DOI: 10.6245/JLIS.201804_44(1).0001

開放取用系統與商業資料庫收錄資料之正確性與重複性

蔡明月
國立政治大學圖書資訊與檔案學研究所教授
E-mail: mytsay@nccu.edu.tw

吳岱欒
國立政治大學圖書資訊與檔案學研究所碩士生
E-mail: 101155002@nccu.edu.tw

關鍵詞：學術評鑑；資料庫正確性；引文資料庫；開放取用引文系統；資料庫重複性；學術傳播

【摘要】

本研究旨在比較商業與開放取用兩種學術傳播工具，透過實際操作商業引文資料庫 Web of Science 與 Scopus 與搜尋引擎 Microsoft Academic，匯集式機構典藏系統 OpenDOAR 和物理學專科特性的 Astrophysics Data System 開放取用引文系統，分析其收錄書目資料的正確性與重複性。本研究以諾貝爾物理學 2001 年至 2013 年得獎者之著作為研究樣本，進行書目資料剖析，包括排序、比對、刪除、聚集與統計，並相互交叉比對分析結果，比較各商業資料庫與開放取用系統之優劣，並根據分析結果提出建議。本研究結果期望能提供圖書館選擇引文索引資料庫與建置機構典藏系統，或引文資料庫與系統未來發展之參考。此外，亦可作為研究人員進行學術傳播與學術單位進行學術評鑑採用之指標與工具之建議。

前言

Van Damme（2001）曾提出，學術研究成果足以作為反映國家競爭力的一種指標。然而，如何評估研究成果的績效，則成為學者與研究機構所思考的問題。為瞭解學者的學術生產力與影響力，必須透過書目計量的方式進行分析研究。以諾貝爾獎得主預測為例，Thomson Reuters 透過「Web of Science」引文資料庫，自 2002 年起每年研究並分析科學研究成果的「被引用」情況，並以此為依據而辨識出化學、物理學、心理學或醫學以及經濟學等各學科領域中最具影響力的研究者，作為 Thomson Reuters 經典引文獎（Citation Laureates）」得主。自

2002 年以來，該經典引文獎已準確預測了 27 位諾貝爾獎得主，由此可見以引文分析方式選出的經典引文獎得主，與諾貝爾獎得主有一定的相關性。因此，許多人將經典引文獎作為諾貝爾獎得主的先行指標（Science Watch, 2013）。

隨著學術出版從紙本到數位的多樣化，加上網際網路的快速發展，使得網路資源急遽成長，進而促使學術傳播管道多元化，不僅改變了學術傳播環境，更進一步推動開放取用運動。學術研究文獻在網路世界大量湧現且快速流動，此種現象，也引發傳統書目計量學研究的擴展和變革，成為網路計量學（Webometrics）的研究。由於網路計量學的研究數據來自於網路世界，無論是網站、搜尋引擎、商業資料庫或開放取用系統等，皆可以是網路計量學的研究對象或研究工具。網路計量學如何精確嚴謹的分析、評價作者與學術文獻為一大課題；而作為網路計量學重要研究基礎的搜尋引擎與開放取用系統，其建置的目的、所具備的系統功能和收錄範圍各有差異。因此，在網路計量學研究中，若欲以搜尋引擎和開放取用系統做為研究工具，則必須先了解彼此的差異，因而引發本研究之研究動機。

近年來，在學術界與圖書館界共同推動「開放取用」的環境下，由學者、學術機構或圖書館等非營利團體所建置的開放取用系統與機構典藏系統，也可能具有引文索引的功能，例如：電腦科學領域的 CiteSeer，其對線上免費學術出版物建立引文索引，其他開放系統尚有典藏物理學、數學、電腦科學和生物計量學等領域研究文獻的 arXiv.org（Thelwall, 2008）。

對學者而言，Scopus 及 Web of Science 等商業資料庫是研究不可或缺的工具，但圖書館必須花費昂貴的代價方能購買資料庫；而免費的開放取用系統是圖書館在商業資料庫之外，可提供的免費研究資源。另外，相對於商業資料庫和開放取用系統，搜尋引擎亦是許多使用者查找資源之重要工具。在商業資料庫、搜尋引擎和開放取用系統之間，從事網路計量研究的學者該如何選擇或取捨？搜尋引擎、社群媒體和各種開放取用系統對各學科領域學者與圖書館而言，彼此差異為何？系統所收錄資源的正確性與重複性等不同面向上，各有何優缺點？

本研究依據以上的背景及動機，以網路計量之研究方法比較商業資料庫和開放取用系統收錄資料之書目資料正確性與重複性。Web of Science 及 Scopus 為本研究所選擇之商業資料庫，開放取用系統則有搜尋引擎 Microsoft Academic，以及屬於匯集式機構典藏系統的 OpenDOAR 和物理學專科特性的 Astrophysics Data System。本研究期望達成以下目的：

一、以諾貝爾物理學獎得主發表之學術文獻作為研究樣本，針對商業資料庫（Web of Science 及 Scopus）與開放取用系統之搜尋引擎 Microsoft Academic；匯集式機構典藏系統 OpenDOAR；學科性開放取用系統 Astrophysics Data System 分析其收錄書目資料的正確性與重複性。

二、綜合前述分析結果，比較各商業資料庫與開放取用系統之優劣，並根據分析結果提出建

議，以作為開放取用系統與圖書館服務改善之依據。此外亦可提供網路計量學研究者、物理學者與研究機構進行學術研究及研究成果評鑑之參考。

Poyer（1984）認為當相同的期刊文章被二個或二個以上的資料庫收錄並製作成索引或摘要，稱為文獻重複性。除了文獻被資料庫收錄之情形外，相同期刊被二種或二種以上資料庫同時收錄之數目比率，亦可稱之為期刊重複性。本研究重複性是指相同的文章被二個或二個以上系統同時收錄之數目的比率。

