國立臺灣師範大學圖書資訊學研究所

博士學位論文

指導教授：謝建成 博士

醫學電子資源大數據分析模式之建構與應用

Model Construction and Its Applications of Medical Electronic Resources Big Data Analysis

研究生：游忠諺 撰

中 華 民 國 一 〇 五 年 二 月

國立臺灣師範大學博士論文通過簽名表

系所別：圖書資訊學研究所

姓名：游忠諺　　學號：898150025

論文題目：醫學電子資源大數據分析模式之建構與應用

經審查合格，特予證明

論文口試委員

邵子順 臺北醫學大學通識教育中心專任教授

林信成 淡江大學資訊與圖書館學系專任教授

曹牧師 國立臺灣大學圖書資訊學系專任副教授

柯皓仁 國立臺灣師範大學圖書資訊學研究所專任教授

謝建成 國立臺灣師範大學圖書資訊學研究所專任教授
論文指導教授

所長簽章：圖書資訊學研究所所長吳美美

中 華 民 國 105 年 2 月 18 日

本論文獲國立公共資訊圖書館博(碩)士論文研究獎助

誌謝

「在研究歷程中，老師能教給你的，不是各種資訊技術，而是看事情的方法與態度。」，在此論文完成之際，首先感謝我的指導教授 謝建成博士，在博士班的求學過程中，給與我許多的寶貴的建議，以及研究過程中的各種指導，讓我能順利的完成論文，僅此致上最深之謝忱。

感謝口試委員林信成教授、柯皓仁教授、邱子恒教授與唐牧群教授在口試期間所給予的寶貴意見與指正，細心審查本論文，方能使之更加完整。

感謝師大圖資所的老師們，陳昭珍教授、卜小蝶教授、吳美美教授、邱銘心教授、謝吉隆教授在修業期間的指導與勉勵，也讓我學習如何從不同的學術觀點，思考與進行各種研究。

感謝我的長官－陽明大學圖書館館長 吳肖琪教授與江秀愛組長，在工作上給與很大的空間與支持；感謝系統組的同事政彥、宴君，讓我在工作與學習之間找到最佳的平衡點。

感謝在師大圖資所遇見的許多人，感謝啟亮、寂絹、亞寧、于第、淇龍、佳香、智惠、忠勤等修課時期的夥伴；感謝苑菁助教在行政事務的協助；感謝恒毅、芳伶、琬姿、毓絜、依婷等學弟妹；感謝嘉豪、苓莉學姐等陪伴我往前進的好朋友們，因為有你們的關懷，讓我能在研究的道路上順利前進。

感謝我的家人，感謝在天上的奶奶 游黃錦女士；感謝我的雙親 游阿龍先生與 李春勤女士，給予我支持鼓勵與悉心關懷；感謝姐姐蓓怡，分享許多求學與研究的經驗；感謝女友詠婷，給予我溫馨的加油和各種協助，家人是我最堅強的後盾，這份榮耀是我們共有的。

最後感謝在生命中遇見的所有人，致上無限的感恩。

游忠諺 謹誌於

國立臺灣師範大學 圖書資訊學研究所

2016 年 2 月 23 日於台北

中文摘要

學術圖書館因考量其電子資源購置成本效益，須透過有效且系統化的方式進行評估，以達成有效管理之目的。故自行建置代理伺服器結合圖書館自動化系統進行使用者權限的驗證，藉以記錄電子資源的使用狀況，但使用紀錄的資料量為動態增加，圖書館無法因應日新月異之龐大使用數據。為解決實務上所遭遇的問題，本研究提出電子資源使用大數據分析模式，模式內容包括：分析原始使用數據、剖析書目暨主題詞等詮釋資料、建立書目及使用紀錄明細資料庫等項目，並以醫學電子資源為研究範圍，做為本研究之驗證個案。

根據醫學電子資源大數據分析模式，本研究採用 PubMed 資料庫的文獻書目與醫學主題詞(MeSH)、期刊評比資料庫-SCImago Journal Rank(SJR)等相關資料來源，自行開發程式進行資料剖析與整理，並採用雲端技術結合圖書館自動化系統與電子資源代理伺服器等系統服務，進行不同層面之應用研究，包括：(1)醫學電子資源使用分析；(2)醫學領域文獻主題分析；(3)醫學領域書目計量分析。其分析結果可提供圖書館做為館藏發展與決策經營等不同層面的參考依據，相關資料可延伸應用至書目探勘、電子資源使用者行為等圖書資訊學相關研究領域，進行更深入與多元的探討。

關鍵詞：醫學電子資源、大數據、書目計量、主題分析、紀錄檔分析、使用者行為分析

英文摘要

In order to effectively manage electronic databases, academic libraries adopt systematic evaluation methods to account for the cost-effectiveness of electronic resource purchases. Proxy servers integrated with automated library systems are used to regulate user access and keep record of electronic resource usage. However, dynamic increase in records and large user data pose strain on the libraries. This study suggests that this matter can be resolved by applying the big data analysis model on the use record of electronic resources. This model consists of analyzing raw user data, scrutinizing metadata such as bibliography and subject headings, and establishing detailed bibliography and use record databases. A verification case study based on medical electronic resources is hereby presented.

In accordance to the big data analysis model, this study adopted relevant resources such as PubMed bibliographic database, medical subject headings (MeSH), and journal ranking database - SCImago Journal Rank (SJR) to develop a program that analyzes and organizes relevant data. Furthermore, cloud technology was used to consolidate system services including automated library system and electronic resource proxy server. Application study was conducted from different levels including: (1) Usage analysis of medical electronic resources; (2) subject analysis of medical references; and (3) medical bibliometric analysis. These results are reference for different functions ranging from collection building to management

decisions. The application of relevant information can also be extended to bibliomining, electronic resource user behavior studies, and other library and information science investigations for more in-depth and diverse discussions.

Keyword: Medical electronic resource, big data, bibliometrics, subject analysis, log analysis, user behavior analysis

目次

中文摘要.....I

英文摘要.....II

目次.....IV

表次.....VI

圖次.....VIII

第壹章 緒論.....1

第一節 研究動機.....1

第二節 研究目的.....10

第三節 研究問題.....12

第四節 研究範圍與限制.....14

第五節 名詞解釋.....17

第貳章 文獻探討.....19

第一節 電子資源評鑑與評估.....19

第二節 電子資源使用調查.....30

第三節 電腦網際網路紀錄檔分析.....34

第四節 醫學領域文獻主題分析.....41

第五節 書目計量學.....65

第六節 書目探勘.....74

第參章 研究設計與實施.....83

第一節 研究方法與研究設計.....83
第二節 研究對象.....84
第三節 研究步驟.....85
第四節 研究架構.....87
第肆章 醫學電子資源大數據分析模式.....91
第一節 醫學電子資源大數據分析模式.....92
第二節 系統建置與資料收集.....96
第三節 資料處理階段.....97
第伍章 醫學電子資源大數據分析模式之應用.....105
第一節 電子資源使用分析.....105
第二節 醫學領域主題分析.....114
第三節 醫學領域書目計量.....134
第陸章 研究結論與建議.....151
第一節 研究結論.....151
第二節 研究建議.....154
參考文獻.....159

表次

表 1 COUNTER 公布的報表類型－期刊、資料庫、圖書、多媒體、標題、文獻 27

表 2 醫學主題詞(MeSH)分類與主題詞個數(2014 年版).....47

表 3 PubMed 資料庫顯示醫學主題詞之樣式－以一般(Summary)與 MEDLINE 格式為例.....49

表 4 書目計量學相關研究－資料來源、分析方法、資料筆數.....71

表 5 歷年點擊下載次數.....106

表 6 期刊跨年度點擊下載次數至少 1 次以上之比例－以 2013 年訂購之期刊為例.....107

表 7 單一年度期刊點擊下載各分列項目與次數－以 2013 年為例.....108

表 8 2013 年期刊點擊下載次數之前十名期刊.....109

表 9 跨年度期刊點擊下載次數－以醫學院為例之前十名期刊.....110

表 10 較為重要之期刊群(A)-以公衛領域科系為例.....112

表 11 特定文獻跨年度點擊次數與讀者之所屬學院/單位之分析.....113

表 12 文獻的重要主題詞(Major Headings)個數統計.....117

表 13 重要主題詞次數統計-主標目出現次數前十名.....119

表 14 下載文獻所屬重要主標目統計-僅針對主標目，不包括含副標目在內..120

表 15 Oncogene 文獻重要主題詞統計-依 MeSH 主題階層.....120

表 16 MeSH 主標目文獻篇數統計-以 Chemicals and Drugs(D 類)為例.....122

表 17 MeSH 主標目文獻篇數統計-以 Amino Acids, Peptides, and Proteins (D12

類)為例 122
表 18 MeSH 主標目文獻篇數統計-以 Proteins(D12.776)類，其文獻超過 1000 篇為例.....123
表 19 MeSH 主標目階層最高者所屬文獻年代分布-以 Tretinoin 為例.....124
表 20 癌症研究於 SJR 期刊與 PubMed 同時收錄文獻之排名數量比例 125
表 21 期刊評比等級 Q1 中癌症研究文獻篇數前十名之期刊.....127
表 22 癌症研究領域文獻之前 20 項核心主題詞.....130
表 23 癌症研究期刊文獻逐年分布統計表.....135
表 24 癌症研究領域出版品類型分佈-按資料型式區分 138
表 25 癌症研究領域出版品類型分佈－按內容型式區分.....140
表 26 癌症研究領域出版品類型分佈－研究報告類.....141
表 27 癌症研究領域期刊文獻分區表.....142
表 28 洛卡定律之計算數據-癌症研究領域期刊(第一作者)(n=1.975 c= 0.6387)145
表 29 洛卡定律之計算數據-癌症研究領域期刊(全部作者)(n=1.975 c= 0.6387)147
表 30 癌症研究領域文獻作者合著數量統計表.....149
表 31 前十名高產量作者與發表篇數.....150

圖次

圖 1 E-Metrics 應用於圖書館的評估面向.....25

圖 2 圖書館資源常見評估方法.....30

圖 3 主題分析與索引概念、類型與方法之互連圖.....44

圖 4 MeSH 醫學主題詞與主題詞編號示意圖 51

圖 5 醫學主題詞(MeSH)架構範例-以 Polyposis Syndrome, Familial 為例...53

圖 6 MeSH 醫學主題詞自動索引圖 55

圖 7 Nicholson 提出的書目探勘資料倉儲資料類型的概念框架 81

圖 8 研究架構圖.....90

圖 9 電子資源使用分析模式.....95

圖 10 網際網路使用紀錄檔資料格式轉換流程.....100

圖 14 Oncogene 文獻出版年代分布 115

圖 15 Oncogene 文獻作者數分布 116

圖 16 Oncogene 文獻的 MeSH 主題詞個數分布 117

圖 17 癌症研究領域期刊的文獻出版年代分布-以期刊評比等級 Q1 為例 129

圖 18 醫學主題詞 Breast Neoplasms 之主題詞路徑與樹狀編號路徑對應圖132

圖 19 主題詞 Breast Neoplasms 之次數統計(以出版年代為區隔).....134

圖 20 癌症研究期刊文獻數量成長曲線圖.....136

圖 21 PubMed 資料庫著錄之出版品類型 137

圖 22 癌症研究領域文獻半對數分布圖.....143

第壹章　緒論

本章為緒論，研究者在這一章說明有關本研究的一些基礎內容與背景知識，共分為四節，分別為：(1)研究動機；(2)研究目的；(3)研究範圍與限制；(4)相關名詞解釋，以做為對於本研究有基本的認識。

第一節　研究動機

本節描述研究動機，分為：(1)圖書館電子資源使用之觀點差異；(2)電子資源的取得管道與身份控管機制；(3)醫學電子資源主題分析之應用；(4)電子資源之書目計量研究面臨大數據之挑戰等，分述如下。

一、圖書館電子資源使用之觀點差異

自 1990 年代起，電子期刊以其跨越傳統載體、不受時空限制、傳播快速等因素而興起，隨著網路興盛與蓬勃發展，電子期刊的重要性早已超越了紙本期刊，學術圖書館徵集期刊的方式，也不再完全以紙本期刊為主，轉而從紙本化朝向電子化，明顯改變學術出版、組織、保存的方法，學術圖書館的館藏發展政策也隨時代更迭持續轉變，但發展至今，圖書館的經費受限於學校或研究機構之總預算逐年減少，除紙本書籍外，電子資源的費用昂貴，圖書館因考量其成本效益，須透過有效且系統化的方式進行評估，以達成有效管理之目的。

電子期刊隨著資訊科技的發達與網路的普及性增加，如今已成為圖書館服

務的重要項目，但其經營模式與紙本期刊大不相同，電子期刊之訂購多以商業資料庫系統套裝方式進行販賣，且來源多元可供圖書館選擇，如出版社、期刊全文匯集資料庫廠商、書目供用中心等，其管理模式相較紙本期刊則複雜許多(王梅玲，2006)。而採以使用為主(Use-centered)之評估時，從館藏評估觀點的角度，圖書館須著重在電子資源的利用情況，以及是否符合使用者之需求程度之探討，其資料來源是根據各種館藏利用之統計資料加以分析，但電子資源其具有多變性高且數量持續增長之特性，過去常用的分析方法，未能完全適用於電子資源，同時圖書館難以從代理電子資源的廠商取得統計資料與使用之原始數據，故在多重的複雜環境中，如何發展其有效測量的方法，在進行電子資源之各項評估時，為圖書館所面臨的重要課題。

電子資源文獻使用(use)探討，可依據使用者類型的不同分為兩種觀點(Rowlands & Nicholas, 2007)：(1)圖書館讀者與(2)文獻作者等兩種身份觀點。從圖書館讀者身份的觀點，認為只要讀者透過圖書館網站於電子資源進行瀏覽、檢索、下載等行為時，即被認定為讀者「使用」該電子資源或文獻；由於圖書館網站會記錄讀者連線至電子資料庫伺服器後的各項行為，圖書館可依據伺服器網際網路使用紀錄檔，判斷某特定電子資源是否被瀏覽、檢索或點擊下載等。此項觀點是以量化數據為主要角度，僅依據網際網路使用紀錄檔是否儲存特定電子資源或文獻相關使用訊息，以判定電子資源是否被使用。但對於「讀者為何要使用該電子資源?」或「讀者檢索該篇文獻的動機與目的為何？」，甚至「如何得知讀者是否真實閱讀下載完成的文獻？」等，相關讀者資訊尋求行為等問題，並未重視。

而從文獻作者之觀點，亦為引文分析之角度思考，意指為當文獻作者在其

學術寫作時，參考他人之觀點或其他資訊來源，並在其文章內容或參考書目註釋中表達其引用之聲明，代表所引用之文獻或其資料來源是有被「使用」的。事實上，無論是一般讀者或文獻作者，其兩種角色是從不同情境下所賦予其意義，就情境而言，因寫作情境的產生，從讀者身份而進階篩選出文獻作者身份的觀點，讓文獻使用的範圍更加的明確，此亦為書目計量學中引文分析的概念。

就圖書館管理層面而言，電子資源的使用是從讀者為出發點進行衡量，其常見延伸發展議題為：(1)館藏使用評估：透過流通研究、館內使用研究、使用者調查、館際互借分析、文獻傳遞測試等方式，評估館藏使用的狀況；(2)發展電子期刊評鑑，以利於決策支援、預算編制、策略規畫、品評鑑等管理議題；(3)使用者行為探索：從使用者的連線（Login）次數、下載次數、連線時間、參考諮詢次數等項目，針對使用者進行統計分析與行為動機方面之探索；(4)書目探勘：從使用者的電子資源下載紀錄為基礎，結合文獻書目資訊，依據圖書館提供的各項服務為情境區隔，發掘各類型的資料樣態，以得知使用者的各項主題偏好等相關議題。

從上述另一觀點來探討文獻使用，常見的探討角度是從引文分析來看整個學術交流體系的變化，是從每年由索引摘要資料庫所推出的引文指標(citation indicator)，來分析各種研究主題的趨勢變化，由於可以看出學科領域中，其所收錄期刊的引用情況與變化，亦言之，都是從文獻本身為出發點，探討期刊文獻如何被「使用」，並從作者與文獻的角度進行研究與評量，但仍有其不足之處，原因在於索引摘要資料庫本身為商業公司所建置，如 Thomson Reuters 公司的 ISI Web of Science、Elsevier 公司的 Scopus，裡面收錄的期刊雖達萬種，但未必圖書館所訂購的電子資源皆有收錄，部份開放取用(Open Access)期刊或由出版

社自行建置的電子期刊皆有可能未必收錄其內，若要得知其書目資訊則須從其他管道得知。

但從圖書館衡量電子資源是否續訂與停訂的角度，雖然是從使用統計報表的數據為考量依據，索引摘要資料庫僅為參考來源之一，但仍有其困難點：(1)主要其資料來源仍來自於訂購的資料庫供應商，對於每一種期刊的使用數據，每一家資料庫供應商皆有提供各自的使用報表，提供的數據報告僅為機構本身購買的資料庫或期刊，在特定期間所成功下載的全文總數，需要圖書館連結至各廠商提供之使用平台進行資料匯出，再重新整理以得到統一的報表格式；(2)圖書館得到之統計數據，僅能得到該館對於該期刊的總下載量，並依據全文檔案的格式-網頁(HTML)或 PDF 兩種格式，按照下載年月進行統計，無法針對機構內的次級機構，進行明確的區隔與統計；(3)從資料庫廠商取得使用原始數據相當困難；(4)無法分析巨量的使用紀錄資料。

另外，學術圖書館為控管電子資源的使用，會自行建置代理伺服器進行使用者權限的驗證，但使用紀錄的資料量為動態增加，資料量之龐大與複雜，以少數主機進行統計分析的傳統思維，已無法因應日新月異之龐大使用數據，故至目前為止，針對電子資源使用之評估模式，未能發展出適當的評估方法與相關研究。

綜觀目前圖書館對於電子資源使用分析之困境，本研究為解決實務上所遭遇的問題，為電子資源使用提出一系統化之分析模式，並以醫學電子資源為研究個案進行模式驗證，說明模式建構的過程中，探討如何分析原始使用數據，建立書目及使用紀錄明細資料庫；以作為圖書館分析探討電子資源使用分析之參考。未來並可將使用資訊整合出如書目計量學、書目探勘與電子資源使用者

行為等圖書資訊學相關研究的資料來源，提供更多的整合應用。

二、電子資源的取得管道與身份控管機制

電子期刊的取得管道，普遍而言分為以下數種：(1)傳統的匯集性全文資料庫(Traditional aggregated full-text database)；(2)期刊閘道和期刊匯集商(Journal gateways and journal aggregators)；(3)出版商的套裝產品(Publisher packages)；(4)紙本與免費電子期刊組合(Free-with-print combinations)；(5)免費期刊(Free journals)與(6)低價期刊(Low-cost alternatives to commercial journals)。圖書館取得電子期刊的管道不外乎上述分類，但電子資源的提供者，為了防止非法的使用者取得資源，必須對使用者的取得方式進行限制（詹麗萍，2005）。

而資訊提供者(如電子資料庫、出版社等)只允許經由授權的單位可以使用其電子資源，在 1998 年，Goerwitz Richard 描述三種權限控制的方式：

1. 發行授權單位的 certificate authorities (CAs)認證：由公正的授權單位發行 CA 認證，採取金鑰制度，一般採取發行公開金鑰(Public Key)與私密金鑰(Private Key)的機制，使用者必須將金鑰儲存在軟碟內，當連結至資料庫時採取金鑰認證，始能使用電子資源。
2. 透過跨機構的資料庫(Cross-Institutional Databases)：由一個公正獨立的圖書館提供跨連結的方式進行認證連結至電子資源。
3. 採用身分驗證系統(Kerberos)：透過身份驗證的方式判定是否為合法的使用者。

時至今日，目前權限控制最常見的方式為：(1)透過認證的使用者與(2)連線 IP 位址篩選，當圖書館購買電子資料庫與所屬期刊的使用權，圖書館提供代表該學校或單位的 IP 地址列表給電子資料庫廠商，藉此識別使用者是否從該學校

或單位所認可的 IP 位址清單，連線至電子資料庫，但使用者雖為學校或單位內的合法成員，但如非在校園或單位內連線，同樣會被辨識為未經授權的使用者。

代理伺服器(Proxy Server)扮演了一個中介的角色，對內，它提供了合法的 IP 位址，並具備使用者驗證的功能，使用者不須考慮連線的本地 IP(Local IP)，而僅須提供個人的驗證資訊，就可透過代理伺服器連結至電子資料庫進行操作。對外，代理伺服器與電子資料庫的進行網路連結時，電子資料庫伺服器僅需認可代理伺服器的 IP，即允許從代理伺服器連線的使用者進行操作。

常見的代理伺服器類型一般分為兩類：(1)URL 重寫型(URL Rewriting)；(2)反向代理型(Reverse Proxy Server)，前者常見的伺服器軟體如 EZproxy，後者為 Squid、Nginx，兩者在使用者設定方面的差異在於 URL 重寫型的代理伺服器，使用者並不需要在瀏覽器進行設定，僅須透過代理伺服器提供的網頁登錄完成使用者驗證的動作，而後者則須在瀏覽器軟體進行代理伺服器的設定，才可進行使用者驗證 的功能，當使用者完成驗證後，其所有操作行為在代理伺服器留下紀錄，電子資料庫伺服器因接收到代理伺服器的連線，同時記錄代理伺服器的所有連線動態與行為。

代理伺服器與電子資料庫伺服器雖同時紀錄使用者的所有操作行為，但兩者仍有其差異性，代理伺服器因具有認證單位內使用者的驗證功能，故能得知合法使用者的認證資訊與連線位置，而電子資料庫因接收的連線來源皆為代理伺服器，經驗證為合理 IP 後即可操作，故電子資料庫僅可記錄該單位的所有連線紀錄與操作動態，如對該單位進行細部分析，其分析項目仍有所局限。

無論何種方式，在讀者下載任何電子期刊時，在 Proxy 伺服器或廠商本身的伺服器皆為留下紀錄，但因為各家資料庫廠商、期刊出版社的網址樣式並非

統一，因此造成了在網址回溯時的困難，在剖析合法(valid)的樣式(pattern)時，皆會增加資料處理上的困難程度，故本研究希冀透過網際網路紀錄檔的分析，從原始資料中進行資料清理與整理，以獲得所需書目紀錄與使用資料。

三、醫學電子資源主題分析之應用

主題分析向來為圖書資訊學研究中，對於分析學術研究發展與典範變遷的一種分析方法，其資料來源主要為學科領域出版的學術文獻，透過對學術文獻內容的分析、分類、歸納等，以探討學科研究領域之範圍、關注議題、時空變遷等不同時期階段之發展重心。

隨著資訊技術的發展，文獻出版的方式，從過去的紙本出版轉變成現今的數位出版，學術出版文獻之數量與日俱增，其研究發展趨勢進而產生了質與量的變化，文獻數量的劇烈增加，不僅增加其主題分析時的廣度與深度，但同時也增加其分析時的困難度，圖書資訊學者除關注各類學科主題的分類體系外，在過去的相關研究中也嘗試透過以期刊文獻為基礎，建構如標題表、索引典等不同呈現方式的主題系統分類，進而應用於特定領域的主題分類與分析，試圖探究不同學科領域的趨勢與變遷。

從文獻內容的角度定義主題分析，主題分析是辨識某作品所包含之知識內容(Intellectual Content)的過程，為依據文獻顯著的特性加以解析，並以數字、符號、名詞、形容詞與名詞的組合或片語標示出文獻中所述及的主題，以做為資料查詢的檢索點(Chu & O'Brien, 1993 轉引自張慧銖，2011)。亦指從傳統的分類編目外，主題分析對於文獻資訊的檢索也具備了重要功能，根據文獻儲存與檢索系統的需要，依照一定的方式對文獻內容進行分析，從中提煉出其主題概念、確定主題類型、並剖析主題結構的過程，而這種代表主題概念的術語，

稱為主題詞(Subject Term)或關鍵詞(Keyword)。通常主題分析的結果以二種方式呈現：(1)數字符號；(2)語言詞彙，前者常見的呈現方式如分類系統，後者則為標題或索引詞 (羅思嘉、陳光華、林純如，2001；張慧銖，2011)。

而從圖書館的技術服務角度探討，主題分析具備的目的有(1)以精確、具體的語彙或符號對於文獻資料提供主題內容之適當描述；(2)揭示出特定主題範圍之研究主題、研究方法、觀念及知識應用上的相關性；(3)對於各學科領域都能提供一般性和專門性詳簡不同的分析款目；(4)提供任何專家學者或一般人習慣的術語，以便檢索其所需資訊；(5)將不同國情與不同學科專家，或因觀念變遷造成的不同標目予以聚集；(6)利用參照方式將具有關聯性的主題做有效地連結，以利讀者蒐集到更完整的資料(陳敏珍，1994 轉引自張慧銖，2011)。

