索引典建置

高秋芳
國研院科技政策研究與資訊中心
98年8月31日

大 綱

- 農業科技索引典編製情形
  （農資中心）
- 科技索引典編製情形
  （科資中心，現稱科政中心）

農業科技索引典(AGRITHESAURUS)編製情形

（民國69年）

索引典定義之一

The vocabulary of a controlled indexing language, formally organized so that the *a priori* relationships between concepts (for example as “broader” and “narrower”) are made explicit.

(ISO 2788---1974)

索引典定義之二

A compilation of words and phrases showing synonymous, hierarchical, and other relationships and dependencies, the function of which is to provide a standardized vocabulary for information storage and retrieval.

(ANSI Z39.19---1974, 1st ed.)

索引典定義之三

A thesaurus may be defined either in terms of its function or its structure. In terms of function, a thesaurus is a terminological control device used in translating from the natural language of documents, indexers or users into a more constrained "system language" (documentation language, information language). In terms of structure, a thesaurus is a controlled and dynamic vocabulary of semantically and generically related terms which covers a specific domain of knowledge.

(UNISIST, UNESCO---1977, 1st ed.)

基本術語

concept
language
- natural language
- artificial language

word
phrase
term
terminology
vocabulary
knowledge

基本術語釋義(一)

- concept 概念

  反映事物特徵的思維單元

- Feature 特徵

  構成概念的任何特點、屬性或關係

- language 語言

  用為溝通的符號系統，通常由字彙與語法規則所組成

- natural language 自然語言

  自然演進所形成的語言

基本術語釋義(二)

- artificial language 人工語言

  根據事先設定的詞彙與其語法規則所編製的結構化語言

- word 字

  語言的最小單元，本身能表達一特定的意義，且可構成句子個別單元

- phrase 片語

  根據一定的語法規則組合起來並具有整體意義的一組詞

- term 用語

  表示一個概念的字或片語

基本術語釋義(三)

- terminology 術語

  某一專門領域的一組用語，其意義及用法已普遍為相關領域所接受

- vocabulary 字彙

  解釋字義或其翻譯名稱之字典，通常按一定順序編排

- knowledge 知識、學識、學問

  the fact or condition of knowing something with familiarity gained through experience or association

索引典必須具備的五個條件

- 索引典是一個集合，集合中的元素是關鍵詞
- 關鍵詞是代表資料內容或主題概念之字或詞
- 需將各關鍵詞依等同、層次及相關等關係組織起來
- 明白規定那些關鍵詞可以在資訊系統中使用，那些不可以，將關鍵詞加以控制的目的，在為文獻作者、索引人員及檢索人員等三方提供一種共同一致的系統語言
- 索引典的內容是隨著資訊系統之成長而成長的，因此索引典必定是動態的

索引典發展情形

目的---追求資訊檢索效率的提升

- 1947-1950　Mooers
- 1957　Brownson
- 1959　杜邦公司索引典
- 1961　Chemical Engineering Thesaurus
- 1960年代　NASA Thesaurus,…….
- 1970年代　歐洲各國發展索引典
- 1980年代　索引典已成為資訊檢索語言主流
- 1985　使用中之索引典約有600部

AGROVOC

- 1980年代早期由FAO & EU Commission 開發
- 多語索引典（英、法、西、中、阿等）
- 涵蓋領域：Agriculture、Forestry、Fisheries、Food及相關者(如Environment)
- 約36,000 terms，每三個月更新一次
- Term關係包括：USE(UF)、BT、NT、RT、SN
- Term 有分類
- 非商業用途可免費下載

CAB Thesaurus

- 由CABI所開發，since 1983
- 英文索引典
- 涵蓋領域：Pure and applied life sciences, technology and social sciences;農業領域包括Agriculture, Forestry, Horticulture, Soil science, Entomology, Mycology, Parasitology, Veterinary medicine, Nutrition, Rural studies
- 約59,000 terms，定期更新
- Term 有分類
- 免費提供非營利用途之學術研究機構