文獻回顧

文獻重複性相關研究最早由 Bradford 於 1937 年提出，其研究主要探討期刊文獻被索引摘要工具收錄的重複性，即文獻在兩個以上資料庫同時出現的情況。1960 年代中期，Martyn（1967）以含有摘要之期刊檢索作者欄位，並計算索引與摘要資料庫間的傳統重複性（Traditional Overlap）。此外，Martyn 也以被子植物分類法（Angiosperm Taxonomy）下的 20 個主題，分析各期刊文獻之多元重複性。Martyn 提出索引與摘要資料庫間重複性的優點，闡述索引與摘要資料庫之重複性為各資料庫建造者間努力的重疊成果。Martyn 最後提出，儘管摘要期刊之收錄範圍僅能分析約 70%的相關資料，但有超過一半的文獻都被收錄在超過一個的索引與摘要資料庫中。Wood, Flanagan 與 Kennedy（1972）針對 Chemical Abstracts、Biological Abstracts 與 Engineering Index 三個資料庫進行重複性研究，由三個資料庫中選出 14,592 種期刊並對其進行題名欄位之分析，結果顯示僅有 1%的期刊同時被三個資料庫收錄，同時被兩個資料庫收錄之期刊有 27%，但此研究僅分析資料庫收錄期刊刊名之重複性，而非收錄期刊文章之重複性。

Nicholls（1989）比較圖書館與資訊科學領域的四個資料庫，Library and Information Science Abstracts（簡稱 LISA）、Library Literature（簡稱 LL）、Education Resources Information Center（簡稱 ERIC）及 Information Science Abstracts（簡稱 ISA），檢索獲得圖書及文獻資源共有 50 萬筆，作者比較並說明四個資料庫收錄範圍之差異。ISA 內含大量研討會論文和專刊文獻。LISA 則收錄大量的外國語文資料及資訊科學期刊，較少圖書資訊專門期刊。LL 收錄之內容包含書籍、評述類文章及少量圖書資訊期刊。ERIC 是四個資料庫當中收錄最完整的。

Hood（1998）針對 Dialog 系統的所有資料庫，以模糊集合論（Fuzzy Set Theory）為主題進行重複性研究。自 Dialog 系統所有資料庫下載包含「fuzzy」的資料，並選出和「Fuzzy Set Theory」相關的文獻，約由 100 個不同的資料庫中選出超過 30,000 筆 1965 年到 1993 年的資料。Hood 移除錯誤及統一各項欄位，最終選出了 15,644 筆資料，針對各資料庫間重複性分布、最多重複紀錄、資料庫內部的複本紀錄及前十名資料庫之間的重複性進行比較研究。研究結果發現資料庫間的重複紀錄有很大的差異，15,644 筆資料中被兩個資料庫同時收錄的有

1,922 筆，占 12.29%；而最多同時被 12 個資料庫所收錄的資料有 5 筆，占 0.03%。此外，Hood 亦發現共有 28 個資料庫含有內部複本，內部重複最高的資料庫為 MATHSCI，共有 239 筆。研究最後指出 INSPEC 和 SCISEARCH 兩資料庫中的相對重複性超過 40%，重複性居所有資料庫之冠。

Read 與 Smith（2000）針對圖書館與資訊科學領域的三個資料庫 LISA、Library Literature and Information Science（簡稱 LLIS）及 ISA 進行重複性研究。此研究利用 Dialog 系統，首先檢索三個資料庫的標題欄位，選出在圖書資訊學領域較廣泛應用的 20 個主題，之後針對此三個資料庫兩兩進行比較，用以評估三個資料庫的重複性。其結果顯示 LLIS 所包含的資料有 30,542 筆最高，ISA 則只有 5,094 筆最低。在收錄資料重複性的比較上，LISA 和 ISA 的重複性為 12.2%，LISA 和 LLIS 的重複性則為 10.3%，ISA 和 LLIS 的重複性為 5.8%。由此發現三個資料庫中同時存在的重複性資料只有低於 3%的比率。

Walters 和 Wilder（2003）針對七個特殊學科資料庫及五個跨學科資料庫進行比較研究。其目的在檢驗單一學科和跨學科資料庫，哪一種資料庫的資料涵蓋範圍較完整？學科中的核心文獻是否會被較多資料庫索引及文獻的重複性是否能夠反映學科的相似性？兩位作者收集了 1990 年至 2000 年間美國及加拿大與 later-life migration 主題相關的資料，其收集到 500 筆相關文獻，經檢驗後選出 155 篇來做研究。Walters 與 Wilder 發現不同資料庫平均重複率占 45%。但結論是，沒有任何一個資料庫是最完美的。

Jacsoó（2005）比較 Web of Science、Scopus 與 Google Scholar 三個引文資料庫與系統，以資訊科學之父 Bush 之名著〈As We May Think〉為例進行檢索，結果發現因該文章發表在非學術性質的期刊《Atlantic Monthly》，因此 Web of Science 與 Scopus 均未收錄。然而，Google Scholar 卻有被引用連結。此外，Web of Science 檢索結果顯示共被引用 712 次，其中 90%集中在 1975 年至 2005 年，在 1999 年被引 45 次達到高峰，其餘 10%分布在 1945-1974 年。反觀 Scopus 只檢索到 267 次，因其參考書目收錄年代自 1995 起。

Löhönen（2010）等人比較精神醫學領域最常使用之 PubMed、Web of Science 以及 PyscINFO 三個資料庫之重複性，以注意力缺失及過動症候群流行率（ADHD prevalence）、精神分裂型人格（Schizotypal Personality）、精神分裂症之腦部斷層掃描研究（brain MRI studies in schizophrenia）與精神分裂症之復原（recovery in schizophrenia）等四主題之相關文獻，比較各資料庫彼此之重複性。研究結果顯示，將任意兩資料庫合併能得出收錄範圍達 77%～94%之結果。作者於結論中建議，因沒有任何一個資料庫能滿足搜索者之所有需求，故使用者在選擇資料庫時，應將搜尋主題及收錄範圍之完整性納入考量。若搜尋跨學科領域主題之文獻，應考慮 Web of Science 及 PsycINFO；若要檢索心理學方面文獻則不可遺漏 PsycINFO。

Esmaeil、Kiaie 與 Ketab（2011）的研究在比較搜尋引擎及資料庫的重複性。此研究群體包含六個開放取用的搜尋引擎，這些搜尋引擎來自「Search Engine Watch」網站所介紹之最常用的搜索引擎。該研究利用數據分析和 Microsoft Excel 計算其頻率分布，再使用百分比和平均數來繪製表格和圖表。研究結果發現，在不同的搜尋引擎中，Yahoo 與其他搜尋引擎之重複率最高，約達 40%。Curry Guide 搜尋物理學門資訊的回現率（Recall Ratio）為 77.1%。此外，該系統與其他系統約有 43.7%的重複率。因此，以物理學門的重複程度而言，Meta Search Engine 是最好的搜尋引擎。