從圖書館參考服務層面而言，良好的主題分類與檢索策略有利於滿足學科專家或一般人的資訊檢索行為，以醫學研究領域中的實證醫學(Evidence-Based Medicine, EBM)為例，實證醫學所強調的證據，是從現有的醫學文獻找出適合的研究結果，以協助醫師進行治療上的佐證，圖書館員除提供醫學文獻上的各種證據外，教導醫師如何有效率在醫學文獻尋找證據亦屬於參考服務的範疇，透過專家學者共同制定的醫學主題詞表，檢索 PubMed 資料庫收錄之醫學期刊文獻，並累積長期檢索的經驗，以制定較佳的搜尋策略，取得適合的醫學文獻(Ann, M. , Angela, E., ＆Susan, M. 著，楊宗翰譯，2008)。

相較與醫學領域具有長期更新且由專家學者制定的主題詞表，過去進行圖書資訊學領域的主題分析時，由於缺乏具有共同制定與長期更新的主題詞表，專家學者採取主題擬定的方式主要分為兩種：(1)針對研究的學科領域範圍，訂定包含整個學科領域的完整類表；(2)事先檢視已蒐集文獻的內容，根據在文獻

中出現的主題，自行擬定主題類目。該兩種方式會因針對其特定的目標，擬定或使用不同的主題分類體系，實務上在運用上述兩種方式進行研究時，各有其使用的時機與範圍，同時隨著時間變遷與環境更迭，在進行研究發展趨勢的分析時，必須在既有的基礎與架構上，擬定新的分類體系，易因為主題概念分析不完全，或難以忠實的反應文獻主題的全貌等因素，影響其分析結果的精細程度與發展彈性(蔡明月、徐韻婷，2009)。

四、電子資源之書目計量研究面臨大數據之挑戰

研究人員為維持其研究動能，須取得不同類型與來源的學術資源，如學術會議的論文、電子資料庫、具有同儕評閱的期刊、教科書等多樣化的資源(Scheckler, 1982)，隨著資訊科技的進步，各類型研究領域的學術產出快速增加，連帶影響學術傳播的發展，故透過量化技術，運用各種數學及統計的測量方法，進而評估現代文明的發展與成長，除可描述學術發展的現況，亦為書目計量之研究內涵與精神。

書目計量學的研究對象和內容主要集中於：(1)文獻本身及其內容資訊；(2)與文獻相關的統計指標（邱均平、文庭孝與宋艷輝，2014）。在資料來源，研究者透過文摘、索引、書目、年鑑等檢索工具，取得文獻出版的數量、時間、類型、出版地、作者等資料，做為統計時的量化單元，同時因為在文獻的利用過程中，產生可用於計量的有用資訊，如文獻申請複印數、書籍借閱數、讀者來源、人數等，增加書目計量學新的資料類型，增加其研究產出的豐富性。

然而，文獻資料收集的完整性，向來為進行書目計量學相關研究的困難點之一，亦影響計量研究之分析結果。如在布萊德福定律的延伸研究，Brookes強調布萊德福分佈的實證與理論若要完全吻合，則文獻的書目必須符合(1)定義

明確的學科主題；(2)書目必須完整，應包含所有相關的文獻與期刊；(3)書目的時間範圍必須有所限制，使得所有期刊生產文獻的機會均等。其論述中其表明，必須收集至完整的書目資訊，始能符合布萊德福定律的規則 (Brookes, 1969，轉引自蔡明月，2003，頁 153)。

由上述可知，過去受限於紙本期刊的載體特性，徵集文獻書目之原始資料取得不易，在資料處理的過程中，必須以人工回溯方式進行資料的整理、清理與統計，資料量的多寡，亦影響計量結果的可信度，研究者在無法取得所有文獻的情況下，採取抽樣方式並推估所需文獻量，其研究結果雖有其研究價值，但亦存在其統計誤差之現象。

而電子資源的發展改變傳統文獻書目資料徵集的便利性，文獻索引資料的數位化與相關電子資料庫的誕生，亦減少研究者在取得書目資料的困難，如常見的引文索引資料庫-Web of Science 與 Scopus，與醫學文獻索引資料庫-PubMed 等，則將其收錄之期刊文獻書目數位化，並提全文連結之功能，大幅改善研究者取得原始文獻與引用文獻之便利性，同時因其收錄資料量具有一定規模，Web of Science 與 Scopus 亦開發不同之計量指標，在資料的質與量層面，文獻索引資料庫具有一定代表性，故從事書目計量之研究者，其資料來源亦多為上述之資料庫提供之書目資料，經資料整理之過程後，再進行計量分析。

第二節　研究目的

綜觀目前圖書館對於電子資源使用分析之困境，本研究為解決實務上所遭遇的問題，為電子資源使用提出一系統化之分析模式，並以醫學電子資源為研

究個案進行模式驗證，說明模式建構的過程中，探討如何分析原始使用數據，建立書目及使用紀錄明細資料庫；以作為圖書館分析探討電子資源使用分析之參考。未來並可將使用資訊整合出如書目計量學、書目探勘與電子資源使用者行為等圖書資訊學相關研究的資料來源，提供更多的整合應用。

主題分析向來為圖書資訊學研究中，對於分析學術研究發展與典範變遷的一種分析方法，為減少研究者須針對特定的目標，在既有的基礎或架構上，擬定新的分類體系，故本研究為得知醫學領域的主題分布，採用美國國家醫學圖書館所公布的醫學主題詞表，做為主題分類基礎，並提出一個醫學電子資源主題分析的架構，進行學術期刊與文獻書目資料的統計，做為從事書目計量學相關研究的基礎。

為了描述學術發展現況，運用數學及統計等量化技術，以評估現代文明之發展與成長，為書目計量之研究內涵與精神。然文獻資料收集之完整性，為書目計量學相關研究主要困難點之一，亦嚴重影響計量研究之分析結果。過去在人工處理資料情況下，資料蒐集之完整性有其難以避免之困境，而今文獻資料內容之數位化建立與普遍性似乎開啟解決此問題的窗口，本研究希冀透過資訊技術的協助，減少人工作業時間與改善資料處理流程，進行電子資源大量資料之書目計量研究，做為未來在從事書目計量學相關研究之參考。

有鑑於此，本研究之目的分為(1)電子資源使用分析；(2)醫學領域主題分析；(3)醫學領域書目計量等三個分析層面；。

首先在電子資源使用分析層面，其目的在於：

1. 建立醫學電子資源大數據分析模式。
2. 建立分析網際網路使用紀錄檔之原始數據之流程。

3. 建立使用者、使用紀錄與書目資料之使用明細資料庫。

4. 依據不同學院系所單位，產生統計數據。

在醫學領域主題分析層面，本研究希冀達成以下目的：

1. 根據美國國家醫學圖書館出版之醫學主題詞表(Medical Subject Headings, MeSH)，建立醫學主題詞樹資料庫。

2. 依據 PubMed 資料庫提供之完整書目資料，建立醫學文獻資料庫。

3. 依據期刊評比資料庫-SJR 提供之期刊清單，建立醫學期刊清單資料庫。

在醫學領域書目計量層面，本研究希冀達成以下目的：

1. 得知醫學研究領域之文獻特性與成長趨勢。

2. 透過布萊德福定律驗證，得知癌症研究領域之文獻分區。

3. 透過布萊德福－齊夫定律驗證，得知核心區期刊數量。

4. 透過洛卡定律驗證，調整斜率值與常數值，試圖求得最適公式，以符合洛卡定律

第三節　研究問題

本研究為解決電子資源使用分析層面，在實務上所遭遇的問題，根據研究動機與目的，所提出主要研究問題包括：

1. 醫學電子資源大數據分析模式為何？

2. 如何分析網際網路紀錄檔之原始數據？

3. 如何進行大量醫學領域文獻之主題分析？

4. 如何進行大量醫學領域文獻之書目計量分析？

本研究因透過不同研究個案驗證，在特定醫學研究領域以癌症研究(Cancer Research)領域為研究範圍，並提出細部的研究問題，分述如下：

在電子資源使用分析方面，以某醫學大學使用 Wiley 電子資料庫為研究情境，分析其使用數據，其研究問題為：

1. 電子資料庫歷年點擊(含下載)之次數分布為何？

2. 電子資料庫實際使用狀況為何？

3. 電子資料庫在特定年度點擊(含下載)之使用狀況為何？

4. 醫學院成員在電子資料庫點擊(含下載)之醫學電子期刊為何？

5. 公共衛生相關科系成員在單一年度之點擊(含下載)狀況為何？

6. 特定期刊文獻之點擊(含下載)情況為何？

在醫學領域主題分析方面，分為兩項：(1)以 Oncogene 期刊為例；(2)癌症研究領域之主題分析，在 Oncogene 期刊為驗證個案方面，其研究問題包括：

1. 使用者點擊下載之 Oncogene 期刊文獻，其文獻出版年代分布為何？

2. 使用者點擊下載之 Oncogene 期刊文獻，其文獻作者數分布為何？

3. 使用者點擊下載之 Oncogene 期刊文獻，其文獻之醫學主題詞次數分布為何？

在癌症研究領域期刊文獻之主題分析方面，以期刊評比資料庫 SJR 收錄之癌症研究期刊清單為研究範圍，其研究問題包括：

1. 癌症研究於 SJR 期刊與 PubMed 均收錄之文獻，其排名數量比例為何？

2. 以期刊評比等級 Q1 為篩選條件，其所屬文獻篇數前十名之期刊為何？

3. 以期刊評比等級 Q1 為篩選條件，其所屬文獻出版年代分布為何？

4. 癌症研究領域之核心主題詞為何？

5. 以特定重要主題詞- Breast Neoplasms 為篩選條件，其所屬文獻次數分布為何？

在醫學領域書目計量分析方面，其研究問題為：

1. 癌症研究相關領域期刊文獻，其文獻成長趨勢為何？

2. 癌症研究相關領域期刊文獻，其合著作者分佈現況為何？

3. 癌症研究相關領域期刊文獻，其文獻類型分布現況為何？

4. 癌症研究相關領域期刊文獻之文獻分布，是否合於布萊德福定律？

5. 癌症研究相關領域期刊文獻之作者分布，是否合於洛卡定律？

第四節　研究範圍與限制

本研究之範圍與限制如下：

一、網際網路使用紀錄檔資料收集時間與來源

本研究在電子資源使用分析與醫學領域主題分析之過程中，皆採用某醫學大學圖書館代理伺服器之網際網路使用紀錄檔，該代理伺服器軟體為 Squid(Squid Developers, 2013)，採取純文字紀錄透過該代理伺服器對外連線的行為。

在醫學領域主題分析方面，為分析使用者點擊(含下載)Oncogene 期刊文獻之文獻特性與主題，其網際網路使用紀錄檔時間自 2009 年 8 月 17 日至 2012 年 7 月 6 日，依據 Nature 資料庫的網址規則，篩選出屬於 Oncogene 期刊的 PDF

網址，共有 15,345 筆全文 PDF 點擊(含下載)紀錄，經進一步去除重覆，共有 5,313 個 PDF 檔。

在電子資源使用分析層面，本研究採用某醫學大學圖書館代理伺服器，篩選 Wiley 電子資料庫使用數據為研究個案，原始紀錄檔案為 1.04GB，共 4,181,850 筆紀錄，記錄時間由 2009 年 8 月 17 日至 2014 年 4 月 21 日，剔除無法完整辨識之紀錄，並以 2013 年之期刊清單進行條件篩選，在未進一步分析文獻書目之詳細資料，共有 1,216,156 篇全文 PDF 點擊(含下載)次數。

二、醫學主題領域資料庫與範圍

在網際網路使用紀錄檔分析方面，本研究分析 Oncogene 期刊與 Wiley 資料庫的點擊(含下載)紀錄，另醫學領域主題分析與醫學領域書目計量分析方面，本研究採用 SCImago Journal & Country Rank(SJR)資料庫中(Scimago Lab, 2013)，主題領域為癌症研究(Cancer Research)且 PubMed 共同收錄的期刊。

三、醫學領域期刊文獻類型

由於各期刊除一般研究論文(research article)外，一般亦包含如期刊主編的說明(editorial)、信函(letter)、校訂(correction)、勘誤(erratum)、綜論(review article)、附註(addition)、摘錄(excerpt)、討論(discussion)等各式文體之文獻，本研究並不事先區隔，僅針對 PubMed 資料庫收錄，且賦予醫學主題詞文獻進行分析與統計。

四、書目資料來源與樣式

本研究的書目資料來源取自 PubMed 資料庫，文獻須被 PubMed 收錄且其文獻狀態為已完成索引(PubMed - indexed for MEDLINE)，即取得該文獻的書目資料，該書目資料以 MEDLINE Abstract 樣式進行著錄，但 PubMed 會不定時更新書目資訊，因此在 Pubmed 未完成收錄索引前，其書目資料僅有基本資料，而未有主題詞。

五、作者權威控制的限制

在作者權威控制方面，由於作者間可能有姓名相同，但實際上卻不同人的情況，或是姓名格式的誤植的情況，現在已有 ORCID 計畫的導入，未來作者辨識的問題將得到解決，故本研究以 PubMed 的書目資料為依據，將每一作者皆視為獨立個體，並不進行作者的權威控制。

六、醫學主題詞的限制

在醫學主題詞方面，由於每一篇文獻的醫學主題詞個數不一，本研究僅篩選該文獻的重要主題詞(Major Headings)進行統計，同時採用 MeSH 2012 年發佈版本進行資料庫的建立。

七、個人資料隱私的限制

因法令因素，個人資訊的收集、處理、使用皆受到法令限制，本研究並不針對使用者個人的使用行為進行統計與分析。

第五節　名詞解釋

為使本研究中所述及之名詞有更準確的定義，以及行文之便利，茲將相關重要名詞解釋如下：

一、電子資源（Electronic Resource）

電子資源指藉由電腦操作而呈現的作品，其利用方式包括直接取用或遠端利用，而有些電子資源則需利用與電腦相連的周邊設備才能使用，電子資源可以每天 24 小時隨時提供使用者服務，突破距離的限制，在本研究指的電子資源主要為電子期刊類型的電子資源，包括了 PubMed、Nature、MEDLINE 等電子資源庫。

二、伺服器紀錄檔(Log)

伺服器在運作時依據功能的不同，分為 Cache.log、Access.log、Store.log、referer.log、useragent.log 等產生不同的 log 文件，其中主要在紀錄使用者的動作的 log，可以在 Access.log 內得知，Access.log 是依據使用者在連網時對外的請求動作進行紀錄，並紀錄了所有的 HTTP 的訪問過程，除了還沒有發送數據就已經斷線的連線動作，Access.log 的欄位依據需求不同，可自行定義所需欄位，本研究採用的伺服器紀錄檔是由代理伺服器軟體 Squid 與 ezproxy 所產生。

三、數位物件識別碼(Digital Object Identifier, DOI)

DOI 為識別數位資源的機制，包括的對象有視訊影片、文件或書籍等不同

數位化物件，它有為資源命名的機制，也具備將識別碼解析為實際網址的通訊協定，目前 DOI 的政策制定由國際數位物件識別號基金會(International DOI Foundation)所負責，本研究在進行取得書目資訊的過程中，除了文獻本身的標題(Title)外，DOI 也為另一個取得書目資訊的辨認途徑。

第貳章　文獻探討

本章為描述與本研究相關之理論框架與研究，分為六節，包括(1)電子資源評鑑與評估；(2)電子資源使用調查；(3)電腦網際網路紀錄檔分析；(4)醫學領域文獻主題分析；(5)書目計量學；(6)書目探勘，分述如下。

第一節　電子資源評鑑與評估

本節試從(1)電子期刊評鑑；(2)電子資源評估；(3)E-Metrics 評估電子資源使用狀況；(4)COUNTER 計畫的報表規範等不同觀點項目，說明過去相關研究提出的觀點與實際上所遭遇的困難，以下分述之。

一、電子期刊評鑑

電子期刊(E-Journals)，與傳統紙本期刊的首要區別，是從傳播載體進行區隔，從廣義上的解釋，是以連續性方式出版，並透過電子媒體發行的期刊，包括各種電子形式，如微縮、光碟、線上資料庫、網路資源等(郭麗芳，1997)。電子期刊雖源自於紙本期刊，但無論是傳播載體或是傳播方式，兩者仍有其差異，對於館藏政策的發展，Mercer (2000)認為在新時代需要重新定義館藏的品質與價值，特別是在電子資源方面，由於傳統的期刊評鑑方式無法完全反映電子期刊的使用全貌，而且圖書館電子館藏的使用資訊並不充分揭露，圖書館需

要因應館藏性質的變化，重新定義電子資源的價值。

從電子期刊評鑑的角度，在資料取得方面，圖書館與代理商網路伺服器記錄保存許多電子期刊使用資料，主要是瞭解使用者如何使用電子資源，但若從圖書館管理的角度進行思考，可分為幾個層面：(1)決策參考：幫助圖書館如何發展電子期刊館藏；(2)預算編製：讓購置預算達到最佳使用；(3)策略規畫：提供各項電子期刊未來推廣策略之參考；(4)品質評鑑：從使用紀錄得知電子期刊與資料庫的質量優劣，協助做為評鑑之參考(王梅玲，2008，頁 99)。就台灣的學術圖書館而言，王梅玲(2008)認為，學術圖書館極需要網路服務評鑑，特別是在電子期刊層面，但相對而言，對於圖書館員要如何進行電子資料庫與電子期刊之評鑑，仍有其技術上的困難，需要有計量的方法與工作協助之需求，藉以達到能採以微觀或鉅觀的不同角度，進階探討單一電子期刊的評鑑、使用者偏好等相關議題。

二、電子資源評鑑

從使用評估角度分析圖書館的電子館藏，詹麗萍(2005)分析電子館藏的使用評估，對於使用(Use)或使用法(Usage)，認為需要多方面的資訊及實際數據，並提出(1)資料收集途徑與(2)以使用為主的分析方法等兩層面的論述，在資料收集途徑方面，除了請廠商提供資料庫使用統計數據外，圖書館自行收集的方法包括：(1)建立連結電子資源的網頁(Intermediate HTML pages)；(2)在網頁伺服器使用特殊軟體(pass-through scripts)；(3)透過線上目錄使用特殊軟體(Using the library’s online catalog with pass-through scripts)；(4)使用獨立的資料庫系統(Database-driven access systems)；(5)使用代理伺服器(Proxy servers)。

透過上述途徑，其主要目的在於(1)透過建立連結電子資源的網頁，透過點

擊該網頁的次數，計算每一項連結電子資源的的點擊次數，即可估算使用者對於該電子資源的連線次數。(2)在網頁伺服器或圖書館的線上目錄使用特殊軟體，如點擊流(Clickstream)側錄軟體，推知使用者在電子資源或瀏覽圖書館網站時，其網站之瀏覽歷程；(3)透過獨立的資料庫系統，當使用者透過其檢索機制，紀錄每次查詢之檢索詞，以得知使用者之檢索偏好；(4)使用代理主機，結合 IP 或自動化系統之身份驗證機制，驗證使用者是否為合法使用者，同時結合上述概念，紀錄使用者之電子資源瀏覽與下載歷程，進而分析使用者之使用行為，如瀏覽、下載、列印、電子郵寄全文等。

在評估方法層面，詹麗萍認為實務上圖書館在進行館藏評估時，以使用或使用者為主的角度，有六種常見的分析方法，分述如下(詹麗萍，2005，頁 170-173)：

1. 引文分析(citation Analysis)：

引文分析法為書目計量學(Bibliometrics)的研究範疇之一，主要利用各種數學及統計方法，將文獻的被引用率進行分析與探討，藉此探究文獻、作者、期刊等之間的引證關係。作者在發表其著作時，若引用某種期刊之文獻，而該期刊又恰為圖書館所訂購之館藏，透過引文分析，圖書館可獲取使用者對其資料使用的線索，進而可以協助圖書館員在館藏發展與服務策略層面的政策制定與調整，以滿足使用者的資訊需求。

2. 流通研究(Circulation Studies)：

分析館內各項流通資料，依不同使用者族群、地點、出版日期、主題分析、使用者在自動化系統所紀錄之交易類型（如：借閱、續借、催還、預約）等加以分析，其結果可用於改善館藏發展實務與修正資源分配，但該方法僅能觀察

流通資料之使用，但禁止流通的館藏並無法進行分析統計，同時未能提供讀者借閱失敗或館藏無法滿足讀者需求之相關資訊。

3. 館內使用研究(In-House Use Studies)：

館內使用研究可透過上架統計、閱讀調查表單、觀察、訪談或問卷調查等研究方法搜集相關資訊，此法常用於非流通性資料的使用研究，結合前述流通研究的方法，可提供館藏使用更正確的資訊，但此法受限於使用者是否能充分合作與願意表達意見，否則影響調查結果之正確性。

4. 使用者調查(User Surveys)：

使用者調查為常見圖書館在取得使用者意見反應時的一種方式，透過口頭訪談、電話訪談、電子郵寄問卷、網頁問卷或傳統的書面問卷實施調查。所獲得的資訊以作為評定館藏是否有效滿足讀者需求的依據，藉此幫助圖書館得知讀者興趣和改變的趨勢。

5. 館際互借分析(Interlibrary Loan Analysis)：

館際互借紀錄可反映館藏的使用狀況，其意義在於表示當讀者在查詢館藏後發現該圖書館缺乏所需要的資料，透過館際互借取得所需要的資料。透過館際互借資料的分析，可確認圖書館本身的館藏領域之不足處，藉以提供修訂館藏發展政策的參考。

6. 文獻傳遞測試(Document Delivery Test)：

用於圖書館在讀者需要資料時能否即時供應的能力，最常使用的方式是找到一份最能反映使用者資訊需求的清單，先檢視館藏目錄，若顯示無館藏時，則另外搜尋其他管道是否能檢索到這些資料，其次則是測試取得資料的時候需時多久。文獻傳遞的測試主要是提供客觀衡量館藏是否滿足讀者需求的方法。

上述六種評估方式對於圖書館的館藏評估與使用模式，皆以紙本館藏的思考為出發點，詹麗萍認為，電子館藏評估雖重要，但遲遲未能發展出適當的評估方法，傳統的評估方法未必適用於電子資源，因為過去的評估方法是以紙本為基礎發展，電子資源並非實體有形之物質，故其認為電子資源無法以出借流通、館際互借或文獻傳遞服務等方式取得，另也未予以分類編目、也無須排架，故傳統的流通研究、館際互借分析、文獻傳遞測試、館藏直接分析等方法均不適用於電子資源評估。

從借閱流通的角度而言，紙本期刊尚可採用借閱次數做為是否使用之依據，但電子資源並非是用傳統借閱的方式進行衡量，同一期刊同時可為多讀者下載閱讀，使用次數是並非限制於圖書館的開館時間與區域，任何具有合法身份能取用(access)圖書館電子資源的使用者，透過網路並通過身份驗證，即可下載期刊文獻，對於「使用」狀態的產生，並不需透過館員即可進行，因此電子資源的使用，是任何時間即都在發生的狀態，必須透過系統性的方法進行分析，始能將複雜繁瑣的原始使用資料，轉換為可加值利用的資訊。

三、E-Metrics 評估電子資源使用狀況

E-Metrics 最初應用於電子商務，為網路企業所用的網站績效評量法，被認為是以科學方法檢驗網站通行數據的一種工具，其定義為:「測量網站成功程度的計量法」，其目的在瞭解網路使用者行為，確定網站是否吸引適當的使用者來訪、評定網站經營策略是否得當。其目的包括：(1)了解網站使用者的行為；(2)確定網站是否吸引適當的使用者來訪；(3)評定網站經營策略是否得當(詹麗萍，2005，頁 192-193)。

E-Metrics 應用於圖書館，依據圖書館服務與館藏評估兩個層面，再細分四

個項目，每一個項目都有各自所屬的元素，分述如下(White & Kamal, 2006)：

1. 以使用者為中心(User-centered)：

以網站使用者做為評估的重點，如包括使用者區別惟一性(Unique visitor)、何時離開網站(Turnaway)、連線時間區段(Session)與瀏覽頁面(Visit)等元素。

2. 以服務為中心(Service-centered)：

以網站頁面的服務點擊次數做為衡量重點，如頁面的檢索、觀看(View)、下載與列印等提供之功能之執行次數。

3. 電子資源(E-resources)：

以資源類型為統計重點，如電子期刊、電子資料庫、電子書與數位檔案(Digital Document)等資源類型的數量。

4. 電子服務(E-services)：

以提供之網站功能為重點，如圖書館之檢索目錄(WebOPAC)、數位參考服務(Digital reference)、整合查詢(MetaSearch)與分散式協定－開放式定位址(OpenURL)等元素。

E-metric 從上述不同類型進行評量，透過量化或質化等不同的資料收集與分析方式，藉此得知圖書館服務與館藏的現況，並結合圖書館服務與館藏評估，做為一個聯結平台的分析方式。

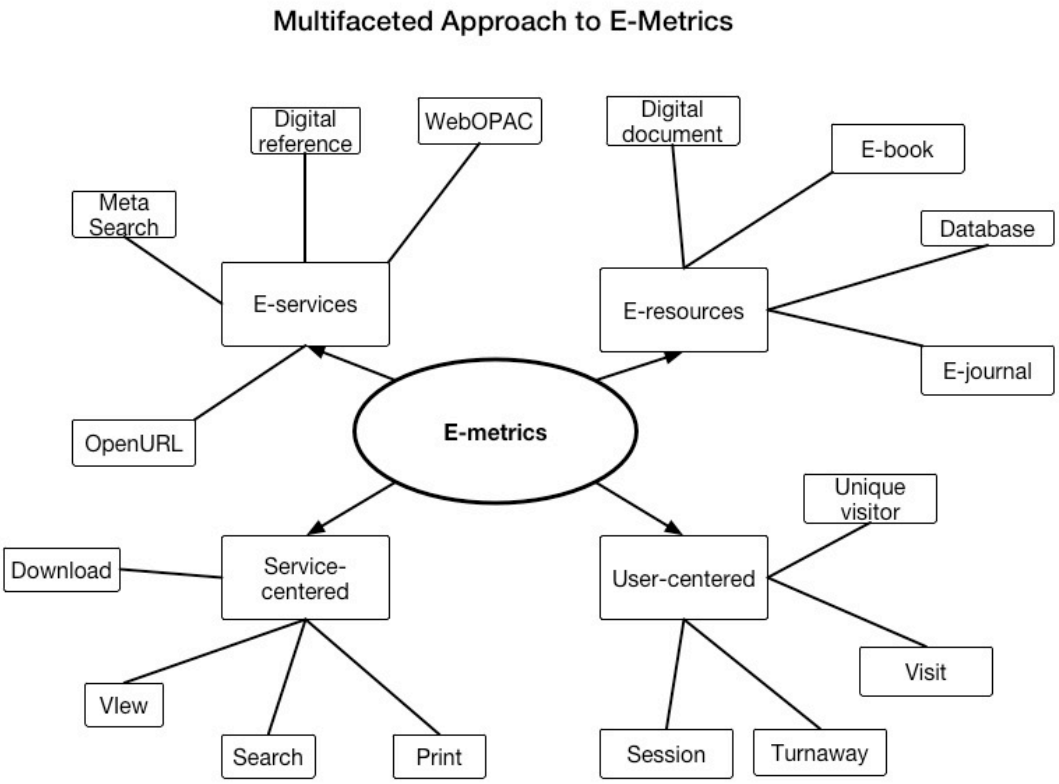


圖 1 E-Metrics 應用於圖書館的評估面向

資料來源：Andrew White and Eric Djiva Kamal, *E-Metrics for Library and Information Professionals* (New York: Neal-Schuman, 2006), 6.