CAB Thesaurus之Term內容

- **BT:** Broader term. One level up from the main term
- **NT:** Narrower term. One level down from the main term
- **RT:** Related term
- **SN:** Scope note. Explanatory notes for the main term
- **HN:** History note. Previous usage notes for the main term
- **UF:** Use for, i.e. a non-preferred term of a preferred main term. In some cases, one of a pair of non-preferred terms (green)
- **AF:** The American non-preferred spelling of a preferred main term (green)
- **PT:** Preferred term. Present if the main term is non-preferred (red). May be the preferred British form of a main term with American spelling. May map to two or more preferred terms, using 'AND' as the separator
- **SE:** 'See' term. The preferred term(s) if the main term is non-preferred (red) and if there are several alternatives. 'OR' is used as the separator
- **CA:** Chemical Abstracts Registry Number. Present if the main term is a chemical
- **EC:** International Union of Biochemistry Enzyme Commission notation for enzymes
- **TN:** Tree number in the classified section
- **LN:** Scientific name for a plant or crop with a common name
- **CN:** Common name equivalent of the scientific name for a common

NAL Agricultural Thesaurus

- 由USDA國家農業圖書館所開發
- 2002, 第一版
- 英文與西班牙文之雙語索引典
- 涵蓋領域：農業，廣義
- 約70,000 terms, 每年更新
- Term關係包括：USE(UF)、BT、NT、RT、SN、Definitions
- Term 有分類
- 可自行下載

農業科技索引典

- 民國69年3月取得聯合國糧農組織的索引典---AGROVOC
- 民國71年9月完成農業科技索引典系統
- 製作與更新皆電腦化
- 應用於農業科技人才、農業科技研究及發展計畫、農業科技文獻等三個資料庫之索引與檢索作業
- 民國77年7月出版農業科技索引典紙本

農業科技索引典

- 主題範圍：農藝、園藝、植物病害與病理學、植物蟲害與害蟲學、土壤與肥料、農業土木與水利、農業機械、水土保持、林業、漁業與水產養殖、畜牧、獸醫、食品科學與技術、農業經濟、農業推廣、生物統計與試驗設計、農業污染與防治等17項主要主題，另包括5項次要主題
- 含中文詞20,115個(系統關鍵詞11,645個，非系統關鍵詞8,470個)、英文詞20,521個(系統關鍵詞11,645個，非系統關鍵詞8,876個)
- 中文詞長度不超過15個字，英文詞長度不超過36個字母

農業科技索引典

- 索引典架構於檢索時之應用
  - 含所有UF: default
  - 含所有層次之NT: <
  - 含所有RT: %
- 索引作業與索引典維護之結合
  - TEMP TERM 之使用

索引典的基礎理論

- 概念邏輯理論
- 語言基礎知識
- 知識分類理論

概念邏輯理論

1. 概念結構

–特徵表理論　C=R (X,Y,....)

C:概念

X, Y, . . :為一類個體具有的共同的定義性特徵

R:整合這些特徵的規則

例： 鳥=合取(羽毛、動物)

－原型理論

主要以原型，即它的最佳實例表現出來

例：鳥

鴿>企鵝

Rosch：最佳實例＋範疇成員代表性的程度

2. 概念的內涵與外延

- 內涵：反映在概念中的對象的本質屬性，即概念的含義，亦是概念的質。

  例：鳥：有羽毛、動物、卵生....

- 外延：具有概念所反映的本質屬性的一切事物，即概念的適用範圍，亦即概念的量。

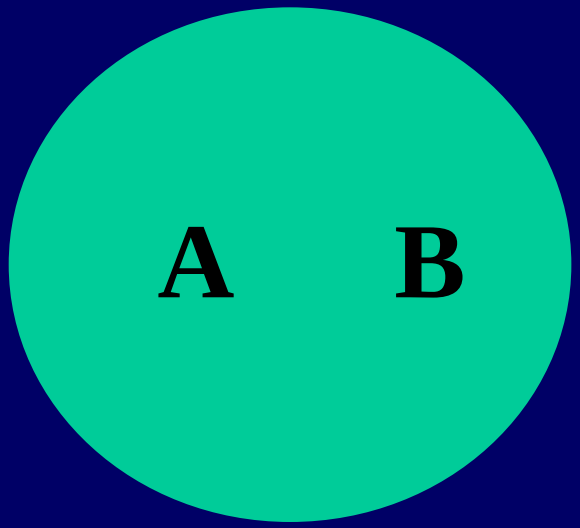
  例：鳥：雞、鴿、鷹、麻雀...