Wang（2012）等人基於網際網路發展迅速，加上醫療健康資訊的取得逐漸普及，因此針對 Google、Yahoo、Bing 和 Ask.com 等四個搜尋引擎以乳腺癌為檢索詞彙進行檢索研究。作者以乳腺癌的六個標準以及五個乳腺癌在醫學臨床術語系統化命名（SNOMED CT）最新發布的定義等資訊進行檢索。其研究結果顯示四大搜尋引擎中乳腺癌六大標準的檢索結果都有擠身前 30 名。就有效性而言，最好的搜尋引擎為 Google，其次為 Bing、Ask.com，最差為 Yahoo。此外，以搜尋引擎的檢索重複性而言，其兩兩搜尋引擎之資料都約有 50%的重複性。

重複性研究範圍十分廣泛，包括各種資料來源，例如：出版社、資料庫及搜尋引擎。各種資料類型如期刊文獻、專利文獻、網路資源等，皆是重複性研究的範疇。重複性的多寡，也是圖書館選擇索引與摘要資料庫的重要依據。

研究方法

本研究收集諾貝爾物理學獎 2001～2013 年共 34 位得主（見附錄 1）所發表之文獻製作成著作清單作為研究樣本，進行商業資料庫與開放取用系統之比較研究。選定商業資料庫 Web of Science 及 Scopus、開放取用系統之搜尋引擎 Microsoft Academic、匯集式機構典藏系統 OpenDOAR 與學科性開放取用系統 Astrophysics Data System 等作為研究對象，於 2016 年 6 月完成各資料庫與系統收錄書目之核對。由於學術傳播在數位環境中，資料類型不限於傳統圖書與期刊，如「OpenDOAR」上列有網站資源等，則不在本研究範圍之內。基於期刊文章為科技學術傳播之主體，因此本研究樣本清單以期刊文章為主，此為本研究之研究限制一。此外，由於各商業資料庫所成立時間之差異，加上開放取用運動興起於網際網路盛行後之 1990 年代，本研究所使用之資料庫與系統，其建置年代分別為 Web of Science（網路版：1997 年）、Scopus（2004 年）、Microsoft Academic（2016 年）、OpenDOAR（2005 年）、Astrophysics Data System（1992 年）。不同的建置時間，可能會造成各系統收錄資料數量的差異性，此為本研究之研究限制二。

為了建立諾貝爾物理學獎得主之著作清單作為本研究之研究樣本，必須先檢索各學者揭露個人出版著作之情形，方能建立著作清單。若學者無提供個人之著作清單，或所提供之著作清單不完整，則必須以作者名稱檢索以取得完善之研究樣本清單。本研究根據諾貝爾得主

姓名，分別檢索 Web of Science、Scopus、Google Scholar、Microsoft Academic、OAIster、arXiv.org 與 Astrophysics Data System 等資料庫與系統，因 OpenDOAR 不具備作者檢索功能，故此步驟不使用於 OpenDOAR 檢索。在辨識並去除同姓名作者後，將資料庫檢索結果匯出並整理，所匯出並分析的欄位分別為題名、作者、出版年、出版社、刊名、卷期、頁數等項目，以便於書目檢核時辨識資料的正確性。

本研究以完整之書目著錄格式下載檢索所得之全部書目記錄，以建立完整之書目資料檔。書目計量所處理之資料大多包含文字及數字，而不同系統之著錄方式不同，所下載之書目資料可能出現著錄不完整、著錄方式不統一之情形，必須經由人工查證、比對，補正書目資訊，例如：書目資料誤植、非物理學領域之研究、同一作者不同之名稱、同一期刊不同名稱著錄方式等問題書目，將龐雜之原始書目資訊標準化，以維護研究數據之正確性，方能透過 Excel 去除不同資料庫檢索所得之重複書目。經過去除重複與書目資料處理程序之書目檔，為利用各資料庫彙整所得之諾貝爾物理學獎得主之著作清單。此著作清單，尚須與各種網路資源、諾貝爾物理學獎得主個人網站、研究團隊或所屬機構網站等資源取得的學者著作資訊交互比對，最後才能建立本研究之諾貝爾物理學獎得主（2001～2013）著作清單樣本。

以此研究樣本為依據，分別以書目題名檢索 Web of Science、Scopus、Microsoft Academic、OpenDOAR、Astrophysics Data System 等五個資料庫與系統，逐一記錄各筆書目於不同系統之書目情形，包括書目著錄錯誤或書目重複等情形，以作為諾貝爾物理學獎得主書目正確性與重複性之評估依據。最後整理、統計與比較五個資料庫與系統之各種研究結果加以綜述討論，並進一步提出結論與建議。本研究資料庫與系統收錄資料評鑑之基本假設為：著錄資料正確性越高，內部複本率越低與收錄資料外部重複率越低其表現越優異。

研究結果

本研究透過諾貝爾物理學獎得主之著作清單為研究樣本，希望藉此瞭解商業資料庫與開放取用系統收錄物理學文獻資源之正確性、內部重複性與外部重複性。以下針對此三項研究結果加以論述。

正確性分析

本研究為了檢視各資料庫與系統著錄書目資料之正確性，以 2004 年諾貝爾物理學獎得主 Frank Wilczek 的著作〈Spin-Singlet to Spin Polarized Phase Transition at $\nu = 2/3$: Flux-Trading in Action〉為例，如圖 1 所示，觀察出各資料庫與系統於「$\nu = 2/3$」著錄時有極大差異，例如：Web of Science 著錄為「NU = （2）/（3）」，Scopus 之著錄方式為「$\nu = 2 3$」，Microsoft Academic 之題名出現亂碼「$\nu=2/3$」，Astrophysics Data System 則將之著錄為「$\nu ={2}/{3}$」，而 OpenDOAR 之書目題名則呈現「$\nu=2/3$」。