雖然電子期刊是 E-metrics 的評量項目之一，但因為電子資源的使用，是不限定任何時間點，在即刻發生的狀態下，所產生的交易紀錄 (Transaction log file)，遠比圖書館自動化系統儲存的館藏資料與書本借閱紀錄量更為龐大，過去採取人工分析的做法，僅能分析小部分的資料集，對於資料集龐大的使用紀錄，根本無法透過傳統人工分析進行分析，並且在合理時間內達到擷取、管理、處理，並整理成為人類所能解讀的資訊，事實上，圖書館在進行電子資源使用分析時，勢必處理使用紀錄等原始資料，透過系統化的分析模式，將原始資料進行清理、整理與整合，始可拆解成各種不同的獨立資料集，可得出許多額外的資訊和資料關聯性，藉此可察覺讀者、單位機構內的更多潛在資訊與使用趨勢。

另外，圖書館員需要分析使用數據的原因：(1)衡量不同的電子資源與其服務價值；(2)協助在訂購期刊時的決策；(3)規畫圖書館資源的基礎建設與分配模式；(4)支援圖書館行銷與服務。另一方面，對於電子資源廠商，使用數據的意義在於：(1)電子期刊與傳統紙本期刊不同，紙本期刊透過複製的份數決定計價，而電子期刊則是用下載次數反映出電子期刊的現有價值，因而廠商藉此嘗試訂定新的計價模式；(2)評估透過不同的資訊來源，來得知該電子資源在市場上的相對重要性；(3)為編輯者提供開發新產品的系統需求與評估依據；(4)改善網站的使用設計與導覽功能，增加收集使用數據的能力，進而能夠在市場分析時，得到更佳的推論與使用分析(Shepherd, 2006)。

因此，電子資源統計報表為圖書館與出版商所採用的一種評量方式，其功能對於圖書館管理方面，可以協助館員得知圖書館所訂購之電子資源在訂購年度之讀者點擊(含下載)狀況、評估電子期刊對讀者的價值、比較各出版商的產品及文獻傳遞方式，藉報表提供之資訊，淘汰不合時宜之電子資源及制定長期訂購計畫等 (葉譓歆，2006)。也因為電子資源使用之統計數據對於圖書館在制定館藏發展的重要性極高，圖書館如何制定採用標準化的方式，持續蒐集與分析數據，乃屬於長期計畫的範疇，故圖書館陸續提出多項評估電子資源的研究計畫，以 E-Metrics 的評估計量方法，應用於電子資源的評估上，並積極從事研究開發，其中如 COUNTER(Counting Online Usage of NeTworked Electronic Resources) (COUNTER, 2002) 、ARL E-Metrics Project 計畫等，另外以 COUNTER 所制定出的報表格式為目前實務上最常使用之類型。

四、COUNTER 計畫的報表規範

COUNTER 計畫規範出版商提供各類型的統計報表，在其計畫進行過程中，自 2003 年公布第一版的統計報表後，分別於 2005 年、2008 年、2012 年公布新的報告類型，截止 2014 年 7 月止，最新的版本為 2012 年 4 月公布的第四版 (COUNTER, 2012)，報告以資源類型為區隔，分為期刊(Journal)、資料庫(Database)、圖書(Book)、多媒體(Multimedia)、標題(Title)等五類，另 2014 年 3 月，COUNTER 另公布了文獻(Article)的統計報表類型 (COUNTER , 2014)，而對於功能層面，則分為「強制要求出版商提供」與「非強制要求提供」兩層級，其資料來源是源自電子資料庫伺服器的網際網路使用紀錄(Transaction Logs)，其紀錄根據 COUNTER 的標準轉化成可利用且易讀的資訊，產生適合圖書館解讀及利用的電子資源使用報表。其報表類型參見表 1。

表 1 COUNTER 公布的報表類型－期刊、資料庫、圖書、多媒體、標題、文獻

名稱	意義
Journal Report 1 (JR1)	依據月份及期刊名稱統計單本期刊內文章全文的有效需求數 (successful requerst)
Journal Report 1 GOA	依據月份及期刊名稱統計 Open Access 單本期刊內文章全文的有效需求數
Journal Report 2 (JR2)	依月份及期刊名稱統計單本期刊被拒絕使用(turnaways)的總次數
Journal Report 3 (JR3)	依月份、期刊名稱與內文格式(Page-Type)統計單本期刊內全文的有效需求數與被拒絕使用的次數。
Journal Report 3 Mobile	依月份、期刊名稱與內文格式(Page-Type)統計單本期刊內，透過行動裝置(Mobile)取得全文的有效需求數與被拒絕使用的次數
Journal Report 4 (JR4)	依月份及服務統計被搜尋的總次數
Journal Report 5 (JR5)	依據出版年(Year-of-Publication, YOP)與期刊名稱統計單本期刊內文章全文的有效需求數
Database Report 1 (DB1)	依月份及資料庫名稱統計資料庫被檢索、檢索結果的點擊與紀錄閱覽的次數

名稱	意義
Database Report 2 (DB2)	依月份、資料庫名稱、類別統計資料庫被拒絕使用的總次數
Platform Report 1	依月份、資料庫名稱、平台統計資料庫平台被檢索、檢索結果的點擊與紀錄閱覽的次數
Books Report 1 (BR1)	依月份及題名統計單本圖書的有效需求數
Books Report 2 (BR2)	依月份及題名統計單本圖書中章節的有效需求數
Books Report 3 (BR3)	依月份、題名與類別統計單本圖書被拒絕使用的次數
Books Report 4 (BR4)	依月份、平台與類別統計單本圖書被拒絕使用的次數
Books Report 5 (BR5)	依月份及標題統計被搜尋的總次數
Multimedia Report 1	依月份及多媒體內容集合統計多媒體的有效需求數
Multimedia Report 2	依月份、多媒體內容集合與項目類型統計其有效需求數
Title Report 1	依月份與標題統計全文與書本章節的有效需求數
Title Report 1 Mobile	依月份與標題統計在行動裝置內全文與書本章節的有效需求數
Title Report 2	依月份、標題與類別統計被拒絕全文使用的次數
Title Report 3	依月份、標題與內文格式統計標題的有效需求數
Title Report 3 Mobile	依月份、標題與內文格式統計在行動裝置取得的有效需求數
Article Report 1	依文獻標題與出版社收集與統計使用資訊
Article Report 2	依文獻作者、月份與 DOI，統計不同來源的有效需求數
Article Report 3	依文獻作者，統計個別文獻有效取得次數的報告綜整

資料來源：COUNTER, “The COUNTER Code of Practice for e-Resources: Release 4”, COUNTER, April 2012, accessed May 28,2014, http://www.projectcounter.org/r4/COPR4.pdf; COUNTER, “The COUNTER Code of Practice for Articles: Release 1”, COUNTER, March 2014, accessed May 28, 2014, http://www.projectcounter.org/documents/counterart_cop_MAR2014.pdf.

從上述 COUNTER 的電子資源統計報表格式得知，主要是以期刊、資料庫、書籍、多媒體內容、文獻等做為分類依據，再配合時間、平台、載具等其他變數進行統計分析，對於出版商而言，透過交易紀錄，足以提供學術單位圖書館之統計數據，且逐年公布報表類型，從期刊、資料庫、書籍，進步至行動載具、多媒體與文獻為統計區隔單位的多種類型報表，但對於各圖書館而言仍有其不足，圖書館在制訂續訂政策時，常需要多種類型的數據以茲佐證，COUNTER 提供了不同類別的的統計報表，但對於學術機構編制內的下層單位之細項統

計，卻受限於紀錄檔之欄位限制而無法進行分析，亦為其在報表格式制定時的先天限制，在實務上，亦有廠商為解決圖書館員須自行下載各不同出版社或電子資料庫的使用數據，結合 SUSHI 協定(NISO, 2013)開發整合平台，提供報表整合與資料匯出等服務，雖可自行設定電子資源與學院科系之間的對應關係，但仍無法完整解決上述圖書館內部單位細項統計分析的問題。

總而言之，從圖書館管理的立場，能夠透過圖書館自動化系統的使用數據，進行使用評估，但取得的數據皆為圖書館借閱與流通的部份，藉以支援館內流通、館際互借、使用者調查、館內使用研究等項目，但目前圖書館提供的服務，電子資源的需求性已日漸增加，過去讀者到圖書館內尋求館員面對面服務的需求已大為減少，相形之下，透過圖書館網站提供的服務，即可完成其資訊需求，但時至今日，較少分析圖書館網站服務的相關研究。整合過去的分析方法，過去詹麗萍提出的以使用為主的館藏評估方法(參見圖 2)，未提出分析圖書館服務網站研究為主的評估方式，而 COUNTER 提出的電子資源使用報表，也無法以機構或學校內下層單位與使用者做為評估單位，故透過圖書館自行建置代理伺服器的方式，雖可補足其不足，但資料量之龐大，且須專人進行長期的資料收集、處理、分析與解釋，則對於圖書館本身，則又受限於人力資源之困窘，故需要一系統化的分析方法以解決人力處理上的困難。

此外，由於現今網路資源為圖書館提供服務的重要來源，故圖書館網站的使用相關研究也相形重要，特別在電子資源使用層面，是過去較少提出分析方式的一環，因而本研究希冀提出一系統化的分析方法，將原始數據進行加值處理，藉以提供更多的整合應用研究。

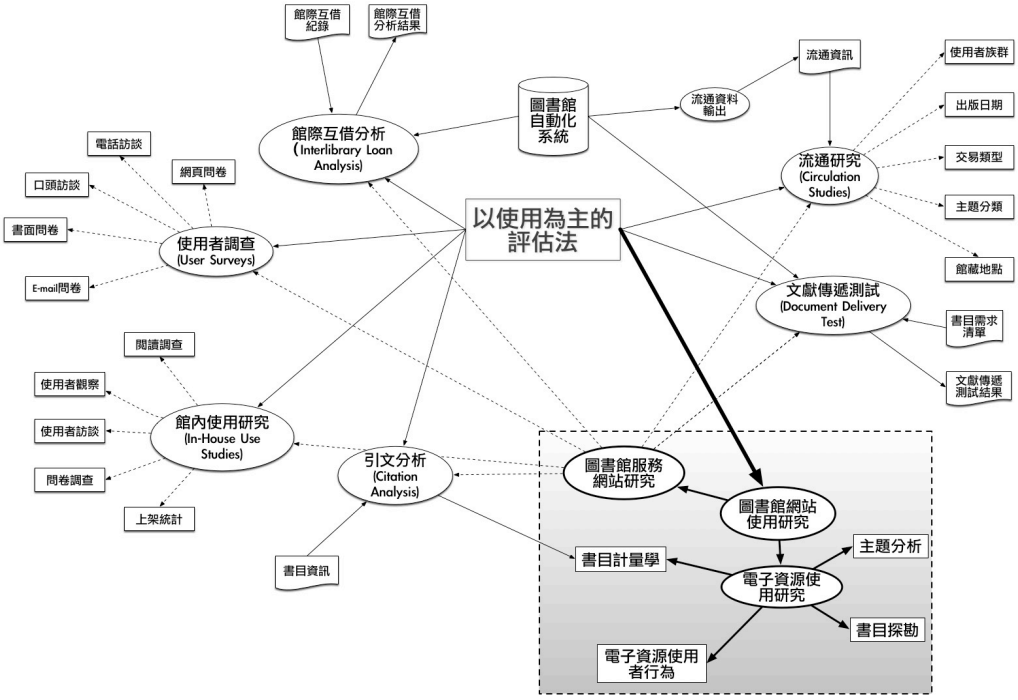


圖 2 圖書館資源常見評估方法

資料來源：本研究自行繪製，其內容整理自詹麗萍，電子資源與圖書館館藏發展(台北市：麥田，2005)，170-74。

第二節　電子資源使用調查

圖書館為了評估讀者的使用狀況，必須先進行電子資源的使用調查研究，才能瞭解使用者的行為，會有不同的使用動機與想法，圖書館清楚使用者的需求與使用目的，才能據此制定妥善的電子資源推廣策略，在過去的相關研究中，從不同的觀點進行調查，顯示圖書館對於電子資源的重視 (羅思嘉，2000)。

一、電子資源影響閱讀期刊行為

電子資源的興起影響了使用與閱讀期刊的行為，Peter Boyce 等人研究科學家使用與閱讀期刊的行為，在其統計數據中發現，科學的期刊使用行為逐年改變，科學家在瀏覽(Browsing)期刊方面，紙本期刊的比例逐年降低，從 1990-1993

年的 100%，至 2000-2001 年的 65.3%，再至 2001-2002 年的 45.2%，而相對瀏覽電子期刊的比例則日漸提高，從 1990-1993 年的 0%，成長至 2000-2001 年的 34.7%，再至 2001-2002 的 54.8%，顯示使用者的取向從傳統的紙本期刊改變使用電子期刊(Boyce, King, Montgomery, & Tenopir, 2004)。

二、電子期刊的期刊偏好

另外電子資源的使用，在經由使用紀錄的累積後，可看出使用者的期刊偏好，Davis(2002)分析 NorthEast Research Library 的 Academic Press IDEAL 的數據中發現，在期刊的使用模式方面，使用者會集中下載某些期刊，其研究數據表示，在全部 206 個期刊中，排名前十名的期刊具有 44%的總下載量，而前 14 名的期刊就具有 50%的總下載量，而排名最後的 123 種期刊，僅占總下載的 10%，顯示在期刊下載方面合於 80/20 法則，而 Morse 與 Clintworth 於 2000 年的研究中，採用 OVID 資料庫的使用紀錄數據，其結果發現在電子期刊的使用量遠超過傳統紙本的使用量，但 Morse 也指出，紙本期刊的使用數據較電子期刊難以取得，而其使用者的期刊偏好於 JAMA、Lancet 與 N Engl JMed 等三種期刊(Boyce et al., 2004)。

三、電子資源的連線使用地點

電子期刊的連線地點，對於圖書館的使用調查，也為重要的一環，在相關研究中，在使用地點方面，洪榮照等人於 2010 年的研究調查中指出，研究生依賴的資訊來源中，紙本圖書與期刊已無法滿足其研究需求，電子資源是其仰賴的資訊來源，且研究生多數皆透過遠端校外連線使用電子期刊，以在家連線檢索為最多，其他包括在服務單位或是研究室與實驗室內使用，其次才是到圖書

館現場使用（洪榮照、鄧英蘭、蘇雅敏，2010）。

Khan、Zaidi、Zaffar Bharati 於 2009 年針對印度 Jawaharlal Nehru 大學(JNU)與 Jamia Millia Isiamia 大學(JMI)等兩所學校的教學人員與研究人員進行調查，結果顯示，在教學人員方面，無論是 JNU 大學(85.71%)或 JMI 大學(64.29%)，皆在教學單位的辦公室內使用電子資源；而在學術研究人員方面，JNU 大學的研究者會在圖書館內使用電子資源(67.16%)，而 JMI 大學的研究者主要會在校外使用(34.33%)，其次會在圖書館連線使用(32.83%)(Khan et al., 2009)。

De Groote 在 2003 年進行的研究中指出，不同類型的使用者在連線地點的偏好也有所不同，學生習慣於圖書館與家中連結至電子資料庫，而教師則習慣於從辦公室或研究室中連結至電子資料庫，在期刊文獻取得的行為分為四種：(1)在辦公室或在家中列印全文；(2)在線上閱讀全文；(3)在圖書館列印線上期刊與(4)影印全文四種，而醫學院的學生會採取直接列印全文或在線上閱讀兩種模式，護理系學生則是會將全文透過電子郵件寄回到他們的信箱，另外教職員的行為則是在辦公室列印全文為最多，其次是閱讀他們個人所訂閱的文獻，其次則使用館際互借取得文獻(De Groote & Dorsch, 2003)。

四、電子期刊的取得行為與使用目的

在電子期刊的取得與使用目的方面，洪榮照等人於 2010 年的研究調查中指出，在使用電子資源的目的方面，以準備學期報告的比率最高，其次為準備學位論文，或是因為教師指定才使用（洪榮照等人，2010）。

另外在 2010 年在印度 Jawaharlal Nehru 大學的研究指出，有 65.15%的受訪者表示，使用電子期刊的目的是為了其研究工作，其次有 26.31%的受訪者表示，在使用電子期刊是為了補充其課程學習，另有 9.21%的受訪者，其目的是

為了教學需要，在使用頻率方面，有 44.73%的使用者是每天都會使用電子期刊，其次有 35.52%人是一週 2-4 次，另有 19.73%的受訪者則是一週一次(Siddique & Ali, 2010)。

五、使用電子資源時面臨的困難

在使用電子資源的困難方面，Siddique、Ali 在 2010 年在印度 Jawaharlal Nehru 大學的研究指出，有 42.10%的使用者認為他們想要的電子期刊卻是圖書館沒有訂購而無法使用的，其次有 36.80%的使用者，認為圖書館應提供更專業的課程，以利於使用電子期刊，而另分別 28.94%的使用者認為連線速度太慢，22.36%的使用者認為面臨的問題在終端使用的介面上(Siddique & Ali, 2010)。

洪榮照等人的研究則認為，在使用電子資源的困難方面，最大的困難是不熟悉電子資源的內容與特色，其次是在校外連線使用時必須設定代理伺服器，第三則是需要從過多的檢索結果再次篩選（洪榮照等人，2010）。

Khan、Zaidi、Zaffar Bharati 於 2009 年針對印度 Jawaharlal Nehru 大學(JNU)與 Jamia Millia Isiamia 大學(JMI)等兩所學校的教學人員與研究人員進行調查，JNU 大學的教學人員認為在使用電子資料庫的問題最多的是下載時間過慢(37.14%)，其次是技術問題(34.29%)，而研究人員則認為受限的電腦連線數目(56.94%)是最常見的問題；JMI 大學的教學人員與研究人員皆多數認為問題出在技術問題(57.14%、41.79%)，顯示電腦的技術與網路速度皆是使用者在使用電子資源時，會面臨的問題(Khan et al., 2009)。

六、使用者的檢索技巧與下載檔案偏好

洪榮照等人於 2010 年的研究調查中指出，在使用電子資源的檢索技巧方

面，是「使用與所需資訊相關之適當詞彙來搜索資料，其次是「利用適當的方式擷取所需資訊」(洪榮照等人，2010)。

Khan、Zaidi、Zaffar Bharati 於 2009 年針對印度 Jawaharlal Nehru 大學(JNU)與 Jamia Millia Isiamia 大學(JMI)等兩所學校的教學人員與研究人員進行調查，在使用電子資料庫時，使用簡易或是進階的檢索選項，在教學人員方面，JNU 大學(42.86%)與 JMI 大學(57.14%)兩間學校的教學人員較常使用簡易的檢索功能，而研究者方面，JNU 大學的研究者較認為是簡易與進階並用(43.08%)，而 JMI 大學的研究者仍是採取簡易的檢索功能為主(52.24%)(Khan et al., 2009)。

在文獻的下載檔案偏好方面，Borrego 與 Urbano 在其 2007 年針對 Barcelona 大學的研究中指出，使用者對於下載檔案的偏好，以 PDF 檔案格式為主，優於以 HTML 的檔案格式(Borrego & Urbano, 2007)。

第三節　電腦網際網路紀錄檔分析

本節在說明紀錄檔分析在圖書館取得電子資源使用扮演的角色與相關研究和限制。

一、網際網路紀錄檔相關研究

紀錄檔(Log Files)是資訊系統為了要紀錄使用者在使用資訊系統的所有行為，自動(autonomous)產生的檔案，源自於 90 年代的線上檢索系統(Online Information Retrieval, IR)，紀錄檔分析其主要目的為：(1)改善資訊檢索系統的效能；(2)改善人類使用資訊檢索系統；(3)了解系統與人類之間的互動；(4)辨識

與區隔使用者，進而針對系統進行改善。在紀錄檔分析的發展歷程中分為三個階段：(1)1960~1970 年代中期：主要評估線上檢索系統的性能，該階段主要是針對資訊系統的軟硬體效能做為研究重心；(2)1970 末期至 1980 年代中期：研究線上檢索系統與一般使用者的搜尋行為；(3)1980 年代末至今：開始有了多元化的應用，針對於資訊檢索的使用者進行資訊行為分析(Peters, 1993)。

由於早期發展線上檢索系統的方式，是採取文字界面的終端機模式，與現今透過網頁界面，在操作上有相當大的不同，同時系統產生的紀錄檔也有所不同，現今使用者多採取透過網站進行檢索，當使用者利用瀏覽器輸入一個網頁的網址或對網頁中的連結進行點擊(hit)等存取網頁資源的行為時，對於網站伺服器而言，即是認為使用者發出一項需求，因所存取的網頁中可能嵌入(embedding)不同類型的檔案，如靜態圖形(still graphs)、音訊(audio)、純文字(Plain text)檔案、PDF 檔案等，故使用者取用網頁的行為，即包含了對網路伺服器提出的需求。

使用者在網頁瀏覽過程中，每一次的資源需求或連結，會在提出資源需求的用戶端(Client)、網站伺服器(Web Servers)或代理伺服器(Proxy Server)中留下要求資源的 URL 位址、用戶端主機電腦主機的 IP 位址、需求時間、訊息長度等紀錄。

伺服器會針對每一項獨立需求，記錄其處理的過程，並產生不同的紀錄檔案，分別為：(1)交易紀錄(transactions logs)、(2)錯誤紀錄(error logs)、(3)參考紀錄(referred logs)與(4)代理程式紀錄(agent logs)。交易紀錄是記錄使用者對於網站伺服器發出每一個需求的相關資訊，包括造訪者輸入之資料(Patron’s entry)、時間、使用端主機位址、系統回應(system response)、存取資源等。錯誤記錄則