3. 概念的擴大與縮小

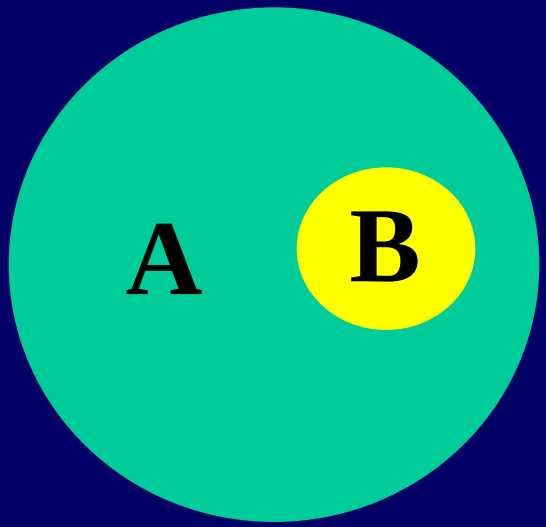
- 擴大：經由減少概念的內涵以擴大概念的外延。從特殊過渡到一般。
- 縮小：經由增加概念的內涵以縮小概念的外延。從一般過渡到特殊。

4. 概念之間的關係

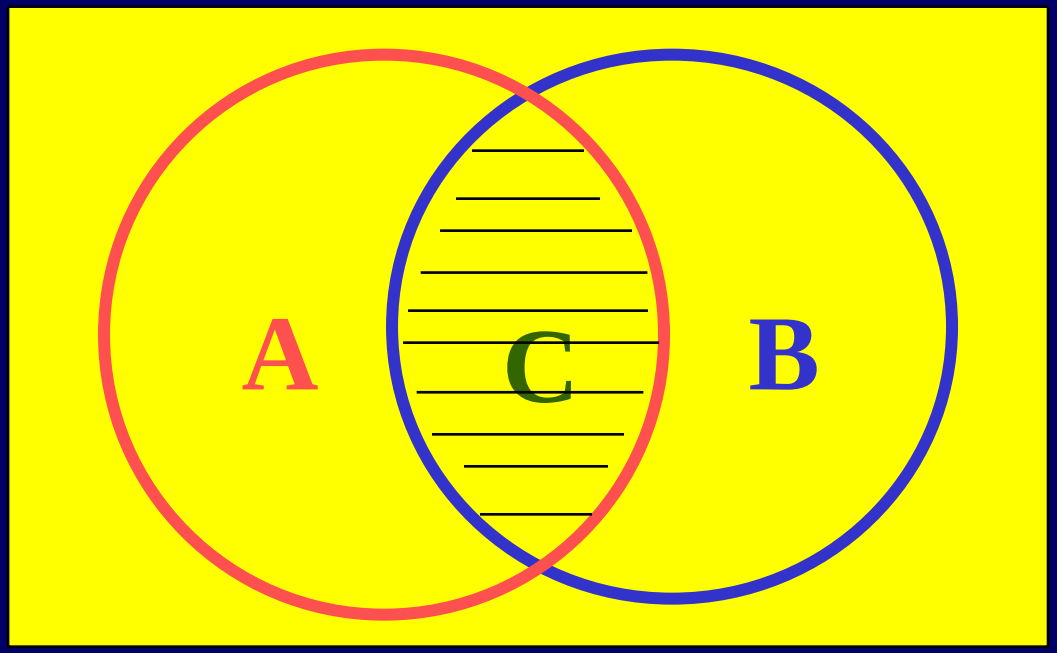
–同一關係：兩個概念的外延完全重合。如蘇軾與蘇東坡。



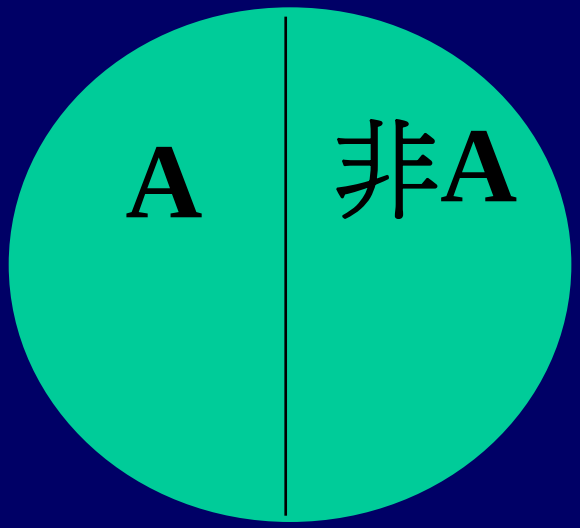
– 屬種關係：一個概念的部分外延與另一概念的全部外延重合。如圖書館與公共圖書館。