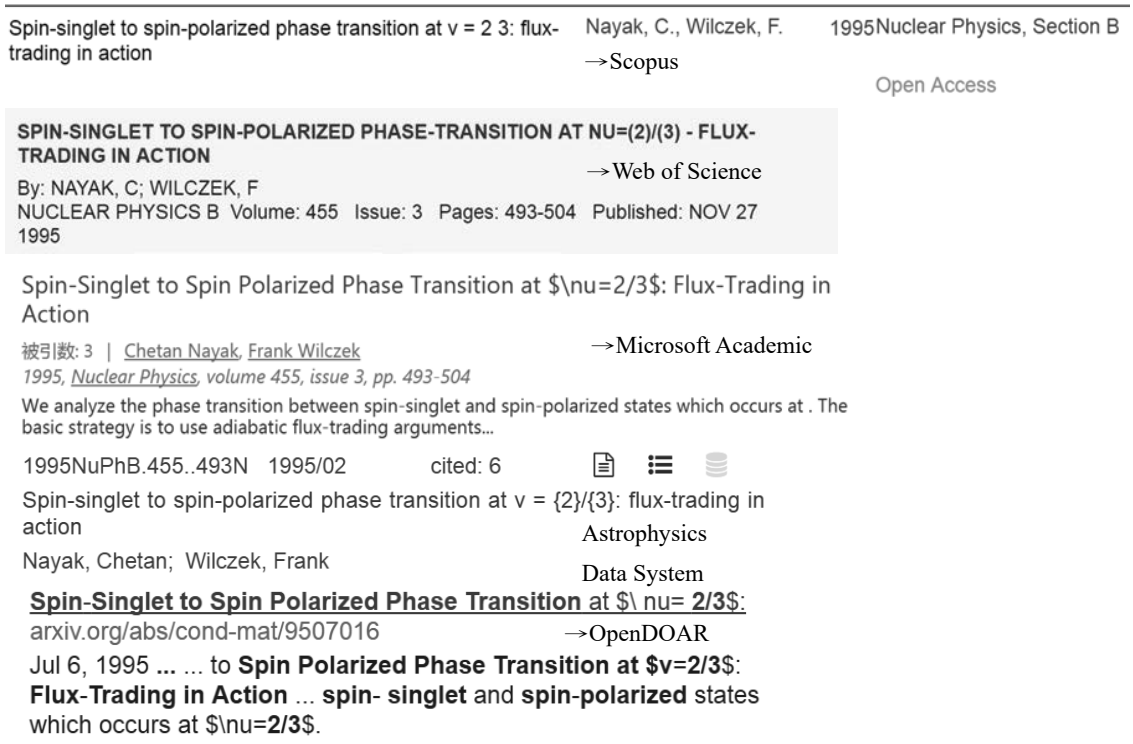

圖 1 〈**Spin-singlet to spin polarized phase transition at ν = 2/3: flux-trading in action**〉**各資料庫書目資料樣態**

相較於商業資料庫 Web of Science 和 Scopus，Microsoft Academic 之文章題名歧異性多且不規律，且書目著錄錯誤之問題比例亦高，但許多著錄錯誤出現在複本之書目上。若以 2011 年諾貝爾物理學獎得主 Saul Perlmutter 之著作〈A Study of 42 Type Ia Supernovae and a Resulting Measurement of ΩM and ΩΛ〉為例，除了 Web of Science 仍舊有特殊符號著錄為拼音之情形外，Microsoft Academic 所搜尋到的兩筆書目，比對兩筆書目之文章題名、作者、出版年份和期刊名，可確定兩者為複本書目且皆有著錄問題：其文章題名分別著錄為〈A Study of 42 type Ia Supernovae and a Resulting Measurement of ΩM and ΩΛ1This work was supported in part by the United States Department of Energy, contract numbers DE-AC03-76SF00098, CfPA, and NSF contract number AST-9120005.1〉和〈A study of 42 type Ia supernovae and a resulting measurement of X M and X K 1〉。一是書目題名後出現不屬於題名之字串，另一則是將特殊符號「ΩM and ΩΛ」著錄為「X M and X K 1」（如圖 2）。

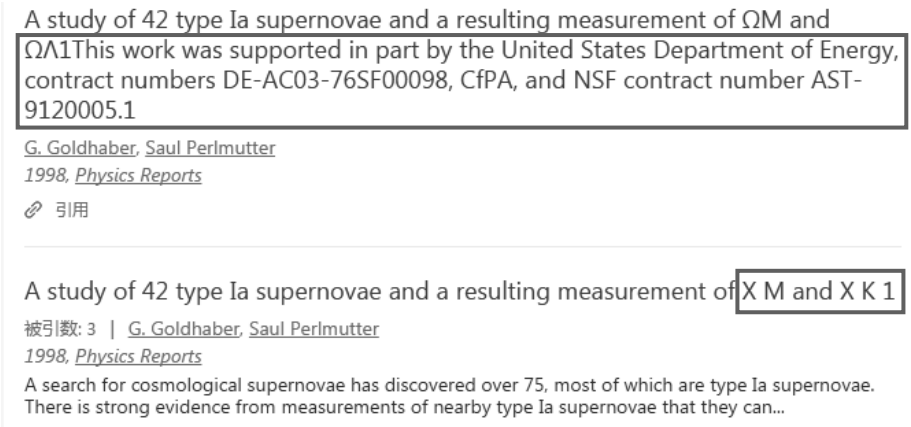

圖 2　書目〈A study of 42 Type Ia supernovae and a resulting measurement of ΩM and ΩΛ〉 Microsoft Academic 問題樣態

前段論述 Web of Science、Scopus 等商業資料庫與搜尋引擎 Microsoft Academic 之書目問題，綜合整理如表 1 所示，三個系統最多見的問題包括書目題名特殊符號或專有名詞著錄方式與其他資料庫不同、書目題名著錄錯誤、書目題名著錄方式與文章題名不同。

Web of Science 之著錄歧異性較其他資料庫多，大多數問題為著錄格式不統一。觀察 Web of Science 問題書目筆數較高之項目為「書目題名沒有著錄冠詞或介系詞」，有此著錄問題之書目多為 2000 年以前出版之書目。至於，書目題名沒有著錄標點符號、羅馬數字著錄為數字、英文數字著錄為數字、特殊符號以拼音著錄四項問題只出現在 Web of Science。由於 Web of Science 容錯性較低，若使用「題名」欄位直接輸入文章題名檢索，假設該書目存在 Web of Science 之著錄問題，則檢索失敗率極高。然而，Web of Science 沒有作者與出版時間著錄錯誤的問題，是其他資料庫與系統不及的優點。反觀，Microsoft Academic 著錄錯誤最多的是出版時間且超出 Web of Science 與 Scopus 甚多，此外，Microsoft Academic 尚有題名出現亂碼的情況。至於，書目題名沒有著錄冠詞或介系詞，或沒有著錄該作者二種問題 Microsoft Academic 沒有發生，此為其強項。相較於 Web of Science 之著錄格式不統一問題，Scopus 則多為題名著錄錯誤，但 Scopus 著錄錯誤之筆數，也僅比 Web of Science 多 12 筆。整體而言，Scopus 優於 Web of Science，Microsoft Academic 亦優於 Web of Science，Scopus 又略優於 Microsoft Academic。