是當網站伺服器對使用者的需求傳回錯誤訊息時，保留在伺服器中的紀錄，這些錯誤通常是使用者所要求的資料不存在或是使用者權限無法存取所需求的資料；參考紀錄則是記載使用者需求是那些網址(Uniform Resource Locator, URL)產生的，如由搜索引擎或是透過其他連結到所需求資源的網頁，亦指使用者在連結至所需求的網頁時，非直接輸入所需求網頁的網址，而是透過其他的網站點擊網址而連線至所需求的網頁；代理程式紀錄則指的是為何種代理程式對網站伺服器發出需求，所謂的代理程式如搜索引擎廠商為了取得網頁內容而開發的網頁機器人(robot)等（Peters et al., 1993；林頌堅，2000）。

紀錄檔文件通常具備一定的紀錄格式，以網站伺服器為例，常見的格式如W3C 格式、NCSA 格式等，其原始數據的收集方法，不同於研究方法中常見的問卷調查或訪談等方式，紀錄檔的產生隱藏在資訊系統背後，在使用者與資訊系統的互動中，記錄使用者在資訊系統中的所有動作，換言之，它是事先所定義，紀錄使用者於何時何地，在資訊系統內執行的所有動作，只要資訊系統為每個使用者定義不同的唯一識別的方式，它就能夠區別每一個使用者的系統內的所執行的任何單一甚至是所有的行為。

但紀錄檔分析但仍有其限制，Kurth(1993)認為無論採用何種研究方法，交易日誌分析仍具有其限制性，回顧過去在收集資料紀錄的狀態，系統人員只要開始收集日誌分析的數據，在長期收集相關數據後，資料量的龐大與複雜性，就會是試圖進行日誌分析的研究人員的嚴格挑戰，基本的問題在於兩個層面：(1)日誌檔的資料元素(data elements) 是否易被量化，如搜尋類型(Search types)和檢索集(Retrieval set)的大小。(2)日誌檔的情境元素(textual elements)，如使用者的搜尋歷程，使用者從開始進行資料檢索的歷程，在日誌檔的紀錄要如何進

行切割與分析則是另一個問題，因此無論是研究人員採用自動化或是人工作業的方法進行分析，對於資料龐大與複雜程度高的交易日誌分析，都是屬於長期與艱難的研究歷程(Kurth, 1993)。

Jansen 與 Spink(2000)在分析檢索系統的紀錄檔時，認為在搜尋期間的主要目的，是確定使用者在搜尋期間內的檢索數量(Jansen & Spink, 2000)。但如何辨認個別的搜尋期間(Sessions)是在資料處理上重要且較為困難的問題，如使用者在檢索界面未輸入任何檢索詞，是否仍須納入頁面瀏覽的統計？或研究者已規範特定的檢索樣式始納入分析，是否會影響最後的分析結果？皆為在進行紀錄檔分析須注意的要項。而紀錄檔文件的時間欄位會是辨別搜尋期間的一個重要的判斷依據，根據系統的不同，時間欄位的紀錄方式也會影響分析的困難度，一般而言，會採取以人可直接判讀與機器判讀的兩種格式做為紀錄的格式，前者如常見的”YYYY-MM-DD”的年月日格式，而後者則是以毫微秒的格式進行時間欄位的註入，雖然易於讓系統解讀，但可讀性較差，如採人工作業時，會讓研究人員無法直接判讀，仍須採用其他軟體或程式進行轉換為人類可讀的格式，才能夠進行分析與區別搜尋期間，且影響研究的時間歷程，這也是進行交易紀錄檔分析時會遭遇的困難點。

Hancock 等人(1990)認為，雖然紀錄檔分析可以提供有用的實證依據，但仍舊有其限制，這些限制包括：(1)缺乏正確的數據類型與相關資訊，以充分告知使用者如何或為何要搜查他們；(2)並非所有的系統都會有記錄的設施；(3)分析與識別搜尋期間(session)的困難性；(4)系統回應(system response)並非每項都紀錄；(5)大量的數據讓分析時增加了困難度。

因此在進行網路使用記錄檔案與分析時，必須事先進行的規畫，其程序分

為五項：(1)規畫數據收集；(2)決定要紀錄的數據元素；(3)選擇紀錄的軟體；(4)處理保密的數據；(5)數據收集與記錄，分述如下(Yu & Apps, 2000)：

1. 規畫數據收集：資料收集的方式不等同於傳統進行量化研究的方法，如問卷調查或是訪談，透過紀錄檔分析，了解使用者與系統互動的行為，這是即時且無法重新再來的過程，透過系統記錄使用者的行為，無法像人的眼睛一樣具有靈活性與適應性，而是忠實的一筆筆記錄使用者在系統內所執行的所有動作，因此研究者必須在研究過程的初始階段就明確設定所需要收集的數據類型。

2. 決定要紀錄的數據元素：依據各種不同的變數決定所要紀錄的資料，分為五類：(1)代理人的行為(Agent of actions)；(2)行為的類型(Type of Actions)；(3)行為的時間(Time of actions)；(4)行為時的目標(Objects of actions)；(5)行為的模式(Modes of action)換言之，記錄的數據應詳細提供有關使用者的行動，包括身份、做了什麼行為、如何進行、與何時何時開始進行，這些都必須在進行確定。

3. 選擇記錄的軟體：選擇軟體方面儘可能能夠將系統與人類之間進行的互動進行全方位的收集。

4. 處理保密的數據：為了要分析個別使用者的使用者狀況與使用情境，必須紀錄一定程度的使用者資料，如使用者帳號、電子郵件地址、機構、身份別等個人資訊，為了讓這些數據嚴格保密，必須採取用代碼來表示個別使用者的身份，才能夠進一步處理。

5. 數據收集與記錄：依據目標的不同，建議紀錄檔必須包括以下欄位：(1)日期與時間；(2)所要瀏覽的網址；(3)使用者身份驗證；(4)使用者的操作行為記錄

透過網站紀錄檔收集原始使用資料，林頌堅(2008)建議研究者可以利用網

站流量資料做為網路使用研究的初步分析資料，在資料記錄之時，研究者對於使用者的實際操作行為，是幾乎沒有任何影響的，研究者若能善用這些原始紀錄進行分析與解讀，對使用者在網路上的活動可以有客觀的認識與分析。透過流量分析，研究者可以知道較多數使用者偏好的網路資源，以推估使用者上站需求動機（林頌堅，2008）。

二、網際網路紀錄檔分析在台灣學術圖書館資訊服務之應用

在台灣，網站統計數據是圖書館最常使用的評估工具之一，圖書館採用紀錄檔分析應用於各項網站服務，其目的主要用來評估電子館藏的使用(Usage)，詹麗萍的研究指出，在台灣的大學圖書館據以瞭解讀者對於電子館藏使用情形的方式主要有四種：(1)請電子資料庫廠商提供數據；(2)分析圖書館主機的(Log Files)；(3)訪談讀者；(4)問卷調查等；其中以「請資料庫廠商提供數據」為主，其次為分析圖書館主機的紀錄檔(Log Files)（詹麗萍，2005，頁 177）。

除上述方式外，圖書館另可透過其他第三方軟體(Third-party software)的方式收集使用數據，包括：(1) 建立連結數位資源的網頁 (intermediate HTML pages)；(2) 在網頁伺服器使用特殊軟體 (pass-through scripts)；(3) 透過線上目錄使用特殊軟體(using the library’s online catalog with pass-through scripts)；(4) 使用獨立的資料庫系統 (database-driven access systems)；(5) 使用代理主機 (proxy servers)，圖書館會將使用統計數據，做為電子資源是否續訂或停訂的決策參考(Breeding, 2009)。

故實際在進行相關研究時，研究者根據所能取得的原始資料進行分析。李美燕(1994)分析國立政治大學圖書館線上公用目錄的交易處理記錄(Transaction Log)，從分析使用行為的角度，進而深入了解該校師生的資訊搜尋型態，但該

研究受限於圖書館自動化系統只保留 10 天的查詢歷史檔(Query log)，故僅能針短期的時間進行研究，未能針對長時間的交易處理記錄進行分析採討(李美燕，1994)。林頌堅(2000)從世新大學圖書館網站為分析個案，分析自 1999-2000 年的連線紀錄，得知該網站的資源存取次數、用戶端的上站次數與 IP，以觀察大學圖書館網站的使用情形 (林頌堅，2000)。

戴玉旻等人(2001)分析 Elsevier SDOS 電子期刊系統為分析個案，藉由分析 Elsevier SDOS 的使用紀錄檔，得知臺灣 SDOS 使用者的使用習慣和資訊尋求行為(Information seeking behavior)，做為提供電子資料庫系統設計與改進、以及圖書館進行資訊服務時的參考(戴玉旻、陳莉君、偌弗·卡克拉爾、柯皓仁，2001)。

由於網站紀錄收集不易，透過第三方軟體(The thirty-party) 進行紀錄檔分析亦為研究人員常採用的方式，林信成、洪銘禪(2010)以 Google Analytics 做為主要分析工具，分析「台灣棒球維基館」(Taiwan Baseball Wiki)，再輔以問卷調查與訪談，以期掌握使用者的多樣面貌，做為網站經營、管理、改進的依據。黃慕萱等人(2013)透過 Google Analytics 網站紀錄分析工具，分析「拓展臺灣數位典藏」(Taiwan Digital Archives Expansion Project)與「數位島嶼」(Exhibition of Cyber Island, Taiwan)，分別針對網站整體歷時表現、國家/地區造訪者來源、造訪者來源網站以及造訪者搜尋關鍵字等面向進行分析，做為網站經營者的改善參考資訊。

吳政叡(2010)分析西文期刊文獻快遞服務(Journal Articles Delivery Express, JADE)-RapidILL 系統之日誌檔，以 2008 年 6 月至 2009 年 12 月做為時間變數進行統計，以得知 RapidILL 系統在輔仁大學的使用現況，其目的為(1)按照使用者類型做為區隔，藉以了解使用者分佈；(2)對照該校圖書館期刊經費分析，

以瞭解透過西文期刊文獻快遞服務在抑制期刊經費成長的成效，以及可能帶來的經濟效益。

總結上述，研究者通常利用交易處理記錄的資料改進資訊檢索系統的內容、系統的使用情形、資訊尋求者的資訊搜尋方式，並給予管理者行政參考的依據。而交易紀錄檔分析(Transaction log analysis)應用於圖書館服務評估，總結在台灣地區的相關研究中，若須長期進行網際網路使用紀錄的追蹤與分析，主要遇到的困難為(1)如何取得與保留長時間的原始數據；(2)龐大的紀錄資料儲存數量如何分析；(3)如何分析機構內的次級單位的使用資訊等因素，為進行研究時所面對的挑戰。

第四節　醫學領域文獻主題分析

本節為說明從事醫學領域文獻分析，所需具備之領域知識，分別為(1)主題分析與領域分析；(2)MeSH 醫學主題詞簡介；(3)MeSH 的階層結構與路徑；(4)MeSH 自動索引；(5)MeSH 主題分析相關研究。

一、主題分析與領域分析

主題分析(Subject Analysis)的定義，Vickery(1968)認為主題分析就是一種「資訊的分析」(Analysis of information)，即自文獻的內容中衍生出文字或符號，形成該文獻要義之簡要描述，其描述可用來辨識文獻、提供檢驗文獻的檢索點、指示文獻的內容，或作為文獻的替代，它主要是根據文獻顯著的特性解析內容，以名詞、或是形容詞與名詞的組合、或是片語或句子，甚至數學等特殊符號，以標示文獻所揭示之重要主題(Vickery, 1968，轉引自陳明來，2002)。

Hanne Albrechtsen(1993)認為，在建立主題資料的過程，涉及到兩個步驟：(1)透過具體的語句描述，藉以表現所感知的資訊；(2)從語句中萃取出詞彙，根據主題詞表或分類法，轉化成適合的主題詞，以利於進行索引之用。故在其立論內容，分為(1)主題分析與索引概念類型；(2)主題資訊的類型與(3)運用的索引方法等三項，參見圖 3，以下分述之。

1. 主題分析與索引概念類型

主題分析與索引概念分為三種：(1)單純化(Simplistic)的概念；(2)以內容導向(Content-oriented)的概念；(3)以需求導向(Requirements-Oriented)的概念：

(1). 簡單化的概念(Simplistic Concept)：將主題視為客觀實體(Objective entity)的角度來切入，可將文獻中的抽象語言，利用數理的統計推導方法，將其概念簡化，並用數字的方式呈現，若根據這項概念，文獻在進行索引時，以萃取(Extraction)詞彙的方法與原則，應可完全採以自動化的方式完成，但其缺點是無法得知文獻之內隱意義。

(2). 以內容為導向的概念(Content-oriented Concept)：以內容為導向的觀點，保有將主題視為客觀實體(Objective entity)的思維，另加上文獻內容的解釋層面，亦指文獻內容的語義隱含其他意涵，並非文字上的直接意義，不能完全以統計推導方式獲取主題詞，透過人工方式實際閱讀文獻，了解其內容，始能進行主題分析與索引。

(3). 以需求為導向的概念(Requirements-oriented Concept)：認為主題只是一種知識轉換的手段，藉此檢索其所需要的資訊或知識，故文獻作為知識傳遞的一種的產出，主題詞做為任何人對於某種特定領域的文獻感興趣時，藉以查找的中介媒體。

2. 主題資訊類型：

主題資訊類型分為外顯(Explicit)與內隱(Implicit)兩類，外顯之主題資訊，其意義在於為可清楚且客觀的陳述資訊，並以文字、數字或圖形等方式進行編碼，易於記錄、儲存、傳播與分享，乃結構化的資訊類型，相對於外顯類型的主題資訊，內隱的主題資訊，其意義具有間接或不直接表達的的意涵，必須透過個人的閱讀、理解、認知的過程後，始能判斷其所屬主題資訊。

3. 索引方法：

索引方法分為萃取(Extraction)與分配(Assignment)兩類，萃取一詞具有「抽出」之意義，應用在索引行為上，可指從文獻內容中，透過某種特定的方法或規則，取得足以識別描述該文獻之詞語描述，以做為該文獻之代表詞彙；而分配一詞，應用在索引行為方面，指將一篇文獻，分配在不同的分類體系內。

結合上述論述，對於上述三種主題分析與索引的概念，首先，若從簡單化的概念的角度切入，屬於外顯的主題資訊類型，如透過文獻之題名(Title)、摘要(Abstract)之內容，透過語法剖析、處理等等過程，取得認為能夠識別的詞語，再進行索引，此一過程並不涉及人類是否閱讀，僅從文本內容進行分析，其次，以內容為導向的概念，在主題資訊的類型方面，涵蓋內隱(Implicit)與外顯(Explicit)兩種類型，在索引方法，亦應用了萃取(Extraction)與分配(Assignment)兩者，代表該取向在內容導向的思維方面，除了從文本分析其主題資料外，文獻亦須由人類閱讀，根據文獻情境判斷其隱含在語句中的意涵，始能將其區分至不同的分類體系中。而延伸至以需求為導向的概念，在進行索引時，文獻內容是否具有外顯或內隱的資訊並非其重點，而在於思考如何使該文獻能夠被感興趣的讀者，透過檢索而發掘該文獻，如何應用結合資訊檢索系統，藉以提高

文獻能見度才是其關注重點。

Concept of Subject Analysis and Indexing	Type of subject information	Indexing method
Simplistic Conception	Explicit information	Extraction
Content-oriented Conception		
Requirements-oriented Conception	Implicit information	Assignment

圖 3 主題分析與索引概念、類型與方法之互連圖

資料來源：Albrechtsen, H. (1993). Subject analysis and indexing: from automated indexing to domain analysis. *Indexer*, 18, p220.

主題分析做為圖書資訊學的一項重要研究方向，根據上述概念，其思維以文獻為出發點，根據文獻儲存和檢索系統的需要，依據一定的分析方式對文獻內容進行分析，從中萃取出其主題概念、確定主題類型及剖析主題結構的過程，但對於學術領域的各項發展，如文獻作者、所屬機構等，屬於領域分析的範疇。

領域分析(Domain analysis)原先應用於軟體工程(Software Engineering, SE)，領域(Domain)一詞於韋氏字典之意義為「a sphere of activity or interest」與「an area of knowledge or activity」兩種解釋，前者意指「活動或感興趣的範疇」，後者指的是「對知識或活動的範圍」，在軟體工程學門，領域最常用來指稱一個應用區域或軟體系統開發範圍，領域分析的名詞最早由加州大學 Irvine 分析的資訊與電腦科學系教授 Neighbors 於 1980 年提出，其定義為「在一個特定的問題領域中，用來辨識相似系統中，物件(object)與函數(function)的活動」，意即把相似系統中，重複使用的物件與函數萃取出來，以避免重複開發的情況發生(Neighbors, 1980)。

因此，領域分析亦被視為一種協助解決系統開發與設計問題的方法，其目的是如何讓資訊被識別(identified)、截取(captured)、架構化(structured)、組織化(organized)和進一步再被利用(reuse) (Prieto-Diaz，1990)。

有別於軟體工程學門的思維，Birger Hjørland & Alberchtsen 認為瞭解資訊最佳的方法即是將知識領域(knowledge-domain)視為一種思想或論述社群(discourse communities)，是社會分工的一部份，認為舉凡知識組織、知識結構、合作模式、語言與傳播方式、資訊系統、主觀及相關準則等議題，皆可以由領域分析的觀點加以分析。故領域分析應用於圖書資訊學門，是藉由分析某個學術領域內進行的學術活動，探索該領域所關心的研究主題，了解研究人員所重視的資訊與各種資訊生產、傳播與產出與使用的情形，故為圖書資訊學界重視的一個研究方向，若單就期刊的論文發表情形，進行領域分析則可以了解期刊的重要作者及研究機構，也可以辨識出期刊內的重要的研究主題，對於期刊編輯而言，可對於收錄的論文主題加以檢視，藉以了解學科發展的趨勢，調整編輯的方向，對於作者而言，亦可了解期刊所涵的研究範疇以及提供在檢索上的參考(Birger Hjørland & Alberchtsen, 1995，轉引自林頌堅，2003)。

事實上，領域分析並非全新的論點，但對於圖書資訊學的研究發展，卻是一項有效的研究方法，由於資訊科學與圖書資訊學的理論與應用經常無法獨立於主題領域之外，Birger Hjørland 於 2002 年曾就資訊科學的領域分析，具體描述圖書資訊學家應具備的 11 項專業能力，包含：(1)建立並評估文獻指南或主題入口網站；(2)建立並評估專門的分類法與索引典；(3)研究專業領域的索引與資訊檢索歷程；(4)具備某領域使用者實證研究的知識；(5)執行並詮釋書目計量學研究；(6)進行某領域資訊架構與服務之歷史研究；(7)熟悉某知

識領域的文獻與類型；(8)了解不同典範、假定、關係之間的認識論與批判研究；(9)具備某領域術語研究、專門用語、論述分析的知識；(10)了解某領域在科學與專業知識傳播中的結構與機構；(11)具備與專業認知及人工智慧相關的領域分析知識。(Birger Hjørland，2002，轉引自國家教育研究院，2012b)

二、MeSH 醫學主題詞

醫學主題詞表(Medical Subject Headings, MeSH)(National Library of Medicine, 2013)是由美國國家醫學圖書館(National Library Medical, NLM)於1960 年首次獨立出版的控制詞彙表(Control Vocabulary)，為目前廣泛應用於醫學資訊檢索的工具之一，源自於美國國家醫學圖書館於 1954 年所出版的標題主題詞規範清單(Subject Heading Authority List)(NATIONAL LIBRARY OF MEDICINE, 2014c)，其用途在於描述各種生物醫學領域的相關文獻、書籍等進行主題編目，以利於使用者在檢索文獻時，能夠精確的檢索出所需要的相關研究文獻。MeSH 亦被世界各國翻譯成不同語言，廣受各國醫學圖書館與其他相關機構歡迎，將其用於索引與編目 (Nelson, S., Johnston, W. D., & Humphreys, B. L., 2001)，NLM 亦於 1960 年代起進行有關儲存與檢索資訊的數位計畫，如醫學文獻分析與檢索系統 (The Medical Literature Analysis and Retrieval System, MEDLARS)，開始讓醫學文獻索引與檢索的速度加快，以利於文獻的散布與擴展 (Lipscomb, C. E.，2000)，MeSH 每年為 NLM 固定出版修訂，時至 2014 年公布的版本，分為 16 個分類，計有 26,582 個主題詞，主要分類與其所屬主題詞個數，參見表 2(NATIONAL LIBRARY OF MEDICINE, 2014b)。

表 2 醫學主題詞(MeSH)分類與主題詞個數(2014 年版)

分類	標題	主題詞個數
A	Anatomy	2,928
B	Organisms	5,197
C	Diseases	11,304
D	Chemicals and Drugs	17,914
E	Analytical, Diagnostic and Therapeutic Techniques and Equipment	4,765
F	Psychiatry and Psychology	1,151
G	Phenomena and Processes	3,429
H	Disciplines and Occupations	514
I	Anthropology,Education,Sociology and Social Phenomena	652
J	Technology, Industry, Agriculture	602
K	Humanities	219
L	Information Science	520
M	Named Groups	259
N	Health Care	2,351
V	Publication Characteristics	189
Z	Geographicals	554

資料來源：National Library of Medicine. (2014). *MeSH Tree Structures – 2014*. Retrieved from http://www.nlm.nih.gov/mesh/2014/mesh_trees/trees.html

由於生物醫學領域發展進步快速，隨著時代的演進，新舊觀點的更迭，讓主題詞與其用法也因應時代朝流而必須持續更新，使得在其結構上擴展的相當快速，自 2001 年後的整體架構，將醫學主題詞表的架構進行更多的修正與擴充，讓包括了術語 (Term)、概念(Concept)與描述符(Descriptor)等物件(Object)的集合，在描述彼此關係時更加的明確，對於編目者與使用者，即可透過簡潔的概念即可描述特定的研究文獻，在實際應用於指涉的文獻時，能夠更清楚表達領域知識之間的關聯。

醫學主題詞表的結構在經過擴充與修正後，它迅速了擴大了詞彙量，並應用於資訊檢索，結合了統一醫學語言系統(Unified Medical Language System，UMLS)之中的泛索引典(Metathesaurus)的概念，應用在針對醫學文獻的主題編目層面，結合兩者而開發出半自動化索引(Semi-automated indexing)的系統，使其讓收錄的文獻能更快速的進行主題編目。

MeSH 醫學主題詞的組成，主要分為三個部份：(1)主題詞本身(Headings)；(2)副標題(Subheading)，亦稱為限定詞(Qualifiers)與(3)補充概念(Supplementary Concept Records, SCRs)，其分述如下。

主題詞(Headings)是醫學主題詞表中的最主要核心，其主要目的為描述文獻或是書籍與某種概念「有關」(about)，亦可言之，主題詞的存在，是提供讀者對於文獻與書籍做某種初步的理解與思考，讓他們知道該文獻或書籍是歸屬於某種概念之內，在實際應用方面，MEDLINE 資料庫將主題詞視為索引詞，反映出概念之意義，亦表示在文獻或書籍主要討論之範疇。

限定詞(Qualifiers)則有不同類型，主要是為主題詞所指示出的概念範疇，再做進一步的限縮，在 PubMed/MEDLINE 資料庫中，是採用副標目(Subheading)一詞，亦為在前述主題詞所標示的範圍中，限制在其中某些特定的概念，目前共有 83 個限定詞(NATIONAL LIBRARY OF MEDICINE, 2014f)。

補充概念(Supplementary Concepts)為 MeSH 每天動態增加的補充紀錄，它應用在 MEDLINE 資料庫的紀錄內，多半都是應用在化學與藥品(Chemicals and drug)層面的主題詞。

PubMed 資料庫對收錄的醫學文獻，若 MEDLINE 進行收錄與索引(index)則會根據文獻的特性，透過專家學者給予適宜的主題詞，在分類過程中，若認

為主題詞不足以完全分類與區隔該文獻的特性，則加上不定個數的限定詞，主題詞與限定詞之間以斜線符號(/)做為區隔，若專家學者認為該主題詞或限定詞其中之意涵足以視為該文獻的重要主題(Major Topic)，則會在該主題詞或限定詞之前或後加上星號(*)做為識別，並著錄於文獻的書目資料內，故使用者透過 PubMed 資料庫檢索至該篇文獻時，即可透過書目資料，得知該文獻的主題詞分布，以限定或擴展其資訊檢索的行為。