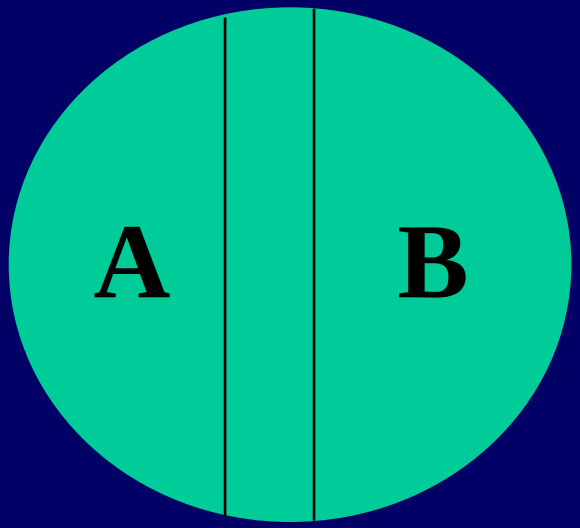
－交叉關係：一個概念的部分外延與另一概念的部分外延重合。如有聲出版品與CD。



－矛盾關係：兩個概念的外延完全相互排斥且外延的和等於其屬概念的外延。如金屬材料與非金屬材料。



－對立關係：兩個具有全異關係的概念同時包含於一個屬概念中，且它們的外延之和小於其屬概念的外延。如貧與富。



概念邏輯理論在索引典中的應用

- 對概念進行劃分(縮小)或概括(擴大)，可區分各種事物。
- 利用概念的劃分與概括過程中所形成的概念等級關係和並列關係，可建立索引典的等級關係。
- 利用具有交叉關係的兩個概念外延的重合部分可以形成一個新概念，反之亦然。

語言基礎知識

- 語法單位
  - 語素：語言中最小的音義結合體。如蠟、燭各是一個語素。
  - 詞：最小的能夠獨立運用的語言單位。
  - 短語：詞和詞的語法組合。

－句子：短語或詞構成，能夠表達一個相對完整意思的語言單位。

- 詞義：詞是代表事物或現象的一種符號，詞義是代表詞在人腦中所呈現的概念。

語言基礎知識在索引典中的應用

- 詞性：以名詞為主
- 結構單元：分為單詞和複合詞
- 關聯詞：要儘量擴大其比例
- 同義詞：要儘量擴大非描述詞比例
- 概念與詞：儘量達成一一對應
- 動態的語言--->維修詞彙--->合時的索引典