表 1　**Web of Science、Scopus 和 Microsoft Academic 題名著錄問題之書目統計**

問題描述 ＼ 資料庫名稱	Web of Science	Scopus	Microsoft Academic
書目題名沒有著錄標點符號	17		
書目題名羅馬數字著錄為數字	39		
書目題名英文數字著錄為數字	99		
書目題名特殊符號以拼音著錄	64		16

（續表 1）

問題描述 ＼ 資料庫名稱	Web of Science	Scopus	Microsoft Academic
書目題名特殊符號（或專有名詞）著錄方式與其他資料庫不同	65	12	15
書目題名沒有著錄特殊符號		2	
書目題名著錄方式與文章名不同	73	15	23
書目題名著錄錯誤	25	37	19
書目題名出現亂碼			18
書目題名沒有著錄冠詞或介系詞	166	3	
書目沒有著錄該作者	4	3	
書目作者著錄錯誤		9	6
書目的出版時間著錄錯誤		2	45

除了檢視 Microsoft Academic 外，另外二個開放取用系統 OpenDOAR 與 Astrophysics Data System 最常見之著錄問題為「書目題名與文章題名不同」，以 Saul Perlmutter 所著之〈New Constraints on ΩM, ΩΛ, and w from an Independent Set of 11 High-Redshift Supernovae Observed with the Hubble Space Telescope〉一文為例，OpenDOAR 所取得之書目題名為〈New Constraints on ΩM, ΩΛ, and w from an Independent Set of Eleven High-Redshift Supernovae Observed with HST〉，比較兩者題名之差異，可見 OpenDOAR 系統之書目題名中，「Hubble Space Telescope」於出版後以縮寫「HST」表示，阿拉伯數字「11」亦改為英文數字「Eleven」。

本研究在檢索研究樣本書目清單過程中發現，OpenDOAR 共有 210 筆書目之題名與文章題名不同，其題名亦與商業資料庫不同。以物理學文獻而言，arXiv 為 OpenDOAR 之主要書目來源，arXiv 為作者上傳之電子預印本，文章從完成到投稿、出版需經過學術出版流程，文章在出版過程中會有修改文章題名之情形，即造成部分 arXiv 之文章題名與出版後之文章題名不同，使用者可透過 arXiv 書目頁面中的 DOI 連結至出版社書目頁面，比對兩者題名之差異。雖然 OpenDOAR 可檢索各大學術機構典藏資源庫，然而因為受到 arXiv 的影響，OpenDOAR 檢索所得之書目亦有「書目題名與文章題名不同」的問題。至於 Astrophysics Data System 此種問題的資料很少只有四筆，另外還有四筆題名著錄錯誤。

重複性分析

資料庫重複性研究可分為複本（單一資料庫內部重複收錄）和重複性二種（資料庫間重複收錄）兩種，以下分別敘述之。

（一）複本分析

複本是指各資料庫或系統中自身所包含收錄複本書目之情形。以諾貝爾物理學獎得主之著作清單於各商業資料庫、開放取用系統之搜尋引擎及開放取用系統中檢索，其期刊書目資料筆數內部複本性如表 2 所示。由表 2 可見，Web of Science 收錄複本比率最低，為 0.15%，而 Scopus 的收錄複本比率稍高於 Web of Science，為 0.46%。然而，收錄複本性程度最高者

為 OpenDOAR，為 46.2%。綜觀而言，商業資料庫之複本收錄比率普遍低於搜尋引擎與開放取用系統。而具有學術性質之 Astrophysics Data System，其複本程度性低於 Microsoft Academic 和 OpenDOAR，其複本收錄比率僅有 1.3%，與商業資料庫僅相差約 1%。

表 2　**商業資料庫、搜尋引擎開放取用系統之複本收錄分析**

複本分析 / 資料庫與系統	書目資料筆數	複本筆數	複本百分比
Web of Science	5170	8	0.15%
Scopus	5460	25	0.46%
Microsoft Academic	5579	662	11.87%
OpenDOAR	2935	1355	46.17%
Astrophysics Data System	5883	79	1.34%

論及諾貝爾物理學獎得主之著作於商業資料庫中的複本收錄程度低的原因，可能在於商業資料庫之書目以人工建置，文章收錄於資料庫前須經過濾，因此無論是 Scopus 抑或是 Web of Science 之內部重複性皆低於 1%。

探究 Microsoft Academic、OpenDOAR 內部重複性高之原因，Microsoft Academic 之複本收錄比例占 11.9%，觀察其內部複本書目發現其中有文章題名是否著錄冠詞、特殊符號以拼音著錄等篇名歧異情形，或是缺乏來源出版品題名、卷期頁次等書目資訊，可能因此系統無法判別是否為同一筆書目，因而未刪除複本書目。而 OpenDOAR 以 Google Custom Search 提供之檢索功能進行資源庫內容檢索，搜尋各開放取用資源庫之內容，其檢索結果亦可能包含數個書目來源。由於上述複本收錄情形，在於有多個書目資料來源，不同書目來源的書目著錄方式不同，因而其書目著錄情形亦有所差異。

（二）重複性分析

本研究透過分別比較 Web of Science、Scopus、Microsoft Academic、OpenDOAR、Astrophysics Data System 等五個資料庫與系統彼此收錄資料的重複比率。

如表 3 所示，Web of Science 與 Scopus 同為商業資料庫，彼此重複的筆數共有 4646 筆，重複性高達 90%。若與搜尋引擎相較，Web of Science 比對 Microsoft Academic 重複性達 93%。Web of Science 比對具學術性質開放取用系統 Astrophysics Data System 高達 98%；與 OpenDOAR 則有 53%重複。

表 3　**Web of Science 與其他商業資料庫及開放取用系統之重複性**

Web of Science / 資料庫名稱		Scopus	Microsoft Academic	OpenDOAR	Astrophysics Data System
書目總筆數 5170	重複書目筆數	4646	4810	2738	5043
	重複書目比率	89.86%	93.04%	52.96%	97.54%