目前 PubMed 資料庫提供七種不同的樣式，顯示不同格式的書目資料，分別為(1)一般樣式，網頁顯示(Summary)；(2)一般樣式，但以純文字顯示簡要資訊(Summary,text)；(3)摘要樣式(Abstract)；(4)摘要樣式，但以純文字顯示簡要資訊；(5)MEDLINE 樣式；(6)XML 樣式；(7)文獻編號(PubMed ID, PMID)清單格式，其以(1)一般與(5)MEDLINE 兩種樣式會顯示文獻的主題詞，但在顯示主題詞與限定詞時，其顯示方式仍有其不同處，在匯出書目資料後，須注意其著錄格式，始能進行資料剖析，本研究採以 MEDLINE 樣式進行文獻資訊的剖析與截取，下述以醫學文獻“*A variation in the value of colectomy for cancer across hospitals: mortality, readmissions, and costs.*“為例 (Fox, J. P., Tyler, J. A., Vashi, A. A., Hsia, R. Y., & Saxe, J. M., 2014)，說明 PubMed 描述書目資料時，一般格式與 MEDLINE 格式在描述文獻的醫學主題詞時的區別，參見表 3。

表 3 PubMed 資料庫顯示醫學主題詞之樣式－以一般(Summary)與 MEDLINE 格式為例

樣式(Pattern)	主題詞/限定詞(含重要主題*)
一般格式	Colonic Neoplasms/economics
(Summary)	Colonic Neoplasms/mortality
	Colonic Neoplasms/surgery*
MEDLINE	Colonic Neoplasms/economics/mortality/*surgery

樣式(Pattern)	主題詞/限定詞(含重要主題*)
(欄位識別: MH -)	

說明：星號在主題詞或限定詞文字的前或後，並不影響其語意，「Colonic Neoplasms/surgery* 」與「Colonic Neoplasms/*surgery 」，其語意是相同的，但在資料處理時因位置不同，易被判定為不同的紀錄。

由表 3 顯示，一般格式對於若出現含限定主題範圍的主題詞時，若主題詞具有多個限定詞時，分別以「主題詞／限定詞」的方式著錄，而 MEDLINE 格式則以「主題詞／限定詞(1)／限定詞(2)／限定詞(3)」方式著錄，若認為該主題詞或限定詞具有足以描述為該文獻的主要特性，則加上星號(*)，就表 3 所示，雖主題詞 Colonic Neoplasms 主題詞同時有 economics、mortality、surgery 三組限定詞，但以重要程度而言，則以「Colonic Neoplasms/surgery* 」為描述最能代表 *A variation in the value of colectomy for cancer across hospitals: mortality, readmissions, and costs.* 一文的主題詞描述，故透過主題詞加上限定詞的概念，藉此讓每篇醫學文獻皆有其專屬分類。

三、MeSH 的階層結構與路徑

MeSH 採用了階層架構呈現出其相互之間的關係，架構中出現的每一個主題詞，都代表一個獨立的概念，以醫學主題詞－大腸直腸腫瘤（Colorectal Neoplasms）為例，其主題詞編號為 C04.588.274.476.411.307，若從 MeSH Term 的分類階層，Colorectal Neoplasms 屬於疾病(Diseases)的類別中，屬於腫瘤(Neoplasms)的其中一類，主題詞至少會對應到一個以上的主題詞編號，而階層(Level)數越小，亦指其分類概念較為模糊與廣泛(generic)，反之階層數越大，代表其分類概念較為精確與專指(specific)，參見圖 4。

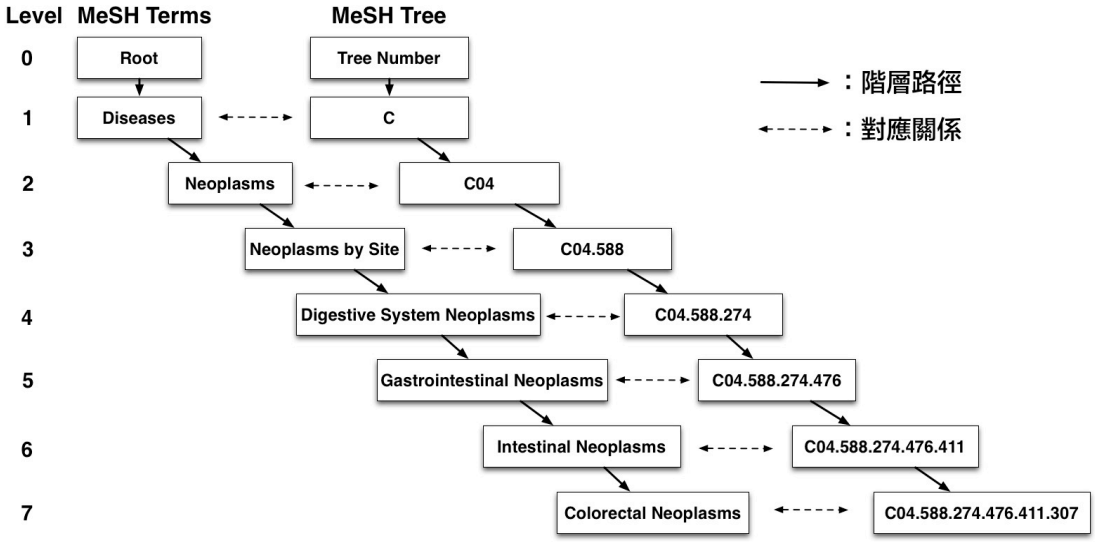


圖 4 MeSH 醫學主題詞與主題詞編號示意圖

資料來源：依據美國醫學圖書館(NLM)之醫學主題表內容自行繪製

主題詞編號的組成規則以英文字母與數字構成，再以英文句點(.)做為分隔符號，每一層固定由三個字元所組織，除編號首碼固定為英文字元外，其他字元皆為數字所構成，而編號首碼的英文字元，其用途為對應到主題分類的 16 個分類，以上圖為例，MeSH 主題詞編號(MeSH Tree Number)第一碼定義為英文字元 C，其意涵對應到主題詞中的疾病(Diseases)，其後第二階層(Level = 2)的結構，則是由首碼英文字元加上兩個數字字元所構成，亦指主題詞編號為 C04，則對應到主題詞中的腫瘤(Neoplasms)，以此類推，對於主題詞編號 C04.588，則對應到 Neoplasms by Site，以本例主題詞- Colorectal Neoplasms，其所在分類階層數為 7，故使用者只要得知特定主題詞，就可以透過其主題詞編號的階層關係，回溯每一層的分類，繪製出 MeSH Tree 的路徑(Path)。

但主題詞編號與主題詞之間並非完全為一對一的對應關係，從主題詞編號的觀點出發，由於編號具有惟一性，一個特定的主題詞編號只會對應到一個主題詞，建構出 MeSH Tree 的上下層級，但另一方面，從主題詞的角度，單一主

題詞由於意涵因素，同一個主題詞會重覆出現在不同的分類階層，亦會產生不同的路徑，故在進行醫學文獻的主題分析時，除須將文獻中的主題詞進行 MeSH Tree 的對應階層與路徑回溯外，在其處理過程中，必須注意到主題詞本身在 MeSH Tree 中的階層分類出現的次數，以 Colorectal Neoplasms 為例，在 MeSH Tree 中出現 6 次，則有 6 條路徑，目前在 MeSH Tree 分布路徑出現頻率最高的主題詞為 WAGR 症候群(WAGR Syndrome)，其出現頻率高達 20 次，其意義為在回溯與建立路徑時，若 WAGR Syndrome 一詞出現，則須建立 20 條分布在 MeSH 樹狀階層的路徑。

MeSH Tree 是藉由主題詞之間的意涵建構，形成上下關係，若將每一個主題詞視為一個節點(Node)，那麼透過路徑中每一個節點，即可了解主題詞的階層分類，以主題詞 Polyposis Syndrome, Familial 為例，如圖 5 所示，會產生路徑 A 與路徑 B 兩條分類路徑：路徑 A 會經過 Colonic Polyps，再到 Colonic Neoplasms，終點為最上層的 Intestinal Neoplasms；路徑 B 則是會從另一個主題詞 Colonic Polyps，再至上一層的 Intestinal Polyps，終點仍舊是主題詞 Intestinal Neoplasms，兩條路徑雖有部份名稱相同的主題詞，但在分類意涵上，兩條路徑的代表意義是絕不相同的。

從上述例子，可得知醫學文獻的分類結合 MeSH 的樹狀階層，其意義在於當 PubMed 資料庫的專家學者在分類眾多醫學文獻時，在由上而下的分類歷程，最終行至決定性的主題詞時，途中所經過的每一個節點，皆可以視為該文獻的廣義或狹義分類主題屬性，而從主題分析研究的角度，當醫學文獻具備主題詞做為其分類屬性，無論選擇 MeSH Tree 中的任何一個主題詞做為起始節點，往上可得知道其廣義的主題類別，而其下層的節點，皆可成為不同的文獻

資料群集，例如若將主題詞 Colonic Neoplasms 為主題分析的起始點，即可得知，在 Colonic Neoplasms 主題之下，具有 Colonic Polyps、Colorectal Neoplasms、Sigmoid Neoplasms 等三個文獻資料群集，若須再細分，則照同一方式持續向下切割文獻資料群集，直至最底層的主題詞，亦指無法再進行分類為止，參見圖 5(Mork, J. G., Jimeno-Yepes, A., & Aronson, A. R. , 2013)。

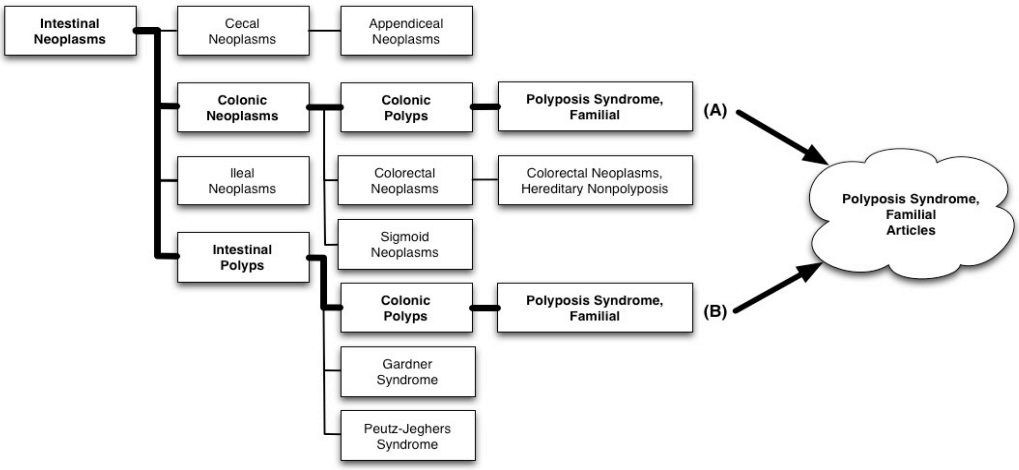


圖 5 醫學主題詞(MeSH)架構範例-以 Polyposis Syndrome, Familial 為例

資料來源：根據 Mork, J. G., Jimeno-Yepes, A., & Aronson, A. R. (2013). The NLM Medical Text Indexer System for Indexing Biomedical Literature. In *BioASQ@ CLEF*. 的內容再行繪製而成。

故上述醫學主題詞的架構可知，當某一個主題出現在不同的分類越多，其主題路徑則會越多，其複雜程度越高，過去傳統透過人工處理分類的方法，並無法批次處理大量的文獻分類，必須透過系統化與自動化的方式將文獻進行處理，始能將文獻進行分類索引，因此，美國國家醫學圖書館開始發展醫學文字索引系統，以利於醫學文獻的分類索引。

四、MeSH 自動索引-醫學文字索引系統

美國國家醫學圖書館於 1990 年代中期，開始發展醫學文字索引系統(Medical Text Indexer System, MTI) ，MTI 主要關注於主動索引(Index Initiative)，

醫學文獻主題詞在進行自動索引(Automation Index)或半自動索引(semi-automated indexing)的概念上分為四個部份：(1)資料來源：醫學文獻的標題與摘要；(2)醫學主題詞的限制(Restrict)與擴展(Extract)；(3)主題概念的分群(Cluster)與排名(Rank)；(4)增加其他的規則(Apply Rules)；(5)主題詞的呈現(Occurrence)，參見圖 6，對於醫學文字索引系統的處理流程與相關技術與概念，以下分述之(NATIONAL LIBRARY OF MEDICINE, 2014a)。

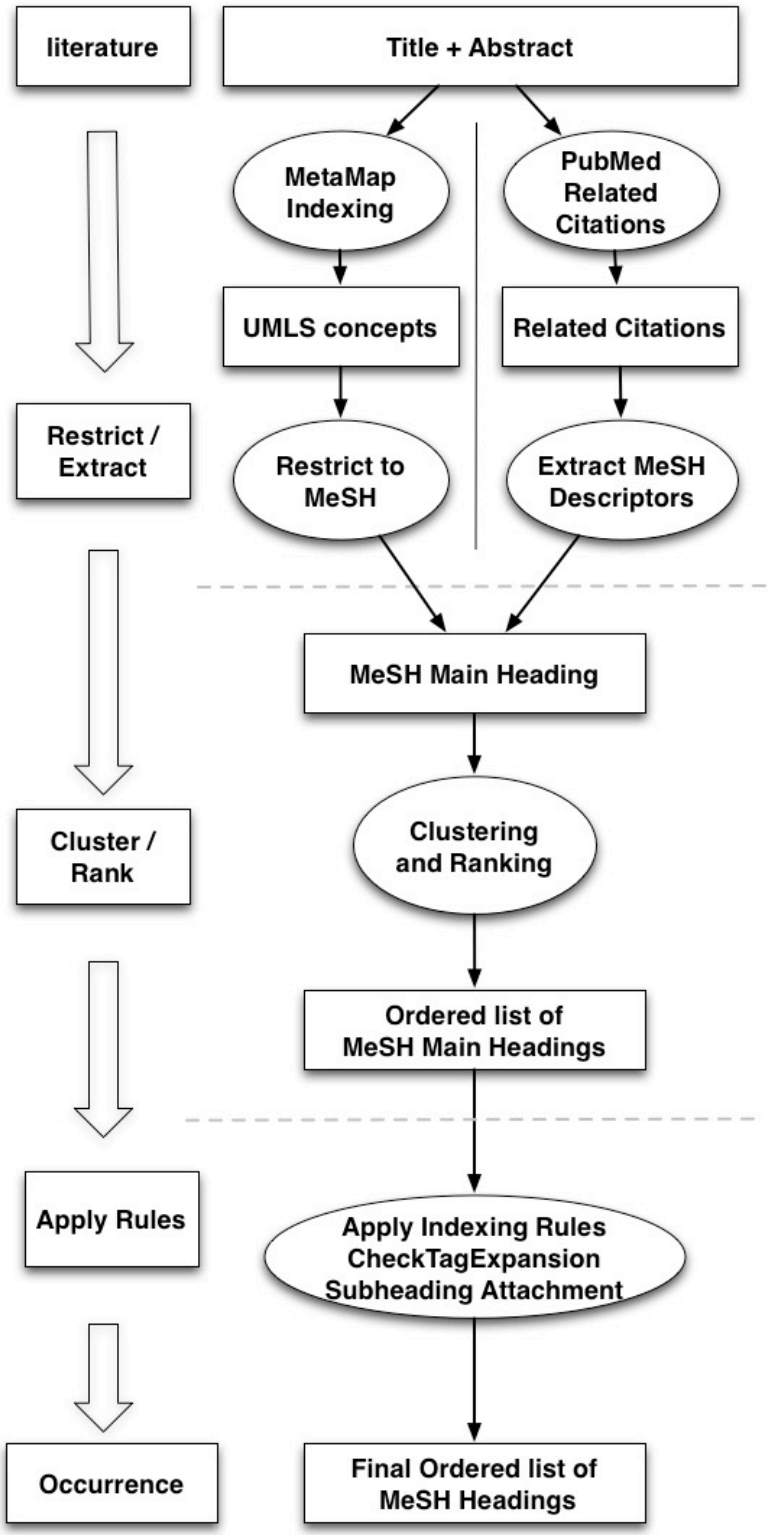


圖 6 MeSH 醫學主題詞自動索引圖

資料來源：National Library of Medicine. (2014). *History of MTI*, Retrieved from http:// http://ii.nlm.nih.gov/MTI/history.shtml，根據內容重新繪製。

從圖 6 可知，醫學文字索引系統須採用多種不同演算法、索引典與工具軟體始可完成，包括統一醫學語言系統(Unified Medical Language, UMLS)、PubMed 相關引文方法(PubMed Related Citations, PRC)、詞彙對應工具－MetaMap 等，下面針對上述三項簡述之。

（一）、統一醫學語言系統(Unified Medical Language, UMLS)

UML 著重在解決二個影響有效檢索的問題，(1)同一概念經由不同的或在不同的資料庫中，可能會有不同的表達方式；(2)資料庫系統分散所造成的檢索不完整的問題，故 UMLS 試圖在讀者問題與資料來源間建立概念上的連結，其架構分為三個元素：(1)泛索引典(Metathesaurus)；(2)語義網路(Semantic Network)；(3)專家辭典(SPECIALIST Lexican)，該三項概念簡述如下 (張慧銖、黃婷鈺，1996)。

1. 泛索引典(Metathesaurus)：

泛索引典是一種術語庫，亦可視為知識庫的一種，泛索引典是由含意(meaning)組成，其基本單元稱之為概念(concept)，此作法將使代表同一個含意的不同名稱，被連結在泛索引典的單一概念，如此可以找出有不同別名的同一概念，或找出不同概念間的關聯性。泛索引典中，保留了原索引典對於概念的定義、階層的連結及各種語意類型，並為收錄的每一個概念建立新關聯，用以連接不同索引典中各概念和詞之間的關係，將不同來源的索引典整合成一個龐大的、受控制的概念知識庫。在 UMLS 的泛索引典，是由生物醫學實務和研究領域中跨語文的多個索引典或分類表中所選出的詞條組成，即其編譯自不同來源的生物醫學命名系統的名稱、關係和聯想資訊，時至 2014 年的釋出版本-2014AB Release，共有 3,091,786 個詞，11,936,143 個概念 (NATIONAL

LIBRARY OF MEDICINE, 2013, 2014g)。

2. 語義網路(Semantic Network)：

語義網路的目的是為了在 UMLS 的泛索引典表達構念時，具有一致性的分類方式，並提供了如何表達這些概念有用的關係集合(NATIONAL LIBRARY OF MEDICINE, 2004, 2006)，語義網路的現今釋出的版本包含 135 種語義類型與 54 種語義關係，可分為實體(Entity)與事件(Event)兩大類 (NATIONAL LIBRARY OF MEDICINE, 2003)，實體指的是物理對象，如生物、物質等，而事件指的是社會活動，如行為、活動、研究過程等。語義類型是分層次的，故存在於等級關係(is-a)，語義類型之間還存在各種相關關係，包括：物理相關(physically related to)、空間相關(spatially related to)、時間相關(temporally related to)、功能相關(functionally related to)、概念相關(conceptually related to)等種類，從上述可知，語義類型可視為具有層次結構的節點，透過語義關係而將這些節點連結，依情境的差異形成富有意涵的語義網路。

3. 專家辭典(SPECIALIST Lexican)：

UMLS 收錄常見的英語單詞、生物醫學術語，與出現在 MEDLINE、UMLS 的泛索引典中的術語等，集結而成為辭典資料庫，並以 NLM 的自然語言處理工具(Natural Language Processing Tools)進行語法上的處理，每個詞條紀錄均詳述自然語言處理系統所需要的辭典資訊，包括句法、詞彙形態和語法結構的各項資訊，並提供工具以確認各種不同語言的詞彙範圍，與識別生物醫學專有名詞和文本(Context)中詞彙的詞型變化，專家辭典的建置，主要在提供各項語詞資料，以滿足對於專家對於詞彙的需求。

（二）、PubMed 相關引文方法(PubMed Related Citation, PRC)：

它是利用 K 最近鄰(k-Nearest Neighbors)方法來檢索 PubMed 資料庫中與要進行索引的待索引文獻相類似的文獻，亦即為 PubMed 相關文獻 (PubMed Related Citation, PRC)，再從與待索引文獻相似度最高的 10 篇相關文獻中的主題詞進行計算，以推論出待索引文獻可推薦採用的主題詞與限定詞等詞群(NATIONAL LIBRARY OF MEDICINE, 2014e ; Lin J., & Wilbur, W. J. , 2007)。

（三）、詞彙對應工具－MetaMap：

MetaMap 為將文字對應到 UMLS 的一種詞彙對應工具，它採用知識密集(Knowledge-intensive)型的方法，包括了符號(symbolic)、自然語言處理(Natural-language processing, NLP)與計算機語言學(Computational-linguistic)等技術，它主要是將生物醫學文獻和 UMLS 中泛索引典中的概念進行對應(mapping)，以達到將醫學文獻以半自動化或全自動化索引的目的，實務上 MataMap 常見的應用範圍包括：資訊擷取(Information Extraction)、分類法與分類(Classification/Categorization)、文件摘要(Text summarization)、問題詢答系統(Question answering)、資料探勘(Data mining)、文獻挖掘(Literature-based discovery)、新聞或文獻中的正文理解(Text understanding)、統一醫學語言系統概念的索引與檢索(UMLS concept-based indexing and retrieval)、生物醫學文獻的自然語言分析(Natural-language analysis of biomedical literature)等(NATIONAL LIBRARY OF MEDICINE, 2014d)。

上述所提及的觀念與工具，運用於醫學文字索引系統時，醫學文獻的標題與摘要是為萃取主要主題詞(Main headings)的資料來源，為求達成主題詞萃取時，概念的限制與延伸，就會採取不同的途徑分別達成其目的，以圖 6 所示，

從醫學文獻的標題與摘要取得後，採用 MetaMap 工具，運用自然語言處理規則法(Natural Language Processing, NLP)，即可從統一醫學語言系統(Unified Medical Language, UMLS)所收錄的泛索引典、專家辭典中的專業術語，透過語義網路的規則，對應到相關的詞彙，將原本語義較為廣泛與模糊的文獻標題，限縮其概念成為主要主題詞。

而另一途徑則是從 PubMed 相關引文方法，從 PubMed 資料庫檢索出已索引完成的其他醫學文獻集，再與待索引文獻兩者透過演算法進行相關文獻的計算，並找出該相關文獻中，著錄完成的主題詞，成為待索引文獻的主要主題詞。該方式時也將文獻彼此之間的關係透過主題詞而產生連結。

從 MetaMap 與 PubMed Related Citations 兩種途徑產生的主要主題詞，經過分群與排名的過程，透過精確率(Precision)與回現率(Recall)的計算與篩選，以產生排序後的 MeSH 主題詞群的備選清單，再增加限定詞的對應等規則，最後成為 MeSH 主題詞的正式清單，透過上述的過程，產生以足以識別醫學文獻特性的 MeSH 主題詞群，除了將醫學文獻進行專業的分類，並且提供 PubMed 資料庫在檢索時的便利性。

五、MeSH 主題分析相關研究

本小節描述採取以 PubMed 資料庫收錄的文獻進行主題分析的相關研究，其採取的方式為－採取特定的疾病為主要檢索詞，收集相關文獻，並進行統計分析。

B-Z Li 等人針對紅斑性狼瘡(Systemic Lupus Erythematosus)的檢索文獻進行檢索，引用 PubMed 資料庫收錄自 2002 年至 2011 年的文獻，其檢索策略為”

Lupus Erythematosus, Systemic[MeSH]”，檢索結果 PubMed 共收錄了 14,053 篇文獻，來自於 1,627 種不同的期刊，其中以 Lupus , Arthritis Rheum , J Rheumatol 等三種為出版文獻數量最多的期刊，而文獻產出最高的國家則是美國，其次是日本與中國；B-Z Li 等人認為，研究紅斑性狼瘡文獻的主要語言為英語，而非英語系發表的僅佔總數的 11.492%，另外，B-Z Li 等人在收集文獻時發現，雖然隨機對照實驗的相關文獻，在實證醫學的證據等級上是較高的，但是在統計文獻時的結果，隨機對照實驗的相關文獻所占比例相對較少，在主題詞分析方面，以病因(etiology)與生理學(physiology)為較主要研究的領域，而傳播(transmission)與獸醫學(veterinary)則是相對較少(Li, Pan, & Ye, 2013)。