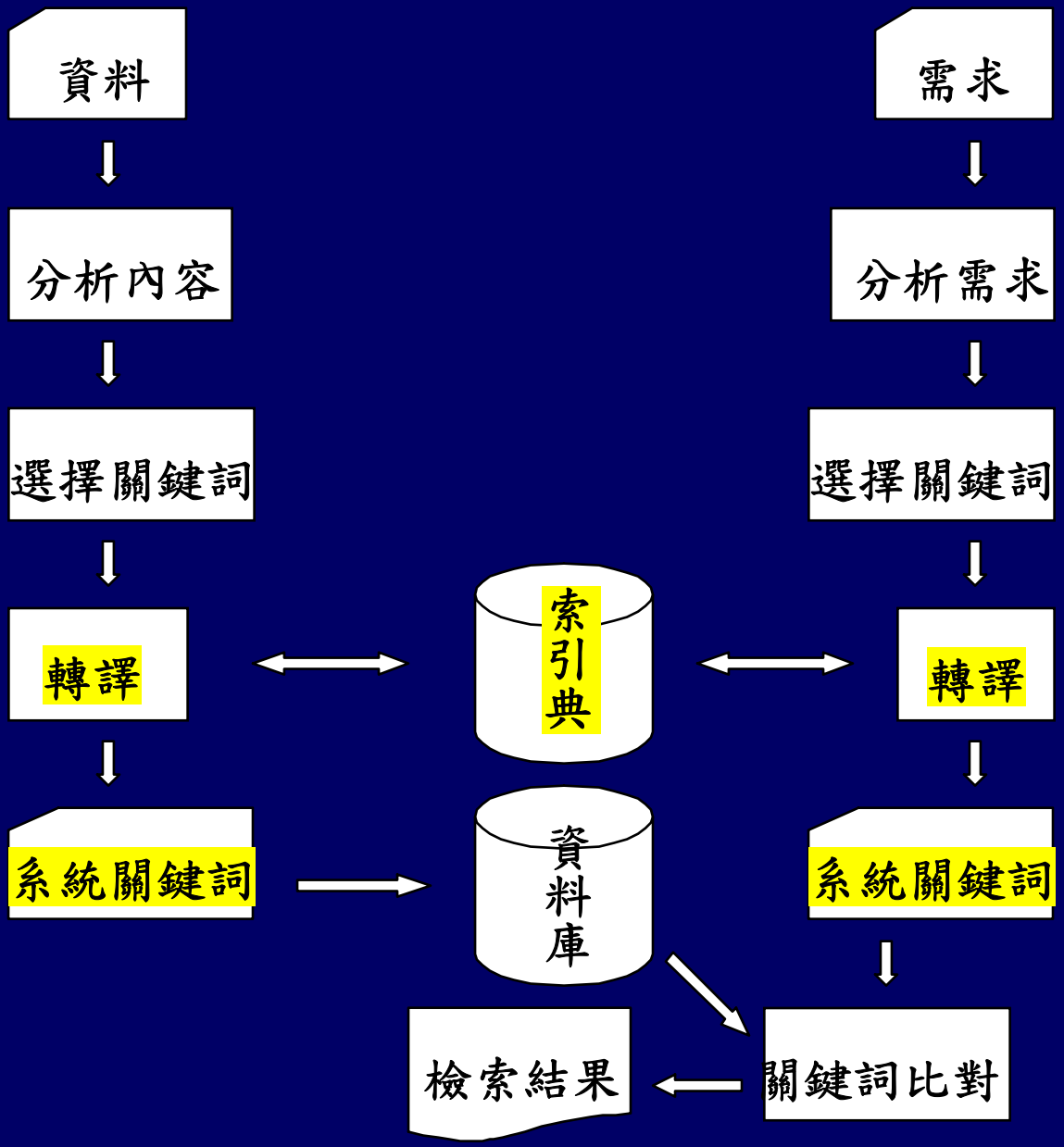
知識分類理論

- 分類：人類思維活動中的一種本能，是人類認識事物的一種重要手段。

知識分類理論在索引典中的應用

- 確定族首詞
- 組織和展示詞族等級
- 劃分和確立範疇類目
- 決定範疇類目的排列順序

索引典功能


索引作業
檢索作業
資料
需求
分析內容
分析需求
選擇關鍵詞
選擇關鍵詞
轉譯
索引典
轉譯
系統關鍵詞
資料庫
系統關鍵詞
檢索結果
關鍵詞比對

索引作業方式

- 控制式詞彙索引(Controlled Vocabulary Indexing)

  索引工具：分類表、主題標目表、索引典

- 非控制式詞彙索引(Non-controlled Vocabulary Indexing)

  －自然語言索引

為何要控制詞彙(一)

用自然語言於檢索系統的缺點

- 同義現象：一義多詞
- 多義現象：一詞多義
- 詞的模糊性與不確定性：借喻、轉義等
- 詞間關係不明晰

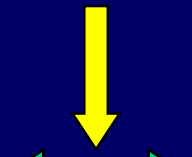
為何要控制詞彙(二)

目的：統一使用者的語言和索引人員的語言

詞彙控制：一種把自然語言加工成資訊檢索語言的過程

過程一：對自然語言的語詞進行壓縮、優選和規範化處理

過程二：對自然語言進行結構化處理



概念體系 ⟷ 術語體系

為何要控制詞彙(三)

詞彙控制原則：

1. 穩定性
2. 正確性
3. 單義性
4. 系統性
5. 簡明性
6. 成族性
7. 兼容性

為何要控制詞彙(四)

詞彙控制範圍：

1. 詞量控制
2. 詞類控制
3. 詞形控制
4. 詞義控制
5. 詞間關係控制
6. 專指度控制
7. 先組度控制

需要熟悉的知識

- 索引作業：分類、賦予關鍵詞
- 資訊檢索：檢索策略、資料庫
- 索引典製作
- 術語學
- 領域知識

科技索引典
(STIC sci-tech thesaurus)
編製情形
(民國77年)