由表 4 可見，Scopus 同為商業資料庫，其書目和 Web of Science 彼此重複的比率達 85%，與 Web of Science 之書目和 Scopus 比對的重複性（90%）相差僅有約 5%。Scopus 若比對其他開放取用系統 Microsoft Academic 及 Astrophysics Data System，其重複率均超過 90%。分別為 91%和 96%。至於與 OpenDOAR 重複性與 Web of Science 相似為 52%。

表 4　**Scopus 與其他商業資料庫及開放取用系統之重複性**

Scopus	資料庫名稱	Web of Science	Microsoft Academic	OpenDOAR	Astrophysics Data System
書目總筆數 5460	重複書目筆數	4646	4989	2853	5231
	重複書目比率	85.09%	91.37%	52.25%	95.81%

如表 5 所示，Microsoft Academic 和學術性開放取用系統 Astrophysics Data System 相較，彼此的書目重複性極高為 97%。若與商業資料庫 Web of Science、Scopus 相比，重複性皆高於 85%（約 86%～89%）。若將 Microsoft Academic 與 OpenDOAR 相比較則有 48%重複率。

表 5　**Microsoft Academic 與其他商業資料庫、搜尋引擎與開放取用系統之重複性**

Microsoft Academic	資料庫名稱	Web of Science	Scopus	OpenDOAR	Astrophysics Data System
書目總筆數 5579	重複書目筆數	4810	4989	2683	5416
	重複書目比率	86.22%	89.42%	48.09%	97.08%

表 6 顯示 OpenDOAR 與開放取用系統 Astrophysics Data System 比對之書目重複比率高達 100%，至於與其他資料庫及系統交互比對的書目重複性皆高於 90%，可見 OpenDOAR 九成以上的資料都可在其他資料庫與系統找到。

表 6　**OpenDOAR 與其他商業資料庫、搜尋引擎與開放取用系統之重複性**

OpenDOAR	資料庫名稱	Web of Science	Scopus	Microsoft Academic	Astrophysics Data System
書目總筆數 2935	重複書目筆數	2738	2853	2683	2935
	重複書目比率	93.29%	97.21%	91.41%	100%

自表 7 可見 Astrophysics Data System 與搜尋引擎 Microsoft Academic 相比對，其書目重複性極高達 92%。Astrophysics Data System 若與商業資料庫 Web of Science 和 Scopus 比對其重複性分別為 86%和 89%。Astrophysics Data System 與 OpenDOAR 書目重複性約為 50%，其與其他資料庫及系統的差距近 40%。

表 7　**Astrophysics Data System 與其他商業資料庫、搜尋引擎與開放取用系統之重複性**

Astrophysics Data System ＼ 資料庫名稱		Web of Science	Scopus	Microsoft Academic	OpenDOAR
書目總筆數 5883	重複書目筆數	5043	5231	5416	2935
	重複書目比率	85.72%	88.92%	92.06%	49.89%

以下進一步，以整體交叉比對進行綜合論述，如表 8 所示。就商業資料庫 Web of Science 與 Scopus 而言，其與機構典藏系統 OpenDOAR 重複率分別為 53%及 52%，Web of Science 略高於 Scopus1%；其與搜尋引擎 Microsoft Academic 重複率分別為 93%及 91%，Web of Science 又高出 Scopus 2%。至於與物理學開放資源 Astrophysics Data System 之重複率仍然高出 Scopus 近 2%（98%與 96%）。即使就二個資料庫彼此互相比較，Web of Science 與 Scopus 重複率為 90%大於 Scopus 與 Web of Science 85%。由此可見，從收錄資料重複比率觀察此二個商業資料庫，可見 Scopus 重複收錄資料比 Web of Science 少，其表現較優。

就搜尋引擎 Microsoft Academic 與機構典藏 OpenDOAR 而言，其與 Web of Science, Scopus 及 Astrophysics Data System 之重複率依次分別為 86% vs 93%；89% vs 97%；97% vs 100%。如同 Scopus，Microsoft Academic 重複率均低於 OpenDOAR 與其他資料庫與系統之重複率。至於二者相互的重複率為 48% vs 91%，OpenDoar 高出 Microsoft Academic 甚多為 43%。因此就收錄資料的表現上 Microsoft Academic 優於 OpenDOAR。

就 Astrophysics Data System 而言，其與 OpenDOAR 及其他資料庫與系統重複之相對比率來看，四個中有三個百分比低於 OpenDOAR；同樣的，Astrophysics Data System 與 Microsoft Academic 與其他資料庫與系統重複率，也是有三個重複比率低於 Microsoft Academic，只有與 OpenDoar 重複率 50%略高於 Microsoft Academic 48%，整體而言，Astrophysics Data System 比 Microsoft Academic 相對較佳。

表 8　**諾貝爾物理學獎得主著作清單於商業資料庫、搜尋引擎與開放取用系統之重複率（橫欄之資料庫與與直欄之資料庫比對書目資料）**

	Web of Science	Scopus	Microsoft Academic	OpenDOAR	Astrophysics Data System
Web of Science	X	89.86%	93.04%	52.96%	97.54%
Scopus	85.09%	X	91.37%	52.25%	95.81%
Microsoft Academic	86.22%	89.42%	X	48.09%	97.08%
OpenDOAR	93.29%	97.21%	91.41%	X	100%
Astrophysics Data System	85.72%	88.92%	92.06%	49.89%	x

結論與建議

根據前述研究結果，本研究歸納出以下結論，並進一步根據研究過程提出相關建議如下。

整體而言，就資料庫與系統書目資料正確性而言，Scopus 優於 Web of Science，Microsoft Academic 亦優於 Web of Science，Scopus 又略優搜尋引擎 Microsoft Academic。開放取用系統 Astrophysics Data System 優於 OpenDOAR。

各資料庫與系統普遍出現書目著錄格式不統一之問題，影響書目品質與檢索效率。Web of Science 之著錄歧異性最多。然而，Web of Science 沒有作者與出版時間著錄錯誤的問題，是其他資料庫與系統不及的優點。反之，搜尋引擎 Microsoft Academic 著錄錯誤最多的是出版時間且有題名出現亂碼的情況。至於，書目題名沒有著錄冠詞或介系詞，或沒有著錄該作者二種問題 Microsoft Academic 沒有發生，此為其強項。相較於 Web of Science 之著錄格式不統一問題，Scopus 則多為題名著錄錯誤。開放取用系統 OpenDOAR 最常見之著錄問題為「書目題名與文章題名不同」。Astrophysics Data System 有問題的資料很少。