José M Ramos 等人對於利甚曼病(leishmaniasis)的研究進行相關論文的計量分析，其資料來源透過檢索 PubMed 資料庫取得，檢索策略為” Leishmania [MeSH] or leishmaniasis [MeSH]”，限定文獻出版時間為 1945 年至 2010 年，且並不限定文獻出版語言與類型限制。其檢索結果共有 20,780 篇文獻，再依據每 5 年做為出版區隔，其研究結果，在 MeSH 主題詞的統計方面，在 1945 年至 2010 年間的主題以 Animal 為最多，有 8,564 篇文獻，占總數的 41.2%，其次是 Leishmaniasis. Vsceral，收錄了 6,216 篇文獻，占總數的 29.9%，文獻類型也以期刊文獻為最多，收錄了 17,982 篇文獻，占總數的 86.5%；如出版地區為區隔，以歐洲為最多，出版了 31.7%的文獻，其次是拉丁美洲與加勒比海地區，有 24.5%；José M Ramos 等人並認為，雖文獻多數採以英文為發表語言，但以法文與葡萄牙文的文獻可能更為重要，因為利甚曼病的流行是以北非、法國和巴西為主，因為語系的不同可能可以解釋因為疾病的流行區域 (Ramos, González-Alcaide, & Bolaños-Pizarro, 2013)。

牛道立等人針對中國放射腫瘤學領域 2004 年至 2008 年間被 PubMed 收錄的文獻進行分析，其檢索策略為：(1)以”Radiotherapy”為 MeSH 主題詞檢索；(2)發表時間從 2004 年至 2008 年；(3)第一作者的的單位和地址(Affiliation)限定為中國。其檢索結果如無限定第一作者的單位和地址，則原始檢索結果共有 23,299 篇文獻，而加上限定作者與地區性的條件，則中國放射腫瘤學論文在 2004 年至 2008 年間數 PubMed 收錄的文獻共 550 篇，分散於 160 種期刊，該研究依據(1)文獻在 MeSH 樹狀結構表；(2)文獻的 MeSH 副標目；(3)文獻研究領域；(4)文獻發表期刊；(5)文獻的出版時間；(6)期刊的 SCI 影響因子；(7)文獻發表的地區等因素進行統計分析，其結果認為其內容主要集中在「放療劑量」和「計算機輔助的放療」(牛道立、甄俊杰與何芬，2010)。

王敏等人針對 PubMed 資料庫收錄自 1948 年至 2009 年 3 月為止，有關人體肝移植研究的文獻，其檢索策略為：(1)以”Liver Transplanation”的 MeSH 主題詞；(2)文獻標題為”Liver Transplanation *”開頭或”hepatic transplantation *”開頭的文獻；(3)限定條件為”Human”，其檢索結果共有 28,452 篇文獻，並針對於人體肝移植研究論文的(1)時間；(2)國別；(3)機構；(4)期刊等四項因素進行統計，其結果顯示，PubMed 收錄的 28,452 篇文獻中分布在 2,002 種期刊，其中 20 種期刊刊載論文數占總論文數的 50.72%，顯示人體肝移植論文的來源期刊比較集中。其研究建議並認為採用文獻計量法研究主題分析，有助於把握國際研究總體趨勢(王敏等人，2010)。

翟俊霞等人利用 MEDLINE 資料庫檢索自 1982 年至 1998 年收錄全部 ras 相關論文，其檢索策略為”Genes, ras OR (Oncogenes AND @ras)”，共有 7,724 篇文獻，其分析方法主要將 17 年的文獻分 4 個時間區間，再依據 ras 基因研究

文獻的 MeSH 主題詞與副標目進行統計，並選擇累積出現次數達 30 次的主題詞為核心主題詞，共有 150 個核心主題詞，另外再針對”Genes, ras”主題詞下，在 7,724 篇文獻中，將有副標目的 2,126 篇文獻進行統計，其研究結果顯示，ras 基因研究涉及眾多的腫瘤主題詞，表示許多腫瘤的發生與 ras 基因相關，而副標目部份，ras 基因研究中的 5 個副標目中，以遺傳學的比例最高(1,328 篇)，其次為藥物作用、生理學、輻射效應，最後是免疫學，顯示 ras 基因研究的文獻中，與遺傳學領域較有相關（翟俊霞、劉玉華與安翠平，2001）。

肖燕秋等人以 PubMed 資料庫為資料來源，檢索 2009 年 3 月至 2010 年 8 月有關甲型 H1N1 流感相關文獻，其檢索策略為”influenza A virus, H1N1 subtype”，並限定條件以 MeSH 主題詞為主，其檢索結果為 2,647 篇文獻，以數量變化、國家地區、語種等進行分析比較，研究結果顯示：(1)H1N1 的相關文獻到 2010 年 1 月為最多，有 253 篇，直至世界衛生組織於 2010 年 8 月宣布流感流行已經結束，文獻數量逐漸遞減；(2)文獻的語種分布方面，共有 23 種語言，最多的前五名是英文、法文、西班牙文、中文與德文，其中發表語言為英文的文獻有 2,351 篇，顯示以 PubMed 資料庫收錄的文獻，仍以英文發表為大宗；(3)文獻類型仍以期刊論文為主要文獻類型；(4)國家分布仍以美國為最多(1,172 篇)；(5)副標目的分類以生理學(981 篇)為最多，其次是免疫學(515 篇)（肖燕秋、徐奎、李世莉與陳克俠，2011）。

紀征等人於 2010 年針對人禽流感研究進行文獻計量分析，其資料來源來自於 PubMed 資料庫，透過檢索 2004 年至 2008 年的相關文章，對其發表年限、文獻種類、期刊類別、作者與國家地區等因素進行統計分析，其檢索詞為”Influenza in birds”，並限定與人類(human)相關，檢索結果共有 1,169 篇文獻，

分散在 400 種期刊內，其中 942 篇文獻由英文發表，其次是法文(67 篇)，而文獻作者國家地區，以美國為最多(452 篇)，其次為英國(192 篇)，而文獻類型以期刊文獻為主(585 篇)，其次為 Review(207 篇)，在主題分析方面，在副標目的分類統計，在出現次數最高的 15 個副標目，以 Epidemiology 為最多(364 篇)，其次為 Prevention and Control(220 篇)，其他則是 Transmission(183 篇)與 Virology(159 篇)，顯示在人禽流感研究的文獻中，較關注上述四項領域的研究（紀征、任淑敏與朱啟貞，2010）。

龍莉艷等人針對肺癌研究進行實證文獻計量分析，透過檢索包括 PubMed 在內的多種資料庫找出相關的文獻，其中 PubMed 資料庫收錄的年代從 1978 年至 2008 年，共 102,922 篇文獻，在基礎研究方面有 14,777 的文獻量，而在臨床研究方面則有 94,937 篇的文獻，其中 PubMed 提供的高證據級別的研究文獻，內容以治療和診斷為研究重點，其次為流行病學；在副標目方面，以診斷、治療與病因學為大宗，皆出現達超過 4 萬次以上，其次為細胞學、病理學、生理學等領域，其研究結論認為，選擇有代表性的原始研究文獻資料庫與二次研究文獻資料庫，將實證醫學的證據級別分類導入文獻的量化分析裡，藉此實現在肺癌文獻中，不同特徵的量化分析，可以分析出肺癌的研究現狀與方向，並建議應用共詞分析與集群分析等分析方法，針對各項主題進行深度分析（龍莉艷、張桂云、周丹、劉麗華與馮小東，2008）。

許昌泰等人針對麻醉學研究文獻進行統計分析，以確定其文獻增長規律與發展趨勢，為研究麻醉學相關議題提供依據。研究者透過 PubMed 資料庫檢索麻醉學(anaestheisa)領域，並以每五年做為一檢索區段，從 1866 年至 2009 年，累計共收錄麻醉學文獻共 281,326 篇，文獻類型包括了 Review, Randomized

Controlled Trial, Practice Guideline, Meta-Analysis, Letter, Editorial, Clinical trial, Other 等類型，其中以隨機對照實驗(RCTs)類型的文獻為最多(22,528 篇)，其次為 Reviews(21,710 篇)；文獻語言以英文為大宗，占總數的 78.81%，在文獻量分布方面，從 1945 年後成長速度較為明顯，在 1941-1945 年有 64 篇，到 1946-1950 年增至 2,768 篇，至 1966 年後，以每五年做為時間區間，其出版文獻均超過萬篇，至 2001-2005 年達到 45,321 篇，為各時間區間最高；而依據期刊分類，在 PubMed 資料庫收錄的 24 種期刊中，以 Anesthesiology, Anaesthesia , Anesth Analg 三種收錄文獻為最多；其研究建議中，研究者認為，文獻計量學在醫學領域中的應用，隨時間的發展也在呈增長的趨勢發展，但有關醫學文獻計量學的論文量占醫學論文數量的比例仍相當低（許昌泰、王明華與徐禮鮮，2010）。

張晗等人對於消化性潰瘍文獻進行分析研究，從 PubMed 資料庫檢索有關消化性潰瘍的文獻，其檢索策略為”pepticulcer[majr]”，限定出版年份為 2004 年至 2006 年，檢索結果共有 1,559 篇，並將出現次數超過 20 次以上的認定為高頻主題詞，共產生 65 個高頻主題詞，其中出現次數超過 100 次的主題詞為「抗潰瘍病藥/治療應用」(181 次)、「螺杆菌, 幽門」(159 次)、「螺杆菌感染/藥物療法」(131 次)、「胃潰瘍/藥物治療」(126 次)、「胃潰瘍/預防和控制」(105 次)，再依據共詞的群集分析方法進行分析，最後得到「消化性潰瘍的藥物療法」、「化學物質誘導消化性潰瘍的病因學研究」、「消化性潰瘍的外科手術」、「幽門螺杆菌感染的藥物治療」、「消化性潰瘍的實驗室診斷」及「消化性潰瘍的流行病學和遺傳學研究」等 7 個研究趨勢（張晗、劉鵬年與崔雷，2007）。

徐路透過 MEDLINE 檢索貴陽醫學院的發表論文，其檢索策略為”GuiYang

Medical College, GuiYang Medical University”為檢索主題詞，並限定至 1996 年至 2005 年發表的文獻進行統計分析，總共有 126 篇文獻被 MEDLINE 收錄，其中中文文獻為 88 篇，英文文獻為 38 篇，外文文獻的發表期刊以 Chin Med J 為最多，共有 8 篇，其次是 World J Gastroenterol(5 篇)，收錄的學科與系所單位主要在著重在感染科(21 篇)、中醫中藥科(18 篇)、病理學(11 篇)與分子生物學與實驗中心(11 篇)，該研究試圖從 MEDLINE 是否收錄該單位的文獻，判斷醫學發展的水準（徐路，2007）。

由上述相關研究可以得知，MeSH 主題研究以特定疾病做為檢索詞，再進行各項統計，包括：(1)主題詞與副標目；(2)出版年代；(3)出版地區；(4)出版語言；(5)文獻成長等。

第五節　書目計量學

一、書目計量學定義

書目計量學(Bibliometrics)一詞為 1969 年 A. Pritchard 所提出，其定義為「書目計量學是將數學統計方法，運用到圖書和其他通訊媒體上的科學」(Pritchard, 1969，頁 348，轉引自何光國，1994，頁 3)。明確將過去對於相關研究歸納出其範圍與賦予其意涵，其定義隨著時間演變而各有其差異，Sengupta(1985)認為「書目計量學為一種可以利用數學、邏輯的統計對所有傳播形式的出版品進行

計算、分類及量化的評估」(Sengupta, 1985，轉引自蔡明月，2003，頁 5)；何光國（1994）認為「凡利用數學、統計學、和邏輯學的理論和方法，對各類型文獻的本質和結構，作數量、品質、和運用上的研究與分析者，通稱為文獻計量學」。定義雖略有差異，但歸納其定義的重點，在於將文獻書目從傳播知識的載體，轉變成為其可採數理統計的計算單元，再進行延伸的相關分析與研究方法，再以量的方式輸出與賦予其解釋。

故書目計量學成為對於研究分析學科知識發展的重要工具。隨著資訊科技進步，書目計量學陸續開創新的計量方法，一般認為，書目計量學的重要定律為布萊德福定律、齊夫定律與洛卡定律等三定律，應用於圖書資訊學的文獻分佈與館藏發展等不同層面，其定義分述如下：

（一）、布萊德福定律(Bradford's Law)：

根據某學科各期刊所含相關文獻數量的多寡寡，依遞減次序排列，這些期刊將可分成為核心區(即第一區)與接連數區，且每一區包含大約相同數量的文獻，則各期刊種數的比例將呈現 $1：n：n^2$ 的關係，其意義在於特定學科之最相關文獻集中於少數的核心期刊之中 (蔡明月，2003，頁 12)。

（二）、布萊德福-齊夫定律(Bradford-Zipf's Law)：

布萊德福-齊夫定律是結合布萊德福定律與齊夫定律延伸發展而得。齊夫定律又稱為「字詞分布定律」，研究在文獻中詞頻和排名之間的關係。發現字詞的使用次數(f)與字詞的使次數排名(r)的乘積，會等於一常數 C。此定律可應用於資訊檢索，使用低頻詞與中頻詞作為控制詞彙，得以提高檢索效益，Brookes 認為布萊德福分布與齊夫定律相似，因而將兩定律加以合併，進而修正發展出

布萊德福-齊夫定律(Bradford-Zipf Law)，該定律主要是以一標準 S 形曲線來表示期刊文獻之分佈狀況，S 曲線分為三個區域，第一區為非線性起始部份，即為核心期刊部份；第三區為直線分佈；第三區則為偏垂現象，表示有關研究主題的期刊論文已擴散至多數期刊 (蔡明月，2003，頁 149)。

（三）、洛卡定律(Lotka's Law)：

洛卡定律又稱為倒數平方律，主要是以最小平方法研究作者與發表論文篇數之間的關係。洛卡在其研究指出，在採以每一篇文獻的 Senior Author 為計算基準，發表 n 篇論文的作者總數，是發表一篇論文作者總數的 n 的平方分之一。其公式為 $f(x)=c/x^n$ ($1.2<n<3.8$, c 為常數)，根據洛卡的發現，只發表一篇論文的作者總數，約是全部作者總數的 60.79% (Lotka, 1926)。洛卡定律常被應用於評估某學科領域內，不同作者生產能力、學術地位等。

透過上述三項定律，書目計量學成為圖書館學、資訊科學及學術傳播的重要分支研究領域，豐富了學術產出的多樣性。

二、國內作者書目計量文獻分析

Hsu and Chiang (2015) 以服務科學、管理與工程(Service Science, Management , Engineering, SSME)領域，透過 Web of Science 引文索引資料庫蒐集自 1991 年至 2012 年共 22 年間，共 4,513 篇文獻進行分析，並按照出版品類型、出版地區、合著作者、機構名稱等變數進行統計，以及計算關鍵詞的頻率，並執行 K-S 以檢驗作者的學術生產力是否符合洛卡定律，其分析結果顯示(1)對於 SSME 領域，最相關的學科為商業經濟學(Business Economic)、資訊科學與圖書館學(Information Science and Library Science)與電腦科學(Computer

Science)；(2)洛卡定律適用於服務科學與服務科學兩個領域，但不適用於服務管理領域，其原因為某篇文獻的作者數過高，影響其驗證結果。

林偉翔與林雯瑤（2012）以北市醫學雜誌為研究個案，分析 2004~2010 年的文章與引用文獻，探討其刊登的論文特質與主題分布、作者合著狀況與高生產力作者等作者特性與其引用文獻特性，共分析 639 篇學術論文與其 12,185 筆引用文獻，結果顯示合著作者的情況有顯著成長的趨勢。

楊喻翔與釋惠敏（2011）以安寧療護研究為主題，透過 SCIE 與 SSCI 蒐集自 1952 年至 2009 年期間共 6,828 篇研究文獻，分析其文獻的數量成長、機構及個人學術生產力，並透過布萊德福定律，找出核心期刊。其研究結果顯示：(1)研究文獻量呈現向上趨勢；(2)其期刊文獻分布符合布萊德福定律，核心期刊為 5 種期刊。

林巧敏與范蔚敏（2010）蒐集台灣地區自 1959 年至 2008 年有關檔案與文書研究文獻共 1,257 篇，分析文獻的數量成長、主題分布、第一作者與期刊分布等特性，其研究結果顯示：(1)檔案研究文獻數量成長為隨著時間化呈現不規則的成長曲線；(2)文獻研究主題以「檔案與歷史研究為主」；(3)作者分析顯示檔案學的作者相當分散，且文獻作者合著比例偏低，但有合著增加趨勢；(4)檔案期刊數量分布，不符合布萊德福定律。

楊燕枝（2010）蒐集 Web of Science 收錄的 13 本科技管理領域的期刊，檢索 1990 年至 2009 年間台灣學者在科技管理領域發表之文獻 265 篇，分析其文獻分布、作者生產力特性，並以洛特卡定律驗證其學術生產力，研究結果顯示，台灣學者在科技管理領域之學術生產力並不高，集中於少數作者。

林巧敏（2009）蒐集臺灣地區 1982~1998 年資訊組織文獻 610 篇與

1971~1998 年資訊組織學位論文 113 篇，分析文獻的數量成長、主題分布、第一作者特性、期刊分布與著作機構分布等特性，採用布萊德福定律驗證其文獻分布的正確性，其研究結果發現期刊文獻數量分布並不符合布拉德福定律，且合著比例偏低。

黃慕萱與賴麗香（2009）以國科會於 1995~2004 年核定圖書資訊學門 168 件專題研究為主體，探討圖書資訊學門之專題研究計畫數與成長、執行單位、生產力與主題等特性，以了解研究計畫的特性及主題發展。

王明玲、杜立中與曾彩娥（2009）採用書目計量方法，分析「五四運動論著目錄」書中的 1,567 筆文獻，探討臺灣 1949~2008 年所出版之五四運動文獻，分析其資料類型、出版年代、主題範圍與作者生產力的概況，並應用布萊德福定律，找出核心期刊，並分析其特質。

蔡明月與賴芊卉（2007）透過 Web of Science 蒐集 1985-2005 年與資訊科學相關之書目文獻 2,270 篇，運用布萊德福定律(Bradford’s Law)和布萊德福-齊夫定律(Bradford-Zipf’s Law)分析資訊科學引用文獻(citing literature)與被引用文獻(cited literature)的文獻分散現象，並比較其差異性，且利用 LISA 資料庫敘述語(descriptor)欄位量化分析引用文獻與被引用文獻的主題分布及變化。

三、國外書目計量學相關研究

Pilsczek (2015)認為傳染病的醫學相關研究對於貧窮國家是相當重要的，透過公開免費的醫學文獻資料庫，對於貧窮國家有助了解與分析傳染病的相關研究，以阿富汗為例，分析 2002 年至 2011 年傳染病相關研究之文獻共 294 篇，並分析其主題類型，顯示該地區的傳染病相關研究主題包括(1)呼吸道感染(Respiratory Infections)；(2)寄生蟲(Parasites)；(3)腹瀉(Diarrhoea)；(4)肺結核

(Tuberculosis)；(5)人體免疫缺陷病毒(Human Immunodeficiency Virus)；(5)耐多藥細菌(multi-drug resistant bacteria)；(6)脊髓灰質炎(Polio)；(7)利什曼原蟲(Leishmania)；(8)瘧疾(Malaria)。

Rajgoli and Laxminarsaiah (2015)分析航太科技(Spacecraft Technology)領域相關研究之作者生產力，選定(1)Journal of Spacecraft Technology;(2)International Journal of Satellite Communications and Networking;(3) Journal of Spacecraft and Rockets 等三種航太科技期刊為樣本，蒐集 2001 年至 2011 年期間，共 1,907 篇文獻，共 4,355 位作者，分析其合著狀況，並採用洛卡定律進行作者生產力的分析，其研究結果顯示洛卡定律並不完全符合，故建議可採取數量更大的樣本，以得出更佳的研究分析結果。

Huang et al. (2015)以 PubMed 資料庫蒐集 42,192 篇中醫領域(Traditional Chinese Medicine,TCM)的研究文獻，分析中醫文獻的文獻成長、文獻類型、出版國家/ 區域、出版語言，其分析結果認為，從 1995 年至 2014 年止，中醫文獻的研究產出逐步增加，英文雖然仍佔出版語言的主導地位，但出版國家多數來自中國，其次為日本、美國與台灣，但文獻成長的增加趨勢，顯示中醫領域在醫學研究中仍富研究潛力。

Sachithanantham and Raja (2015)以 PubMed 蒐集自 1950 年至 2014 年共 495 篇有關在狂犬病在印度的研究文獻，分析其文獻增長、印度在狂犬病領域的貢獻程度、作者生產力與合著作者的合作模式，並以布萊德福定律驗證期刊分布狀況，其研究結果布萊德福定律不適用於印度狂犬病相關研究的期刊分布。

Pinto, Isabel Escalona, Pulgarín, and Uribe-Tirado (2015)分析伊比利美洲(Ibero-America)地區的研究者，在資訊素養(Information Literacy)領域的學術生

產力，透過 Web of Science 與 Scopus 兩大索引摘要資料庫進行文獻資料收集，共有 105 種期刊，共 340 篇文獻，採用布萊德福定律、洛卡定律進行驗證，其結果顯示：(1)從 2005 年至 2011 年間，每年約有 30%的趨勢呈指數成長；(2)西班牙與巴西為該區域中最富學術生產力的國家；(3)期刊分布與作者生產力分別符合布萊德福與洛特卡定律。

Dehdarirad, Villarroya, and Barrios (2015)以 Web of Science 蒐集 1991 年至 2012 年期間，以女性在科學與高等教育(Women in science and higher education)的參與為研究個案，並以(1)出版生產力(Publication productivity)、(2) 學術與科學中的性別議題(Issues related to gender in academic and science)與(3)性別偏見的影響因素(Factors related to gender bias)為檢索主題，檢索出 595 種期刊，共 1,415 篇文獻，分析其文獻成長、文獻分布等特性，研究結果顯示：(1)文獻成長模式符合指數成長模型；(2)作者生產力亦符合洛特卡定律；(3)文獻分布符合布萊德福定律，可分為 4 區。

Patra and Mishra (2006)以 PubMed 蒐集自 1990 年至 2004 年期間，16,471 篇生物資訊學(Bioinformatic)領域的文獻，分析其文獻成長、文獻語言、文獻分類、出版國家、作者生產力等特性，其研究結果顯示：(1)以布萊德福定律公式，從 1,806 種期刊，找出 20 種核心期刊；(2)作者生產力亦符合洛特卡定律；(3)美國為出版國家最多的(42%)；(4)合著作者方面，以單一作者為作者分布比例最高(23%)，其次為 2 位作者(21%)。

表 4 書目計量學相關研究－資料來源、分析方法、資料筆數

作者(年代)	個案領域	分析方法	資料來源	資料筆數	年代區間

作者(年代)	個案領域	分析方法	資料來源	資料筆數	年代區間
Hsu & Chiang(2015)	服務科學、管理與工程(SSME)	洛卡定律	WoS	4,513	1991-2012
林偉翔、林雯瑤(2012)	醫學	作者生產力、引文分析	北市醫學雜誌	639 篇(文獻)、12,185 篇(引文)	2004-2010
楊喻翔、釋惠敏(2011)	安寧療護	布萊德福定律	WoS	6,828	1952-2009
林巧敏、范蔚敏(2010)	檔案與文書研究	文獻成長、布萊德福	中華民國期刊論文資料庫	1,257	1959-2008
楊燕枝(2010)	科技管理	文獻分布、作者生產力、洛卡定律	WoS	265	1990-2009
林巧敏(2009)	資訊組織	文獻成長、主題分布、布萊德福。	中華民國期刊論文資料庫	610(期刊論文)、學位論文(113)	1971-1998
黃慕萱、賴麗香(2009)	圖書資訊學	主題分析	國科會專題研究計畫查詢系統	168 件	1995-2004
王明玲、杜立中、曾彩娥(2009)	五四運動	布萊德福定律	五四運動論著目錄(書)	1,567	1949-2008
蔡明月、賴芊卉(2007)	資訊科學	布萊德福定律、布萊德福-齊夫定律	WoS	2,270	1998-2005
Pilsczek(2015)	傳染病	主題分析	PubMed	294	2002-2011
Rajgoli and Laxminarsaiah(2015)	航太科技	洛卡定律	期刊	1,907	2001-2011

作者(年代)	個案領域	分析方法	資料來源	資料筆數	年代區間
Huang 等人(2015)	中醫	主題分析	PubMed	42,192	1995-2014
Sachithanantham and Raja(2015)	狂犬病	布萊德福、作者生產力、合著作者	PubMed	495	1950-2014
Pinto 等人(2015)	資訊素養	布萊德福定律、洛卡定律	WoS、Scopus	340 篇；105 種期刊	1985-2013
Dehadarirad 等人(2015)	女性在科學與高等教育	文獻成長、布萊德福定律、洛卡定律	WoS	1,415 篇；595 種期刊	1991-2012
Patra and Mishra(2006)	生物資訊學	布萊德福定律、洛卡定律	PubMed	16,471 篇；1,806 種期刊	1990-2004