科技索引典

- 科技專有名詞之中文譯名採用教育部公布者，若無，則參考其他書籍，最後經學者專家審查
- 參考日本JICST索引典、美國MeSH及TEST、英國ROOT、聯合國科教文組織SPINE等
- 系統於民國81年完成，紙本於同年6月出版

科技索引典

- 主題範圍：理、工、醫、農等所有科技領域，共分為21大類、182中類
- 含中文詞49,270個(描述語37,701個，非描述語11,569個)、英文詞49,447個(描述語37,701個，非描述語11,746個)
- 中文詞長度不超過26個字，英文詞長度不超過70個字母

索引典設計前考慮因素

1. 資訊系統方面的考量
   (1)學科範圍
   (2)資料類型
   (3)資料量
   (4)系統使用人數及頻率
   (5)檢索提問(Query)的類型
   (6)資源：財力、人員、設備

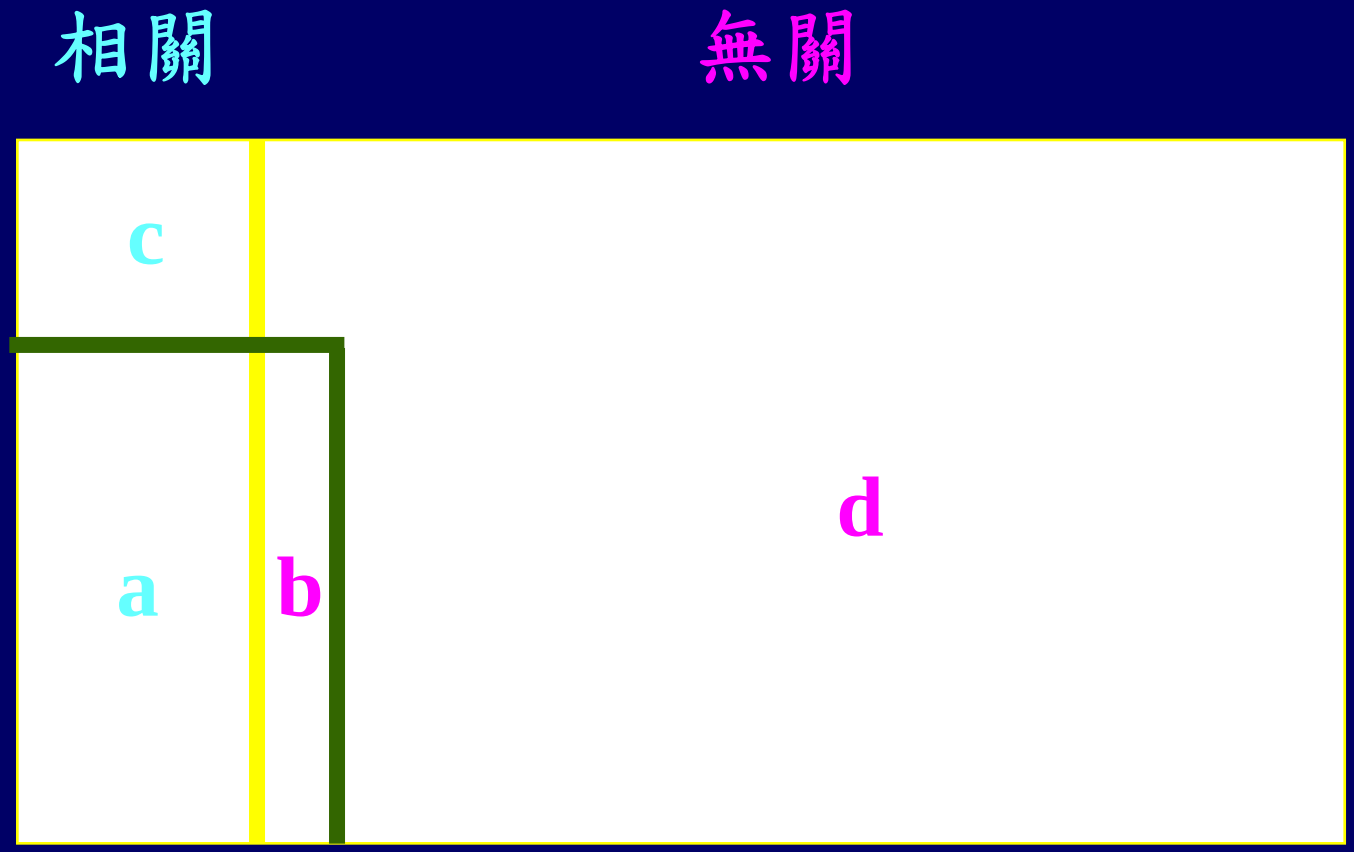
索引典設計前考慮因素

2. 資訊檢索效率的考量

(1)查全率與查準率

(2)詳盡度與專指度

查全率=a/(a+c)　　查準率=a/(a+b)