就各資料庫及系統自身收錄複本書目記錄而言，商業資料庫複本比率最低，OpenDOAR 複本比率最高。本研究五個資料庫與系統複本收錄比率最低者為 Web of Science，只有 0.15% 優於 Scopus 0.46%。其次是 Astrophysics Data System 的 1.34%。內部重複性最高者為 OpenDOAR，高達 46%，分析其原因，可能在於其以 Google Custom Search 搜尋各開放取用資源庫之內容，其檢索結果亦可能包含數個書目來源。

各資料庫與系統相對重複率比較結果，最優者為 Scopus，其餘依次為 Astrophysics Data System、Microsoft Academic、Web of Science 與 OpenDOAR。OpenDOAR 比對 Astrophysics Data System 的重複性比率為 100%最高。另外，在搜尋引擎 Microsoft Academic 也可找到其他資料庫或系統 90%以上之資料。因此，以物理學文獻而言，使用者可透過搜尋引擎 Microsoft Academic，以及具備學科特性的開放取用系統 Astrophysics Data System 檢索文獻，能檢索到較完整之物理學文獻。

本研究發現三種開放取用系統皆有全文鏈結功能，Microsoft Academic 若索引到文獻 pdf 全文檔和網頁全文檔，會提供使用者鏈結使用全文檔案；而 Astrophysics Data System 除了掃描早期物理學文獻以提供全文外，亦提供開放取用期刊文獻之全文鏈結。若此三個系統無法提供文獻全文，使用者應使用商業資料庫購買文獻以閱覽全文內容。

本研究以諾貝爾物理學獎著作清單為研究樣本，比較商業資料庫、搜尋引擎與開放取用系統，然而現代電腦科技處理大數據之技術已然成熟，未來網路計量研究，以及不同資料來源之重複性研究，若能以資料探勘（Data Mining）或是文本探勘（Text Mining）等方法進行研究，將有助於大量資料之計量分析。

本研究結果期望能提供圖書館選擇引文索引資料庫與建置機構典藏系統，或者引文資料庫與系統未來發展之參考。此外，亦可作為研究人員進行學術傳播與學術單位進行學術評鑑採用之指標與工具之建議。

參考文獻

Esmaeil, S. M., Kiaie, R. M., & Ketab, F. (2011). A comparison between search engines and meta-search engines in retrieving information related to physics and the extent of their overlap. *Library and Information Studies, 22*(3), 130-140.

Hood, W. (1998). *An Informetric study of the distribution of bibliographic records in online databases: A case study using the literature of fuzzy set theory (1965-1993)* (Unpublished doctoral dissertation). University of New South Wales.

Jacsoó, P. (2005). Google scholar: The pros and cons. *Online Information Review*, *29*(2), 208-214.

Löhönen, J., Isohanni, M. Nieminen, P., & Miettunen J. (2010). Coverage of the bibliographic databases in mental health research. *Nordic Journal of Psychiatry, 64*(3), 181-188.

Martyn, J. (1967). Tests on abstracts journals: Coverage overlap and indexing. *Journal of Documentation, 23*(1), 45-70.

Nicholls, P. T. (1989). Bibliometrics of the laserdiscs applications literature. *Laserdisk Professional, 2*, 106-109.

Poyer, P. K. (1984). Journal article overlap among Index Medicus, Science Citation Index, Biological Abstracts, and Chemical Abstracts. *Bulletin of the Medical Library Association, 72*(4), 353-357.

Read, E. & Smith, C. (2000). Searching for library and information science literature: A comparison of coverage in three database. *Library Computing, 19*, 118-126.

Science Watch (2013). Successful predictions. Retrieved form http://ppt.cc/xA6s.

Thelwall, M. (2008). Bibliometrics to webometrics. *Journal of Information Science, 34*(4), 605-621.

Van Damme, D. (2001). Quality issues in the internationalisation of higher education. *Higher Education, 41*(4), 415-441.

Walters, W. H. & Wilder, E. I. (2003). Bibliographic index coverage of a multidisciplinary field. *Journal of the American Society for Information Science and Technology, 54*(14), 1305-1312.

Wang, L., Wang, J., Michael, L., Yong, L., Wang, Y. & Xu, D. (2012). Using internet search engines to obtain medical information: A comparative study. *Journal of Medical Internet Research, 14*(3), 74.

Wood, J. L., Flanagan, C. and Kennedy, H. E. (1972). Overlap in the list of journals monitored by BIOSIS, CAS, EI. *Journal of the American Society for Information Science*, *23*(1), 36-38.

附錄 1　諾貝爾物理學獎得主（2001-2013 年）

年份	諾貝爾物理學獎得主姓名
2013 年	François Englert and Peter W. Higgs
2012 年	Serge Haroche and David J. Wineland
2011 年	Saul Perlmutter, Brian P. Schmidt and Adam G. Riess
2010 年	Andre K. Geim and Konstantin Novoselov
2009 年	Charles Kuen Kao, Willard S. Boyle and George E. Smith
2008 年	Yoichiro Nambu, Makoto Kobayashi and Toshihide Maskawa
2007 年	Albert Fert and Peter Grünberg
2006 年	John C. Mather and George F. Smoot
2005 年	Roy J. Glauber, John L. Hall and Theodor W. Hänsch
2004 年	David J. Gross, H. David Politzer and Frank Wilczek
2003 年	Alexei A. Abrikosov, Vitaly L. Ginzburg and Anthony J. Leggett
2002 年	Raymond Davis Jr., Masatoshi Koshiba and Riccardo Giacconi
2001 年	Eric A. Cornell, Wolfgang Ketterle and Carl E. Wieman

資料來源：All Nobel Prizes in Physics. Nov. 24, 2013. Retrieved form:
http://www.nobelprize.org/nobel_prizes/physics/laureates/index.html.