資料來源：本研究整理

總結相關研究如表 4 所示，從事書目計量學的研究限制為資料量的多寡，常見的限制與篩選條件，包括(1)該主題領域相關期刊種類數目；(2)文獻出版總篇數；(3)出版年代分布；(4)出版語言；(5)人口地理區域變數(性別、人口、國別等變數)；(6)書目資料來源；(7)書目資料取得方法等因素，使得研究者取得計量之文獻樣本有限，以致於影響統計分析結果。

而研究者自選定研究個案後，從書目資料的來源進行探討，資料的代表性與完整性是首要要件，以電子資源為例，採用文獻索引資料庫做為書目資料的取得來源，為現今書目計量學研究者常見的方式，如 Web of Science、Scopus、PubMed 等資料庫，Web of Science 與 Scopus 兩者皆提供文獻與其參考文獻，PubMed 則專注於生物醫學文獻，雖未有提供參考文獻的連結，但提供相似文

獻的機制，皆提升文獻的廣度與深度。

第六節　書目探勘

一、書目探勘

Bibliomining 為 Nicholson 與 Stanton 於 2003 年首先提出，在 Bibliomining 一詞提出之前，相關研究文獻是以「資料探勘於圖書館(data mining in library)」，將資料探勘視之為圖書館眾多工具技術中的一種應用，偏重於圖書館與資料探勘技術。然而圖書館運用資料探勘技術是要以圖書館作業相關資料為核心，這也是圖書館書目計量著重的地方，因此，Nicholson 與 Stanton 提出以「Bibliomining」此一新辭以詮釋「運用統計與型樣辨識工具於大量圖書館系統相關資料，藉以協助圖書館訂定決策或調整其服務」，亦或是「資料探勘、書目計量、統計與報告的工具的組合，運用於圖書館系統中，以萃取出使用行為事實的型樣」 (Nicholson, 2003，轉引自謝建成，2010，頁 3)。

書目計量是一種以文獻為基礎(document-based)學術傳播之量化研究，主要是由書目資料(bibliographic data)中探索出某一領域論著作與文獻引文間之關係。書目計量資料包括論著作者、引用文獻、領域主題、登出期刊等，藉由這些資料可以讓研究者知道論著產生之脈絡、引用論著之影響等。

而圖書館服務主要是以讀者為對象，為了瞭解讀者利用圖書館資源的行為模式，可以協助圖書館增進其服務品質。在企業行銷運用範疇，資料探勘技術被廣泛應用在探討客戶關係管理諸多議題，如顧客忠誠度、顧客購買行為模式等，以提供企業有效行銷之決策參考。因此利用資料探勘技術發掘讀者使用圖

書館資源的模式，不僅可以提供個人化讀者資訊或是提供其資訊服務，亦可協助圖書館經營決策之參考。

除針對個別讀者之外，或是基於讀者個人隱私維護，圖書館對於讀者社群，如不同年齡族群、不同居住地、不同學院系等，服務需求與資源使用模式仍必須予以重視。防範個人隱私最基本的方式是透過人口統計代理者(demographic surrogate)替換讀者個人，以避免特定讀者資料的揭露。而人口統計代理者主要就是以其具有共同特徵，如同一學系、同一性別等所構成的群體集合來代替個人。透過資料探勘技術運用，可以得知讀者社群如何使用圖書館資源而保有基本讀者隱私。

書目探勘的實施步驟為(謝建成，2010，頁 51-55)：

1. 決定焦點的範圍：

設定焦點的範圍可能是圖書館某一特定問題或是一般決策所需。如以資料探勘技術分類，書目探勘包含預測性與描述性兩種形式，前者是以從過去的歷史資料或是當前的狀況中推測未來可能發生的事件，這在過去的測量技術或是資料量較少的情況中是無法達成的；後者則是在探究當前的狀況。另外書目探勘可分成直接資料探勘與間接資料探勘兩種，直接資料探勘是以某一問題為焦點，並驅使探索進行。而非直接資料探勘則是反覆利用許多工具以找出可用的型樣(patterns)，找出的型樣必須對圖書館員是具意義的，如「不同的系所或是讀者對於電子期刊的使用有何差別？」，藉以瞭解一般性的議題。

2. 確定內部與外部資料資源

書目探勘所需的是異動、非整合、低階層的資料。內部資料資源是指已存在於圖書館系統或圖書館維護者。如讀者借閱資料、館藏資料、讀者資料、圖

書館網路資源使用紀錄檔、館際合作資料等，為圖書館相關業務所產生的大量資料；而外部資料資源是非存在於現有圖書館系統或是非屬圖書館維護者，如教師研究專長、教師授課資料、學生成績資料等。書目探勘在資料庫領域中是屬於垃圾進垃圾出(garbage in garbage out)的資料處理作業，資料的正確與否、資料是否足夠，對於探勘的結果有絕對性的影響。

3. 蒐集、淨化與匿名化資料形成資料倉儲

為考量讀者資料的隱私性，在資料倉儲的建立的過程中，必須將原始資料庫中具有個人性質的資料予以匿名化處理，如身份證字號的轉換、姓名以代號替換、以系所名稱取代個人資料等作法，館員透過 SQL 查詢語言，從原始作業層資料庫萃取所需的資料，以產生資料倉儲資料庫，這項步驟是極為耗時且循環不斷的，會占據整個書目探勘過程最多的時間，同時也是書目探勘成功與否關鍵性的步驟。

4. 選擇適當的分析工具發現型樣

當資料倉儲資料庫建立之後，則可以進行分析的工作，可透過傳統的統計聚集工具，計算並產生許多不同的統計報表；亦或利用各類具有人工智慧的資料探勘工具，對龐大的資料倉儲進行探勘找出適合的型樣，資料探勘工具會因為所支援的探勘方法不同而有所差異。

5. 分析和實作結果

其結果產出後，無論是統計報表或是決策模式或型樣，都必須經由熟悉領域的圖書館館員或專家，以未經使用過的資料測試驗證或是經由一段時間的驗證評估，以強化或是修正結果。

因此對於書目探勘應用的作業流程，一開始依據圖書口管理決策需求，由

圖書館專家、圖書館館員與圖書館管理者確定資料資源的範圍，接著藉由資料處理工具對於原始資料進行萃取、清理、轉換、載入等，以建立所需之資料倉儲，或是擷取其部分需求資料以形成資料超市，再運用不同演算法資料探勘工具發掘可用之資料型樣或是產生報表，或是以線上分析工具進行多維度資料審視，最後經由圖書館館員或是管理實際驗證、評估、解釋結果。

二、書目探勘應用範疇

資料探勘對於圖書館的應用與協助，可以讓圖書館在經營管理或是服務議題上得更多的決策支援，進一步得知圖書館使用者的行為，以利於制訂更符合使用者需要的服務項目與制度，Schulman 認為，利用資料探勘支援決策，藉此了解圖書館使用者的行為，就可以讓圖書館重新規劃館藏發展方向與制定相關政策，並能根據所了解的使用者行為來設計圖書館相關活動，Schulman 並認為，當圖書館內的資料庫持續的成長與更新時，絕不可能再利用傳統人工的方法，來掌握不斷變動的使用者行為模式及趨勢，必須考慮建置一個決策支援系統，藉此調整出最佳的館藏發展方向，並即時提供最新的資訊服務 (Schulman，1998，轉引自謝建成，2010，頁 58)。

Nicholson(2003)提到圖書館人員若想要在資料庫中獲取資訊，可以利用敘述統計方法或是資料探勘技術，然後發掘出使用者的行為模式，了解這些行為模式的目的：(1)對於館藏政策有更好的決策；(2)讓圖書館的服務機制更合理化，且可以朝向個人化服務；(3)制訂適切的圖書館服務項目以因應網際網路的時代；(4)圖書館管理者在進行決策時，建立更有價值的參考資訊(Nicholson, 2003)。

Christos、Sarantos、Michalis 與 Alexandra(2003)等人提到，因為圖書館資

料的數位化與網際網路的出現，使得研究圖書館使用者行為的分析迅速發展，藉由利用資料探勘技術分析數位化紀錄資料，研究者可以找出使用者的共同行為，建立出一套有意義的群組關係提升資訊的取得，分析的結果也可以找出圖書管理重要的工作項目以及建議工作的可行性方式，所建立的使用者行為群組關係，意味著使用者的資訊需求與諮詢行為，對於圖書館的助益在於：(1)協助管理者重新制訂符合各類使用者需要的館藏政策；(2)協助管理者在作圖書館決策時有一個有效的參考資訊；(3)以及藉由推薦相似的主題協助使用者易於找出有興趣的資訊(Papatheodorou, Kapidakis, Sfakakis, & Vassiliou, 2003，轉引自謝建成，2010，頁 59)。

三、圖書館書目探勘資料資源

由於資料資源對於圖書館進行資料探勘之關鍵因素，圖書館必須釐清且選擇適當的圖書館相關資料作為書目探勘的資料資源，才可獲得正確且合於需求的資訊，基本上資料資源分為五類，分別為(1)圖書館自動化系統；(2)電子資源使用紀錄；(3)圖書館網站使用紀錄檔；(4)研究參考文獻；(5)圖書館管理與服務紀錄，分述如下：

(一)、圖書館自動化系統：

Macini(1996)、Atkins(1996)與 Peters(1996)均提到利用圖書館自動化系統內的資料，可以有利於獲取管理資訊，提升管理績效以及制定管理決策項目。而圖書借閱紀錄向來是讀者實際使用圖書館資源的證據，也是讀者滿足個人資訊需求的行為結果，其中潛藏大量圖書與讀者間互動的歷史紀錄，以及有意義的關係或規則。因此圖書借閱紀錄能反映讀者的實際資訊需求，對於掌握讀者興

趣，進而作為加強圖書館資源利用之基礎，具有一定的參考價值(Macini,1996; Atkins,1996;Peters,1996)。

(二)、電子資源使用紀錄

BanerJee(1998)認為即使紙本資料未來仍是圖書館重要的館藏形式，但是電子形式出版品的資訊量將日漸大量的增加，尤其電子資源是跨越時間與空間限制的一種館藏形式，透過網路使用圖書館資料庫檢索、電子期刊的使用者，將是有增無減的。而 Peters(1996)指出電子資源使用紀錄可以讓圖書館進行電子資源採購的效益評估，以作為圖書館經營者管理資訊系統的基礎，而以往圖書館憑藉著直覺與經驗來執行館藏發展工作，而現在可以透過此一科學的方法協助進行(BanerJee,1998; Peters, 1996)。

(三)、圖書館網站使用紀錄檔

現今圖書館讀者透過網際網路可以隨時且迅速地擷取圖書館相關館藏與服務資源或是訊息，圖書館網站事實上已成為了圖書館服務的延伸，因此讀者於圖書館網站上的使用紀錄，也是了解使用者使用圖書館資源的重要參考依據。館員可經由網站紀錄檔資料探勘的分析，了解電子資源使用狀況、讀者服務需求狀況，讀者使用或是資訊尋求模式等，以提供圖書館增進服務與館藏之參考。

(四)、研究參考文獻

Lawerce、Giles & Bollacker 等人指出圖書館可以利用研究論文之參考文獻作探勘分析，以了解目前的研究趨勢，以及確認新興的研究領域，同時也可以了解參考文獻使用頻繁狀況，進一步利用所得結果作為研究領域的分類，並依此推薦讀者相關書目參考，此直接且客觀的研究資料分析，具有極大的參考價

值：了解文章之間的相關性質，讓研究者在文獻出版時更注意文獻校正與修改，在前期研究工作進行時可以得到有意義的參考文獻提升研究內容，可以縮短找尋資料時所需耗費的時間成本等(Lawerce、Giles & Bollacker,1999)。

（五）、圖書館管理與服務紀錄

書目探勘在圖書館管理經營議題的分析探討，亦為需要業務相關資料資源，事實上圖書館業務運作時，會因為業務不同而產生不同類型的紀錄，如門禁紀錄、館際合作紀錄、館藏採購紀錄、館務會議紀錄、參考服務紀錄等，對圖書館在進行不同議題時的資料探勘時的資料需求，可以提供更完整的資料資源，以增進書目探勘的成效，以支援更正確的圖書館業務管理決策。

（六）、圖書館讀者相關資料

對圖書館要進行書目探勘時，圖書館讀者借閱資料是一項重要的資料資源，若要進一步分析讀者使用圖書館行為，則必須需要更多的讀者相關資料，但也必須考量個人隱私的層面與保護，以保障個人讀者的權益。

Nicholson 於 2006 年對於書目探勘的資料倉儲的資料類型，提出了一套概念框架，它涵蓋三個部份：(1)來自於圖書館館藏本身的資料，如作者、書名、主題、其網址或是館藏地等書目計量資料等；(2)有關圖書館讀者資訊，如除讀者本身個人統計資料，讀者使用網際網路之 IP 等；(3)圖書館做為中介體，在讀者與資訊互動時所產生的資料，如讀者資料搜尋紀錄、流通紀錄、參考服務紀錄、館際合作紀錄等，(Nicholson, 2006)，參見圖 7。

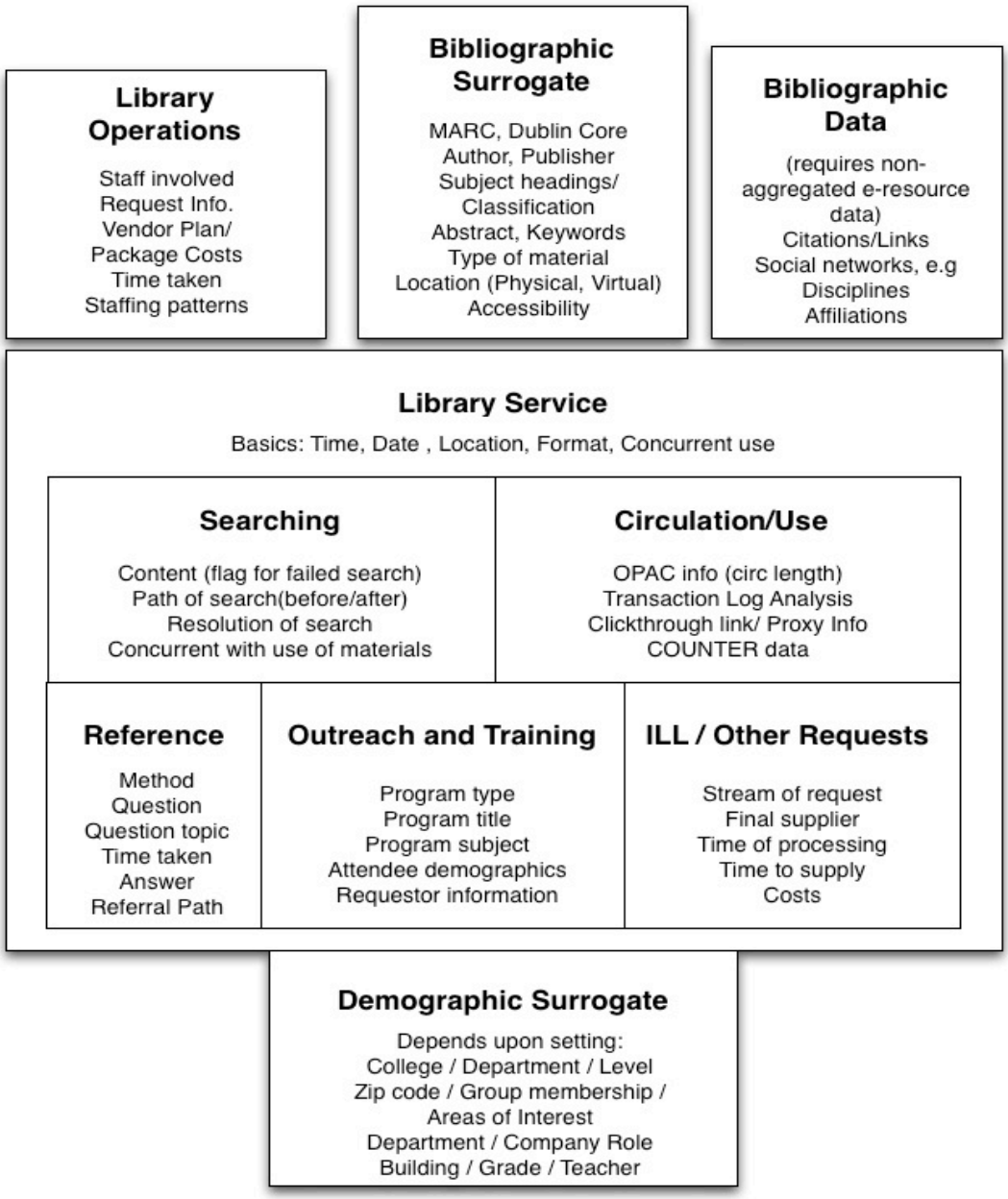


圖 7 Nicholson 提出的書目探勘資料倉儲資料類型的概念框架

資料來源：根據 http://bibliomining.com/nicholson/nicholsonbibliointro.html，根據內容重新繪製。

第參章　研究設計與實施

本研究為建構一醫學電子資源大數據分析模式，並分別對(1)電子資源使用分析；(2)醫學領域主題分析；(3)醫學領域書目計量等項目進行探討，故分為(1)研究方法與研究設計；(2)研究對象；(3)研究步驟；(4)研究架構等方面進行說明，分述如下。

第一節　研究方法與研究設計

本研究為解決以往電子資源使用分析的問題，採取實證研究方法，建構醫學電子資源大數據分析模式，並針對(1)電子資源使用分析；(2)醫學領域主題分析；(3)醫學領域書目計量等不同項目進行探討。

在電子資源使用分析方面，本研究在使用資料紀錄分為兩個來源，即代理伺服網路使用紀錄檔與圖書館自動化系統。前者是使用者在下載電子資源的行為紀錄，而後者是圖書館所紀錄的所有館藏、讀者資訊、借閱紀錄等資料。在從網際網際使用紀錄檔的資料清理與處理過程中，分析與回溯原始文獻與點擊(含下載)檔案之間的關聯，並自行開發爬蟲程式與應用程式介面(Application Program Interface, API)結合，連結期刊出版社或電子資料庫，取得文獻或期刊書目資料，在經資料整合後，再進行使用者點擊(含下載)電子資源之使用統計分析。

在醫學領域主題分析方面，除了從電子資料庫或原出版社的網頁提供引文

資訊外，可透過該學科領域的書目資料庫，獲取電子期刊文獻的書目資料，配合所屬主題詞表，據以分析該學科領域的研究主題與趨勢，故本研究透過自行開發網路爬蟲(Web Crawler)程式，透過 PubMed 資料庫的提供的應用程式介面(Application Programming Interface，API)，取得書目資料，並自行建置(1)文獻的引文資訊；(2)醫學主題詞樹(MeSH Tree)與(3)文獻醫學主題詞路徑(Document MeSH Tree Path)資料庫。為求文獻期刊資料之代表性，本研究從期刊評比資料庫-Scimago Journal Rank(簡稱 SJR 資料庫)，取得癌症研究領域的期刊清單，亦建立醫學期刊清單資料庫，以利於透過自動化方式在最短時間能獲取所需要的書目資料與統計數據。

在醫學領域書目計量方面，本研究自期刊評比指標資料庫-SCImago Journal Rank 與醫學期刊文獻索引摘要資料庫-PubMed 取得資料，前者是限定主題領域範圍，後者是蒐集該領域之書目資料，在取得癌症研究領域期刊清單後，與 PubMed 資料庫提供的期刊清單進行比對，將符合條件的期刊進行檢索，取得完整的醫學文獻書目資料，採用書目計量學的布萊德福、齊夫與洛卡等三大定律進行驗證，並分析文獻成長趨勢、出版類型等特性。

第二節　研究對象

本研究的研究對象分為(1)使用者；(2)醫學文獻書目資料；(3)醫學研究領域期刊與電子資料庫等三類。

在使用者方面，本研究採用某醫學大學的電子資源代理伺服器網際網路使用紀錄檔做為研究對象，從資料清理與整理、資料整合等過程中，得知醫學大

學不同學院系統單位之使用現況與偏好。

在醫學文獻書目資料方面，本研究以 PubMed 資料庫收錄的書目資料為基準，分析其著錄欄位與樣式，並分析醫學主題詞表之組成結構，以建立文獻書目計資料庫與主題詞表資料庫。

在醫學研究領域期刊與電子資料庫方面，本研究的研究對象分為(1)Wiley 資料庫收錄之醫學領域期刊；(2)期刊評比資料庫－SJR 資料庫收錄癌症研究領域的期刊；(3)Nature 資料庫出版之期刊，本研究以 Oncogene 期刊為例。

第三節　研究步驟

依據研究設計，本研究的研究步驟分別為：文獻及蒐集分析、代理伺服器網際網路使用紀錄檔資料清理與整理、醫學電子資源與書目資料的取得、使用數據與書目資料計量分析、研究成果撰寫等步驟，其研究架構參見圖 8。

一、文獻及資料蒐集分析

閱讀電子資源使用分析、網際網路使用紀錄檔分析、書目計量學與醫學主題詞相關研究的文獻，得知目前研究的現況，以進行研究架構的規劃，另一方面，必須先行建置代理伺服器，始能進行紀錄檔內容的清理與資料預處理，本研究採用 Squid 代理伺服器軟體，代理伺服器其用途為二：(1)確認使用者身份是否合法使用者；(2)使用者電子資源使用紀錄的收集。

二、代理伺服器網際網路使用紀錄檔資料清理與整理

透過自行程式開發，將代理伺服器紀錄檔的內容進行資料剖析與篩選，結

合圖書館自動化取得使用者資料，並依據點擊(含下載)之文獻檔案網址，找出文獻與所屬期刊或電子資料庫之原始網址的關係，並取得其詮釋資料，以進行後續書目資料取得的準備工作。

三、醫學電子資源與書目資料的取得

在醫學電子資源選擇方面，分為三個部份：

1. Nature 資料庫收錄之醫學期刊文獻：

本研究以 Nature 資料庫收錄的 Oncogene 期刊為例，從原始紀錄檔中篩選出實際點擊(含下載)Oncogene 期刊的文獻使用紀錄，剖析其詮釋資料再至 PubMed 資料庫取得書目資料，做為醫學領域主題分析的原始資料集。

2. 期刊評比資料庫 SJR 的癌症研究期刊清單：

本研究採用 SJR 內收錄的癌症研究的相關期刊清單，將其連至 PubMed 資料庫進行檢索與確認，取得書目資料後，做為(1)醫學領域主題分析與(2)醫學領域書目計量的原始資料集。

3. Wiley 資料庫收錄之期刊：

本研究採用某醫學大學的代理伺服器的網際網路使用檔，篩選實際點擊(含下載)Wiley 資料庫之使用紀錄，剖析其網址，進行資料比對後之使用數據，做為電子資源使用分析之原始資料集。

四、使用數據暨書目資料計量分析

根據上述取得的書目資料，自行開發程式撰寫與分析原始數據，依據不同需求與目的建立資料庫，再進行相關數據的計算與分析，包括文獻之出版年代、期刊、主題詞、作者數等因素等。

五、研究成果撰寫

將上述的研究成果進行論文撰寫，包括(1)醫學電子資源大數據分析模式與資源相關資料處理；(2)醫學電子資源使用分析；(3)醫學領域主題分析與(4)醫學領域書目計量等研究成果，並提出研究結論與建議。

第四節　研究架構

本研究之研究架構，分為三個部份：(1)使用者電子資源使用分析；(2)醫學領域主題分析；(3)醫學領域書目計量，分述如下，參見圖 8。

一、使用者電子資源使用分析

其研究架構分為：

1. 使用者下載行為：

使用者點擊(含下載)醫學電子資源文獻的行為發生後，在代理伺服器紀錄檔即會產生原始使用紀錄。

2. 網際網路使用紀錄檔資料清理與預處理：

本研究分別以(1)以 Nature 資料庫收錄之 Oncogene 期刊與(2)Wiley 資料庫為例進行電子資源使用分析的流程，從使用者點擊(含下載)之使用行為產生使用紀錄後，自雜亂的原始使用紀錄檔，篩選符合網址樣式(Pattern)的紀錄，進行(1)網址回溯暨分析作業與(2)結合圖書館自動化系統，取得相關資料等作業，前者目的在得知醫學文獻書目之詮釋資料，後者則是得知使用者相關紀錄，以區別個別使用狀況。

3. 自 PubMed 資料庫取得詳細書目資料

將上述獲得之原始文獻的詮釋資料(Metadata)，包括 DOI 與文獻標題等欄位等資料做為檢索詞，連結至 PubMed 資料庫進行檢索，得到以 MEDLINE Abstract 樣式的書目資料，再進行書目詮釋資料欄位的剖析，獲得詳細書目與醫學主題詞等資料。