詳盡度與專指度：

詳盡度：對文獻中有參考價值的各種主題概念，索引人員能夠從不同角度加以判斷並把它們一一索引出來的程度

專指度：所用的索引詞與文獻中主題概念的切合程度

索引典設計前考慮因素

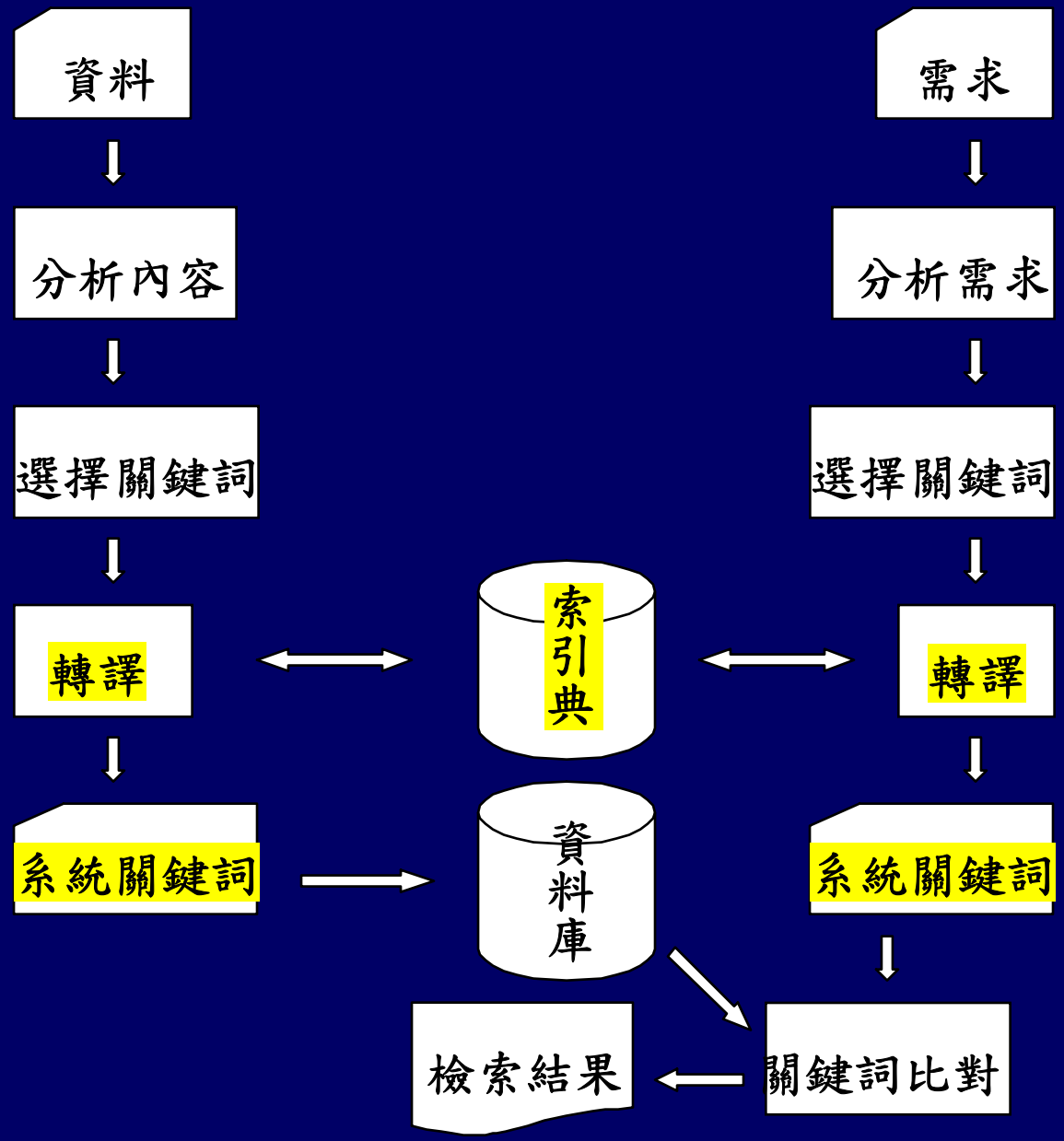
3. 索引典方面的考量

(1)採用已有者

(2)自建

經 驗 交 流

索引典功能



不能忘記索引典的功能

索引典必須具備的五個條件

不能忘記索引典的特性

- 索引典是一個集合，集合中的元素是關鍵詞
- 關鍵詞是代表資料內容或主題概念之字或詞
- 需將各關鍵詞依等同、層次及相關等關係組織起來
- 明白規定那些關鍵詞可以在資訊系統中使用，那些不可以，將關鍵詞加以控制的目的，在為文獻作者、索引人員及檢索人員等三方提供一種共同一致的系統語言
- 索引典的內容是隨著資訊系統之成長而成長的，因此索引典必定是動態的

索引典是…

不能忘記索引典的特性

- 索引典為某一特定資訊系統之索引與檢索的工具
- 索引典與資訊系統所收錄內容極有關係
- 索引典內的term只能算near natural language
- 索引典內term的specificity與文獻關鍵詞的specificity有落差

詞彙控制範圍

目前分散式作業如何管理

1. 詞量控制
2. 詞類控制
3. 詞形控制
4. 詞義控制
5. 間關係控制
6. 專指度控制
7. 先組度控制

索引典的管理與維護

日後內容維護之考量

1. 對詞及其相互關係進行增刪與修改
   - 增詞
   - 刪詞
2. 建立使用情況資料，如
   - 索引詞頻率統計
   - 檢索詞頻率統計
3. 需要專業人力且耗時
4. 新詞之Time lag問題

選詞原則

如何做好選詞工作

文獻索引與檢索的
實際需要是總原則

1. 要根據主題範圍選詞
2. 要根據索引典的實際用途選詞
3. 要注意基本詞的選擇
4. 所選的詞要概念明確
5. 要根據索引頻率選詞
6. 要根據檢索頻率選詞