DOI: 10.6245/JLIS.201804_44(1).0001

Bibliographic Data Accuracy and Overlap in Open Access System and Commercialized Database

Ming-Yueh Tsay
Graduate Institute of Library, Information and Archival Studies,
National Cheng-Chi University, Taiwan (R.O.C.)
E-mail: mytsay@nccu.edu.tw

Tai-Luan Wu
Graduate Institute of Library, Information and Archival Studies,
National Cheng-Chi University, Taiwan (R.O.C.)
E-mail: 101155002@nccu.edu.tw

Keywords: Academic Assessment; Accuracy of Bibliographic Date; Citation Index Database; Open Access System; Overlap of Database Coverage; Scholarly Communication

---

【Abstract】

This study investigates scholarly communication systems' data accuracy, internal duplication, and between-system overlap in coverage through a comparison of bibliographic records from three open access systems and two commercial databases. The former group of open access systems includes an Internet search engine Microsoft Academic, an aggregating institutional repository OpenDOAR, and an academic open access system in physics the Astrophysics Data System; and the latter group includes Web of Science and Scopus. For the study, searches for the publications by winners of Nobel Prizes in Physics from 2001 to 2013 that were conducted in the five abovementioned citation systems. Bibliographic records representing the sampled publications were retrieved and downloaded from each system. The analytical tasks of sorting, matching, elimination, aggregation, computation, and comparison to assess individual systems' data accuracy, internal duplication rates, and the percentages of overlap between systems were then performed. The findings of the study may provide: (1) valuable information for libraries to acquire citation index databases, build institutional repositories, or create future citation index systems on their own; and (2) a useful resource in scholarly communication and academic assessment.

【Long Abstract】

Introduction

The rapid development of the Internet has facilitated the diversification of scholarly communication channels, which has not only altered the scholarly communication environment but also further promoted the open access movement. The “open access” environment has been collectively developed by academic communities in recent years. Open access systems constructed by scholars and non-profit organizations, such as academic institutions and libraries, may also possess the function of citation indexes as the commercialized citation index database proposed. The purpose of this study is to compare two types of scholarly communication tools, commercialized ones and open access ones. Through practical operation and utilization of commercialized citation databases including Web of Science, Scopus and Microsoft Academic search engine, aggregative institutional repository system OpenDOAR, and Astrophysics Data System, an open access citation system with specialties in physics, the accuracy and overlap of coverage in their retrieval results is analyzed.

Research Methods

In the study, a list of works by Nobel Prize Winners from 2001 to 2013 was created as the research sample. After being collected and processed, the bibliographic data were manually verified to ensure their accuracy by sorting, eliminating and aggregating these bibliographic data. The bibliographic titles of the Nobel Laureates were searched in the two commercialized databases and three open access systems studied. The search results for each title on each database and system were then individually recorded, including the descriptive errors and bibliography overlaps, which served as the basis for evaluating the accuracy and overlap of the bibliography of the Nobel Laureates. Finally, the results and reviews of the studied databases and systems were organized, statistically assessed, and compared. Traditionally, the overlap of journal articles was defined as the same journal article being collected and indexed or abstracted in two or more databases. Overlap also refers to the rate at which the same article is collected in two or more databases simultaneously; this is the definition of overlap used in this study. The commercialized citation index database of Web of Science and Scopus were established since 1900 and 1970, respectively. However, the open access movement emerged in the 1990s after the Internet began to flourish. The open access systems used in this study were constructed in the following years: Microsoft Academic, 2016; OpenDOAR, 2005 and Astrophysics Data System, 1992. The differences in the years of construction might lead to differences in the quantity of the data collected in each database and system; systems constructed earlier might cover data published in earlier years than those constructed later.

Research Results

Errors in bibliographic data are a general problem in all five systems selected, affecting their quality and retrieval efficiency. Web of Science exhibits the most bibliographic errors and inconsistencies. However, it outperforms the others by having zero errors in bibliographic elements of author and publication date. Most bibliographic errors that occur in Microsoft Academic are those of publication dates and occasionally garbled titles, but it does not have the problem of missing author names or articles and prepositions in titles, which is its advantage. The most common issue with both Scopus and OpenDOAR is input errors in titles, while the Astrophysics Data System has very few errors in its bibliographic data.

Overall, in terms of data accuracy, Scopus outperforms Web of Science, and Microsoft Academic is superior to Web of Science, with Scopus performing slightly better than Microsoft Academic. Between the other two open access systems, the Astrophysics Data System outperforms OpenDOAR.

When it comes to internal duplication of bibliographic records, both commercial databases have the lowest duplication rates and OpenDOAR has the highest. Among the five systems, the one with the least duplication is Web of Science with only 0.15%, which is better than Scopus's 0.46%, followed by 1.34% of the Astrophysics Data System. The highest rate of internal duplication, 46%, belongs to OpenDOAR, possibly because it searches every open access repository for content using Google Custom Search and therefore may include several bibliographic sources in the retrieval results.

In terms of between-system overlap in coverage, Scopus is superior, followed by the Astrophysics Data System, Microsoft Academic, Web of Science, and then OpenDOAR. The overlap ratio of OpenDOAR over the Astrophysics Data System is the highest, 100%. In addition, more than 90% of the records in the other systems can be found in the Microsoft Academic search engine. For example, users can retrieve more complete data through the Microsoft Academic search engine and the Astrophysics Data System.

This study finds that all three open access systems have full-text link functionality. The Index in Microsoft Academic provides links to full-text files, by following which users may retrieve full-text pdf documents or access full-text web pages. The Astrophysics Data System, in addition to scans of early physics documents for full-text retrieval, also provides full-text links to open access journals. If the three open access systems are unable to provide the full-text contents of certain documents, users have the option of purchasing these documents from the two commercial database vendors.

Conclusions

In this study, works published by 2001-2013 Nobel Prize Winners in Physics were selected as the sample for a comparison of data accuracy, internal duplication, and between-system overlap in two

commercial databases and three open access systems. Computer scientists have continually developed technologies to process big data. If data mining or text mining can be used for research, it will be much more helpful to quantitative analyses of big data in future network measurement research and overlap research of different data sources.

By comparing data accuracy and overlap coverage of physics resources in commercial databases and open access systems, it is expected that the findings of this study can provide valuable information for readers to select citation index databases and open access systems when they seek information in physics. It can also serve as an informative resource for libraries to evaluate commercial citation databases and open access systems and provide suggestions for the establishment of future institutional collections. Hopefully, the study's findings will prompt database vendors to correct bibliographic data errors to improve retrieval efficiency and data accuracy.

【Romanization of references is offered in the paper.】