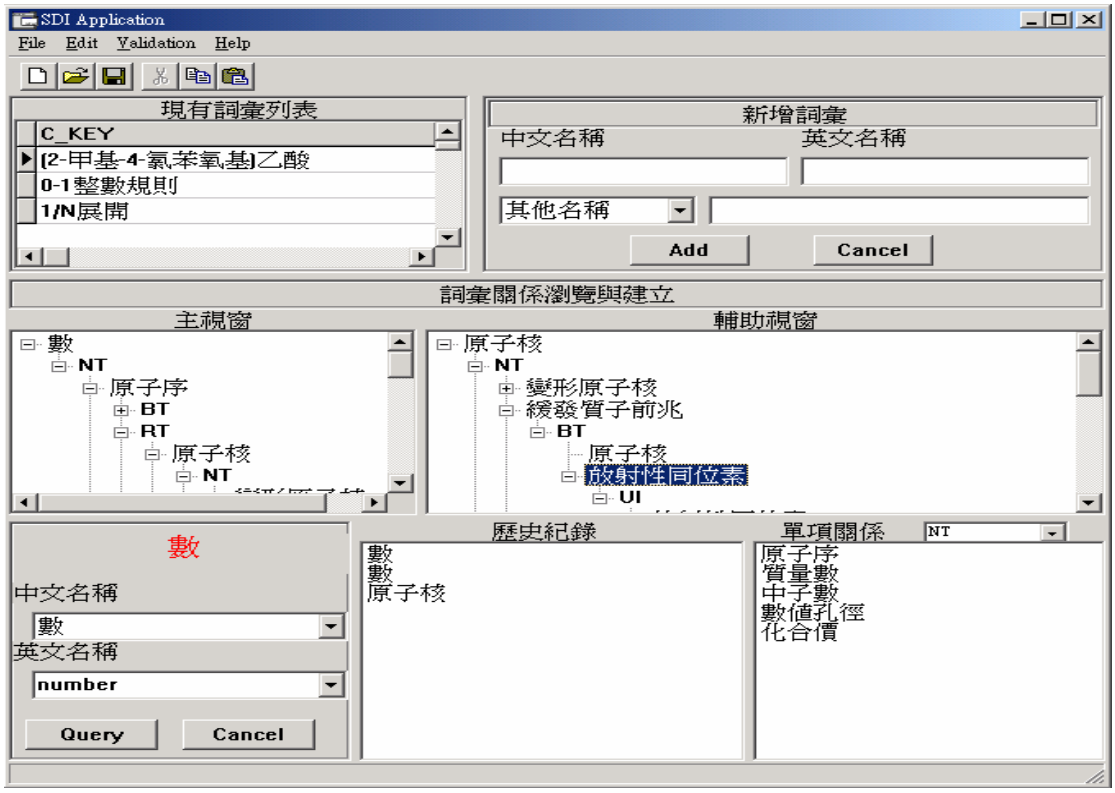
圖十二：驗證詞彙關係的畫面



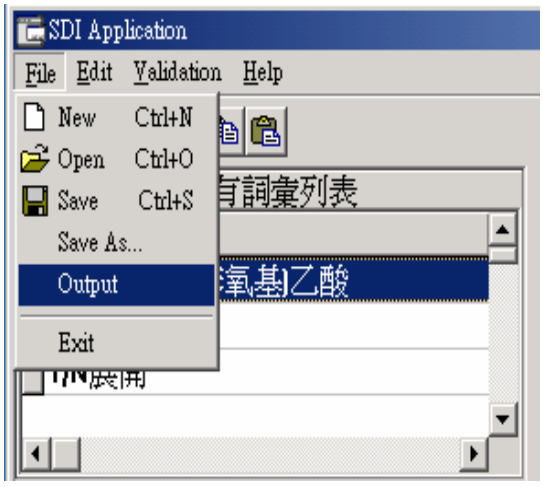
圖十三：檢索詞彙的畫面



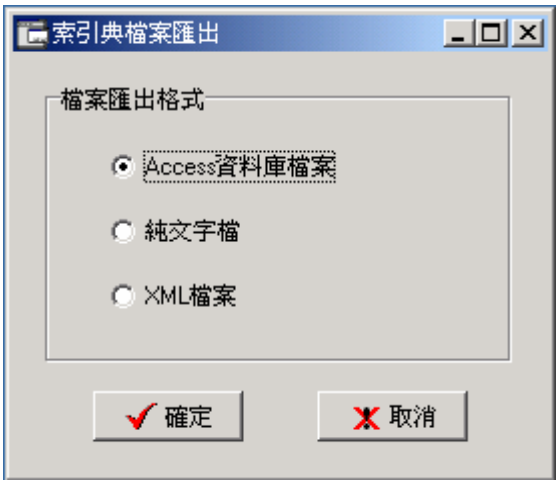
圖十四：檢索詞彙結果的畫面



圖十五：檢索詞彙結果的畫面之二



圖十六：匯出索引典



圖十七：匯出格式選擇

```
<THESURUS>
    <ENTRY>
        <TERM>
            <語系 A>...</語系 A>
            <語系 B>...</語系 B>
        </TERM>
        <BT>
            <entry>
                <語系 A>...</語系 A>
                <語系 B>...</語系 B>
            </entry>
            ...
        </BT>
        <NT>
            <entry>
                <語系 A>...</語系 A>
                <語系 B>...</語系 B>
            </entry>
            ...
        </NT>
        <其他關係>
        </其他關係>


    </ENTRY>
    ...
</THESAURUS>
```