複合詞的選用

複合詞訂定的困擾

1. 使用頻率高的複合詞：如「資訊服務」
2. 專有名詞：如「牛頓定律」
3. 組合檢索時發生多義者：如「雷達偵察」、「偵察雷達」
4. 分解後失義的複合詞：如「酒瓶椰子」

*判定原則：「核心成分」與「限定成分」，如「水泥橋」---不分解、「鳥類遷移」---分解

詞間關係

詞間關係的困擾

詞間關係：

1. 等同關係(同義、類同義)
2. 層次關係(上位、下位)
3. 相關關係

詞間關係的作用：

1. 詞義明確
2. 索引與檢索的用法一致
3. 利於擴檢與縮檢

同義關係的處理

1. 簡稱與全稱：一般以全稱為描述詞
2. 舊稱與新稱：一般以新稱為描述詞
3. 俗稱與學科用語：視使用者的習慣
4. 外來詞及其翻譯：視通用情況
5. 地名、國名、機構名、人名：採通用者或以通稱的詞替代

類同義關係的處理

1. 近義詞之間：一般以科學、通用者為描述詞
2. 正義詞與反義詞之間：一般以正義詞為描述詞
3. 肯定詞與否定詞之間：一般以肯定詞為描述詞
4. 專指與泛指的詞之間

層次關係的處理

處理上位概念與下位概念的關係，
分四種類型：

1. 屬種關係：類稱詞與成員詞，鳥類與鸚鵡
2. 包含關係：集合概念與單獨概念，河流與大甲溪
3. 整部關係：生物系統及其器官、地理與行政區域、學科及其分支、組織及其分支
4. 多層次關係：一個概念同時屬於幾個詞族

相關關係的處理

1. 含義有部分重疊的詞：ship與boat
2. 學科/研究領域/研究對象
3. 過程/工具
4. 事物/性質或事物/動作
5. 原因/結果或行為/結果或行為/受體
6. 概念/計量單位
7. 學科、學說、社會團體/人物

……

以Fuzzy Lyapunov function為例

- Fuzzy Lyapunov function是descriptor？
- 與Lyapunov function的關係？
- 與Liapunov function的關係？
- BT/NT關係，還是RT關係？
- 正確作法？

請討論