圖十八：匯出 XML 檔

```
索引典檔案      匯出日期:2001/10/23
檔案名稱:stam.mdb
//-----------------------------------------------------------//
中文:詞彙 1      英文:Term A      其他語言…
    Relation:   BT:
                    中文:詞彙 2      英文:Term B
                    中文:詞彙 3      英文:Term C
                    …
    Relation:   NT:
                    中文:詞彙 4      英文:Term D
                    …
    Relation:   其他關係
    …
中文:詞彙 5      英文:Term E      其他語言…
…
其他詞彙
```

圖十九：匯出文字檔

柒、結論

索引詞彙事實上扮演二個角色，其一是讓資料著錄者選擇適當的詞彙，以表達所處理的文獻資料的主題；其二是讓資料檢索者下達適當的詞彙，以檢索經適度處理的文獻資料。索引典或是標題表承載者控制詞彙，可以精確地描述文獻資料，如果數位圖書館／博物館或是資訊檢索系統，除了使用自由詞彙之外，還能夠使用控制詞彙，則使用者能夠檢索出更符合其資訊需求的文獻。如何有效地讓資訊系統的使用者，瞭解系統知識架構，妥善地使用檢索詞彙，讓索引典適度扮演前述的二個角色，是系統建置者必須考慮的重要議題。

雖然索引典的重要性，已獲得普遍性的認同，亦有許多學者專家發表許多研究論文，然而，目前整合索引典的資訊檢索系統或數位圖書館／博物館並不多，原因是建構索引典是一件大工程，耗費人力經費甚多。本文提出一個索引典建置工具程式，可以協助索引典建置者輸入詞彙、編輯詞彙、建構詞彙關係、自訂詞彙關係、瀏覽詞彙關係、檢索詞彙、以及匯出索引典。除此之外，更提供其他索引典建置工具少見的詞彙關係的邏輯驗證，以及多語系的考量。目前本工具程式仍在持續發展中，未來將加入更多的功能。

致謝

作者感謝研究助理林志岳先生與吳俊輝先生的辛勤工作與細心討論，才能有本索引典建置工具程式的初步成果。

附註：

註 1： 陳光華，「新資訊時代的啓發性資訊服務」，21 世紀資訊科學與技術的展望學術研討會論文集，世新大學，桃園（民國 87 年），頁 195-208。

註 2： Liptez, B.-A. “Catalog Use in a Large Research Library,” Library Quarterly 42 (1972): 129-139.

註 3： 黃慕萱，「線上檢索指令分析 -- 以國立臺灣大學之終端使用者爲例」，國立台灣大學圖書館學刊第 11 期（民國 85 年 12 月），頁 47-62。

註 4： 陳光華，「數位圖書館中權威控制系統的設計」，政治大學圖書與資訊學刊第 34 期（民國 89 年 8 月），頁 51-71。

註 5： Smith, Alastair. “Search features of digital libraries.” Information Research 5(3) (2002). URL: http://www.shef.ac.uk/~is/publications/infres/paper73.html (18 Nov, 2002).

註 6： 曾元顯，「共現索引典之自動建構、評估與應用」，國立台灣大學圖書資訊學系四十週年系慶學術研討會論文集，台北（民國 90 年），頁 87-106

註 7： MultiTes, URL: http://www.concentric.net/~Multites/ (18 Nov, 2002).

註 8： THESshow, URL: http://www.mu.niedersachsen.de/cds/etc-cds_neu/thes_slide_home.html (18 Nov, 2002).

註 9： Bridgewell, URL: http://www.bridgewell.com/ (18 Nov, 2002).

註 10： 李連輝，「索引典與索引方法」，圖書館學與資訊科學 3 卷 2 期（民國 66 年 10 月），頁 48。

美國資訊科學學會臺北分會編，索引典理論與實務（台北市：編者，民 83 年）。
張嘉彬，「索引典及其於資訊檢索上之探討」，書苑 36 期（民 87 年 4 月）：46~59。
莊雅蓁，「資訊檢索之索引典研究」，中國圖書館學會會報 63 期（民 88 年 12 月）：77~89。
黃惠株，「淺談索引典」，佛教圖書館館訊 5 期（民 85 年 3 月）：2~7。
黃慕萱，資訊檢索（台北市：臺灣學生，民 85 年）。
蔡明月，線上資訊檢索--理論與應用（台北市：臺灣學生，民 80 年），頁 166~169。
Chen, H. and D. T. Ng., "An Algorithmic Approach to Concept Exploration in a Large Knowledge Network (Automatic Thesaurus Consultation): Symbolic Branch-and-Bound vs. Connectionist Hopfield Net Activation," Journal of the American Society for Information Science 46:5 (1995): 348-369.
Crouch, Carolyn J., "Experiments in Automatic Statistical Thesaurus Construction," in Proceedings of the 15th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval, Copenhagen, Den, Jun 21-24 1992, 77-88.
Schutze, H. and Pedersen, J. O., "A Cooccurrence-Based Thesaurus and Two Applications to Information Retrieval," Information Processing & Management 33:3 (1997): 307-318.
Wan, T.-L. et al., "Experimetns with Automatic Indexing and a Relational thesaurus in a Chinese Information Retrieval System," Journal of the American Society for Information Science 48:2 (1997): 1086-1096.