對於上述取得使用者資料與書目資料的整合應用，可進行電子資源使用分析，以得知使用者對於醫學電子資源的使用偏好。

二、醫學領域主題分析

醫學領域主題分析之資料來源分別為：

1. 醫學主題詞表(Medical Subject Headings)

本研究採用的主題架構為美國國家醫學圖書館發佈的醫學主題詞表，以 2012 年發佈的為基準。

2. 醫學文獻書目詮釋資料：

本研究採用的醫學文獻書目詮釋資料，以 PubMed 資料庫所收錄的文獻為準，由於 PubMed 資料庫為美國國家公開與免費之文獻索摘資料庫，故可證明其文獻資料的專業性與完整性。

3. 醫學研究領域期刊清單：

本研究採用期刊評比資料庫－SCImago Journal Rank 資料庫提供之癌症研究領域之期刊清單做為範圍，以證明其專業領域期刊的代表性。

根據上述資料，本研究分析取得其詮釋資料，並依醫學主題詞表的原有分層架構，根據其架構原則自行建置 MeSH 主題詞樹資料庫，以便於主題詞的檢索與欄位關聯，利於處理與分析主題的速度。

三、醫學領域書目計量

本研究按照上述步驟取得醫學領域文獻之書目資料,採用書目計量之三大定律－布萊德福定律、布萊德福－齊夫與洛卡定律進行書目計量之相關研究，以驗證醫學領域的發展趨勢。

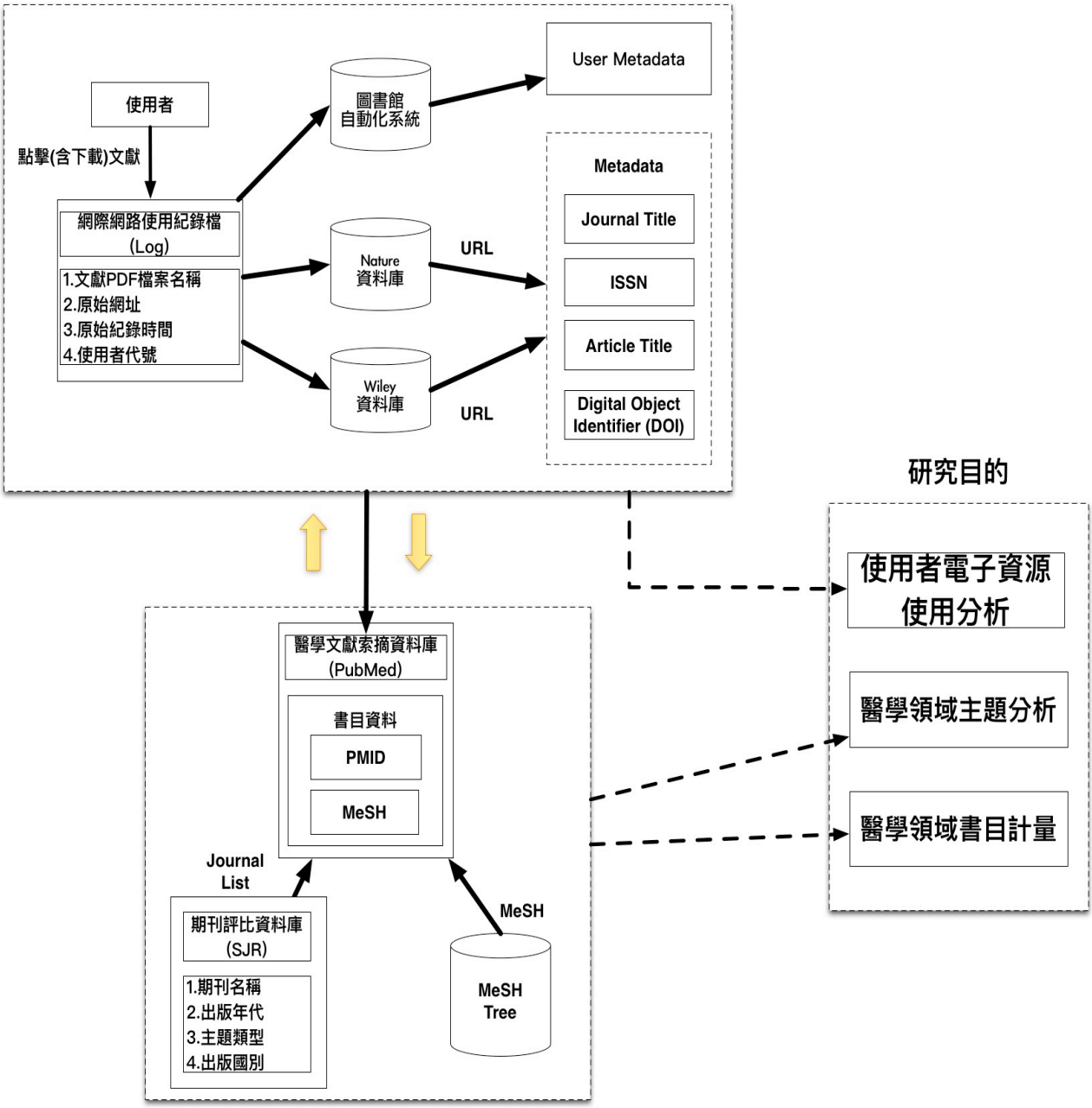


圖 8 研究架構圖

資料來源：本研究整理

第肆章　醫學電子資源大數據分析模式

本節旨在說明論文之主要貢獻-醫學電子資源大數據分析模式內容。過去在從事書目計量學、書目探勘或是電子資源使用者行為等相關研究，無論是原始的使用數據、期刊文獻的書目資料、主題詞等，在資料的取得有其困難性，受限於取得的資料量，故進行相關研究亦有其局限。而在資料處理角度方面，以往之相關研究之處理方式亦採以人工處理為主，未能發展出適當的評估方法，若採以過去少數主機進行統計分析的傳統思維，已無法因應現今日新月異的龐大資料量，故為解決電子資源使用分析上的困境，本研究提出電子資源使用分析模式框架，其功能與貢獻包括：

1. 本研究提出之電子資源使用分析模式，在經由資料取得與整理的流程後，其分析項目包括(1)電子資源使用分析；(2)醫學主題分析；(3)醫學領域書目計量等議題。

2. 針對紙本期刊資源之使用分析，在現今資料採數位化儲存的方式，本研究可採回溯方式取得數位化的書目資料，解決處理書目資料取得不易的問題，亦能處理大量期刊文獻資料，增加研究結果的多樣性與完整性。

3. 研究者收集原始使用數據，經資料處理後建立資料倉儲，為資料探勘的重要處理流程，過去若針對圖書館電子資源進行資料探勘，因取得電子資源使用數據有其難度而無法進行相關研究，本研究採取代理伺服器與圖書館自動化系統驗證之紀錄方式，可即時紀錄電子資源使用分析數據，在經資料處理等流程後，即可建立資料倉儲，有利於書目探勘的相關研究。

4. 過去進行文獻之主題分析，容易因研究者的主觀意識而影響分類之結果，易造成分析結果的研究爭議，為減少書目資料來源的爭議性，本研究之資料來源採用美國國家醫學圖書館公布的醫學主題詞表與文獻索引摘要資料庫-PubMed，根據其主題詞架構建立醫學主題詞樹資料庫，若更換其他研究領域，亦可採取同樣方式建立其他領域之主題詞樹資料庫，利於增加資料分析與檢索的速度。

5. 透過本研究所提出電子資源大數據分析模式，與過去相關研究之差異，本研究自行開發程式，對於網際網路使用紀錄檔的網址資料回溯與書目詮釋資料進行資料剖析，具備處理大量資料之能力，相對證明其研究結果亦比過去之相關研究能更加完整。

下面說明醫學電子資源大數據分析模式之內容，分為(1)醫學電子資源大數據分析模式；(2)系統建置與資料收集；(3)資料處理階段等三部份，分述如下：

第一節　醫學電子資源大數據分析模式

為解決以往電子資源使用分析的問題，本研究提出一電子資源使用分析模式，首先必須自行收集使用數據，故須自行建立代理伺服器，其後資料處理與分析的過程，分為四個階段，分別為實體資料來源、資料預處理、資料呈現與應用領域，最後為了易於保存與分析數據，須自建使用明細資料庫儲存處理完成的資料，以利於透過自動化的方式在最短時間能夠獲取所需的統計數據，以下針對本研究之資料來源、分析模式內容、系統建置與資料收集和資料處理等過程進行描述。

本模式分為四個概念：(1)實體資料；(2)資料預處理；(3)資料呈現；(4)應用領域，以下分述之，請參見圖 9。

一、實體資料

在使用資料紀錄分為兩個來源－代理伺服器網際網路使用紀錄檔與圖書館自動化系統，前者是使用者在下載電子資源的行為紀錄，而後者是圖書館所紀錄的所有館藏、讀者資訊、借閱紀錄等資料，在本研究中，圖書館自動化系統可提供讀者資訊，以利於在分析使用資訊時，建立各使用者的區隔。

二、資料預處理

本概念主要結合上述兩資料來源進行資料的預處理與整合，從前一階段處理完成之 並依目的之不同，其資料處理之過程與步驟較為繁複，區分為：(1)原始資料轉換；(2)文獻書目資料取得；(3)使用者識別與資料取得；(4)處理資料匯出；

三、資料呈現

使用資料在經由資料清理、預處理等過程後，可依據不同目的，而篩選出所需要資料，分為五種：(1)文獻原始書目；(2)文獻參考書目；(3)文獻主題詞；(4)使用者與其下載文獻明細資料；(5)使用者下載電子資源使用統計，本研究並建構一文獻書目暨使用明細資料，為不同領域需求，提供使用數據或書目文獻，以做為資料分析之用。

四、應用領域

本研究除建構醫學電子資源大數據分析模式外，其所處理後的資料，可應用於書目計量學、書目探勘與電子資源使用者行為之用。

此四項概念的重點在於，如何透過系統化的分析方式，清理與整理使用數據資料，匯出成可再進階利用的分析資料，為不同的應用領域，提供使用明細與相關書目資訊，為圖書資訊相關研究領域提供新的資料來源，後述說明實際資料預處理與整理之過程。

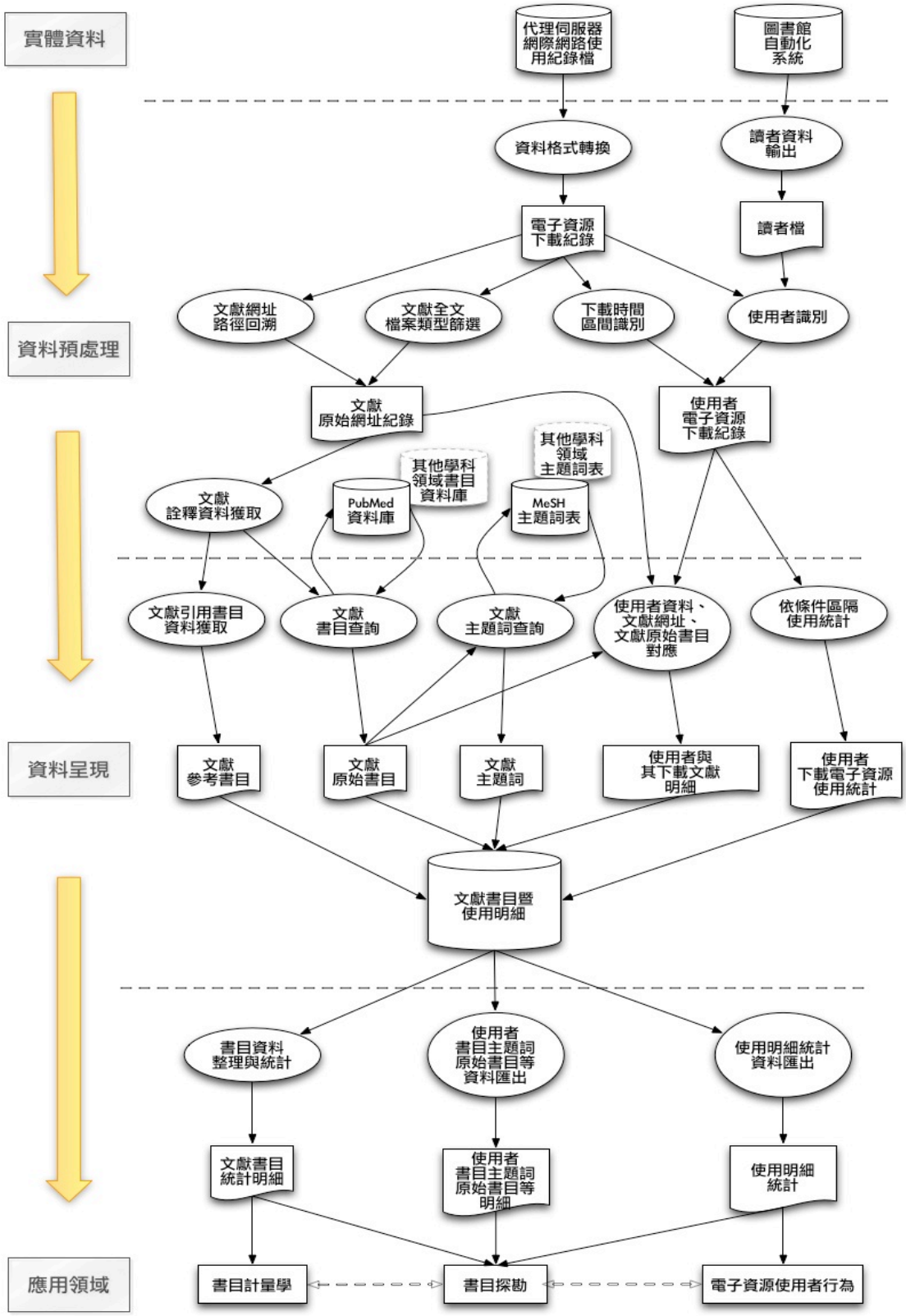


圖 9 電子資源使用分析模式

資料來源：本研究自行整理

第二節　系統建置與資料收集

圖書館提供電子資源服務，如何判斷是否為合法使用者，始能允許文獻的取用，實務上的做法，一般多採取使用(1)讀者身份；(2)讀者來源 IP 位址等兩種驗證方式，前者主要目的是識別是否為該單位機關具合法存取權限之使用者，後者目的是在進行文獻取用時，廠商會根據來源的 IP 位置識別是否為合法單位 (Breeding, 2001，轉引自詹麗萍，2005，頁 180)，判定是否可以取用該文獻，但電子資料庫廠商並無法得知取用文獻之原始使用者為何，僅能得知是否為具取用權限之單位，故在電子資源使用分析方面，廠商提供的統計報表，僅能得知全體使用統計數據，圖書館無法針對機關之下屬單位進行細項統計分析，故為達能進行細項使用者統計分析的目的，圖書館須自行建置一電子資源代理伺服器，其主要目的有二：(1)判定機構的合法使用者與所屬權限；(2)與電子資料庫廠商伺服器進行網路介接，換言之，圖書館建置之代理伺服器為一中介窗口，提供身分驗證與取用文獻之服務，透過此方式收集單位內的使用紀錄，進而與原有的圖書館自動化系統結合，獲知使用者身份與連線來源位置，始能進行細項的資料處理與分析等後續階段。

實務上，圖書館在建置代理伺服器，提供給使用者有數種連線方式：(1)使用者須在連線瀏覽器進行額外的參數設定，將代理伺服器的 IP 位置填入，經身份驗證後，才能正確的連線與取用所需文獻；(2)使用者不須事先填入代理伺服器的 IP 位置，但須連線至圖書館提供的電子資源管理系統進行身分驗證，再連線至所需文獻的網站；(3)透過虛擬個人網路(Virtual Private Network ,VPN)連線方式。

另外，在收集使用紀錄之後，由於代理伺服器提供的紀錄檔，其格式未必統一，在進行系統分析與判讀前，須得知該伺服器軟體採用的紀錄檔格式，始能依據紀錄檔的欄位樣式，讀取所需資料進行分析。

第三節　資料處理階段

從資料探勘的觀點，取得原始資料後，須進行資料預處理(Data Preprocessing)的動作，包括資料清理(Data Cleaning)、資料轉換(Data transformation)、資料縮減(Data reduction)、資料整合(Data Integration)等過程(Jiawei, H and Micheline , K.著，王派洲譯，2008，頁 50-52)，本研究經下列數個步驟進行處理：

一、資料清理

資料清理主要是在清除原始資料的雜訊(noise)，因任何收集到的原始資料，易有如遺失值、雜訊與資料不一致等情況，本研究在資料清理階段，由於網際網路紀錄檔的記錄格式，特別在網址紀錄的部份，資料庫廠商或期刊出版社，除了文獻網址外，可能會插入特殊識別碼，讓網址資料的長度無法完全記錄，容易造成網址欄位之紀錄不完整，故在資料清理階段，須清除無法識別的資料。另因本研究主要以期刊全文檔案，且以 PDF 檔案格式為區隔方式，故在網際網路紀錄檔中若非上述 PDF 檔案格式之全文檔案紀錄，則在此階段皆一併處理。

二、資料整合

本研究因為處理跨年度之紀錄資料，必須將各年度與不同月份的原始紀錄資料進行整合，本步驟主要將不同時期的紀錄檔，整合成一個跨年度與月份的原始紀錄資料檔。

三、資料轉換

取得代理伺服器網際網路使用紀錄檔，至轉換為電子資源下載紀錄，為求在資料處理的便利，須將原始紀錄進行資料格式的轉換，其理由為：(1)紀錄檔若非在紀錄時間的格式，多採用毫秒(microsecond)的格式儲存，利於運算處理但不利於人類的辨讀。(2)當紀錄檔資料量發展至海量資料的規模時，以檔案做為儲存方式，會在資料檢索與運算的效能不彰。

在資料格式轉換的過程中，分為五個動作，參見圖 10：

1. 轉換時間戳記：代理伺服器在時間戳記(Timestamp)欄位採取機讀格式進行時間欄位的註記，其註入方式採取 Unix 紀元(UTC 1970-01-01 00:00:00)以來的秒數，並採取毫秒的方式輸出，在這個階段中，須將毫秒的儲存型式轉換為人類能判讀的格式，本研究採用紀錄檔分析報表軟體-SARG(Squid Analysis Report Generator) (SARG, 2002)，該軟體提供格式轉換功能，易於在處理原始資料時間戳記。

2. 更改時間、日期樣式：第一次轉換時會以歐美常用的「MM/DD/YYYY」輸出，為求與資料庫管理系統中，時間欄位的樣式一致，再進行二次轉換，以「YYYY-MM-DD」方式輸出，在本階段採用 Unix 系統常見的編輯器 sed (stream editor)，將紀錄檔的時間格式透過正則表達式方式輸出。

3. 轉換資料欄位分隔符號：，將原始的欄位分隔符號，轉換為以定位點（tab-delimited）分隔，降低在資料處理時判讀的錯誤機率，本階段採用 Unix 系統上的樣式掃描與處理語言 awk，進行紀錄中的欄位分隔符號轉換工作。

4. 確認資料轉換前後的正確性：若在不同的作業系統進行格式轉換的行為，如 Mac/Linux/Unix/Win，其轉換過程中會因判斷字元編碼的因素而導致失敗，本研究取得的紀錄檔之字元編碼，為求利於跨平台時的處理，字元編碼皆統一以 utf-8 格式進行儲存，在轉換後再進行資料筆數的統計與樣式的驗證，以確保資料的完整性。

5. 輸出電子資源下載紀錄：將驗證正確後的紀錄輸出至檔案，以做為資料預處理階段中主要的資料處理來源。

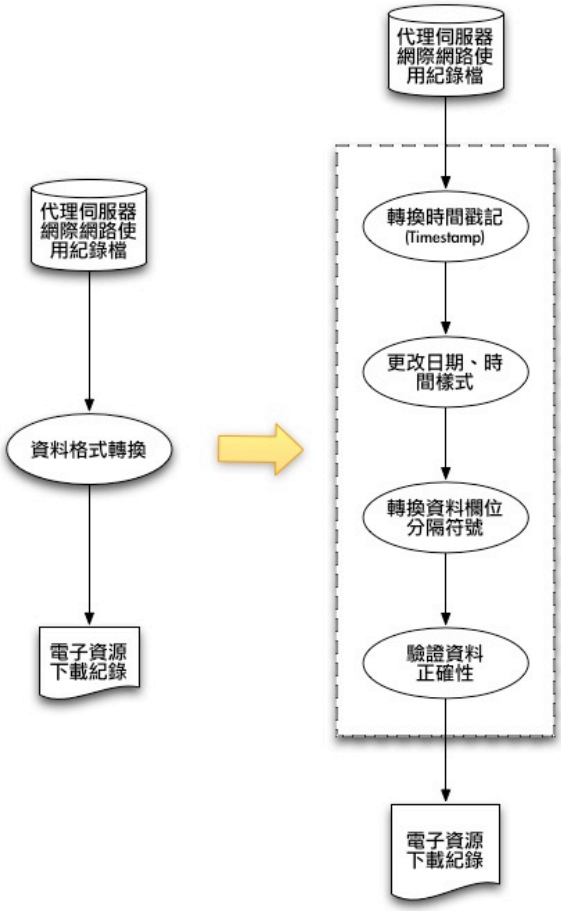


圖 10 網際網路使用紀錄檔資料格式轉換流程

資料來源：本研究整理

從前述的資料格式轉換流程中，從原始難以判讀的紀錄檔，整理並匯出為電子資源下載紀錄，下一階段為求(1)使用者身份；(2)原始文獻相關資訊等兩項目的，須分別進行不同的處理流程，並衍生其他相關檔案，以下分述之。

(1). 使用者身份方面：主要從圖書館自動化系統中的讀者檔取得讀者之相關資訊，在使用紀錄檔中，必須著錄能足以識別讀者身份之資料，始能透過與讀者檔進行比對，實務上圖書館自動化系統習慣以(1)學號或職員編號；(2)個人身份證或居留證號碼做為惟一識別之依據，再著錄其他相關資訊，如科系、所屬單位或部門，若遇多重身份之讀者，則須經由其他欄位之輔助加以識別。

(2). 原始文獻網址紀錄 ：每一出版社或電子資料庫廠商之網址樣式(Pattern)皆有不同，有的為期刊縮寫，有的為 ISSN，有的由電子資料庫廠商或出版社自行編碼，如同讀者身份，在資料清理時，必須事先分析區隔原始文獻網址之樣式，並清理其他妨礙辨識文獻網址之雜訊，始能著錄至使用明細資料庫時，增加檢索效率。

(3). 使用者電子資源點擊(含下載)紀錄：在代理伺服器之網際網路使用紀錄檔案內，除了記錄上述兩者，另須如來源 IP、點擊時間等其他欄位，以資識別與分析。

6. 資料呈現：依據需求目的之不同，產出各類型的資料檔，如文獻原始書目檔、文獻參考書目檔、文獻主題詞檔、使用者與下載文獻明細檔、使用者下載電子資源使用統計等，以下分述之。

(1). 文獻原始書目：在回溯文獻原始書目的過程，其資料來源有二：(1)電

子資料庫或出版社；(2)PubMed/MEDLINE 資料庫。由於從下載紀錄中回溯的網址，為電子資料庫或出版社為該文獻所著錄之書目內容，因而可做為文獻基本書目資訊的來源，待確認回溯網址至文獻後，其書目資料輸出格式有二：(1)網頁內嵌的詮釋資料(Metadata)；(2)提供的引文書目檔案 (Citation file)，如以 Nature 資料庫為例，Nature 提供書目管理軟體 Reference Manager、Endnote 相容的 RIS 檔案，但並未針對文獻本身進行主題分類，為求書目資料之完整性，則將文獻標題與 DOI 等可供識別之欄位資訊，透過 PubMed/MEDLINE 資料庫進行查詢，取得該文獻之書目資訊，並以 MEDLINE 格式輸出成純文字檔儲存，電子資料庫或出版社與 PubMed/MEDLINE 資料庫提供之書目資訊內容最大其不同處，為 PubMed/MEDLINE 另提供醫學主題詞(MeSH)，作為主題分類與檢索之用途。

(2). 文獻參考書目：透過前述取得文獻原始書目等資訊後，可藉此為基礎取得該文獻之參考書目，其來源有二：(1)電子資料庫或出版社提供之參考文獻資訊檔案(Reference Files)；(2)ISI Thomson Reuters 公司之 Web of Science(WOS)資料庫，提供之參考文獻清單檔案。兩者皆提供利於書目管理軟體 Endnote 之匯入格式，或是採用 Tab 分隔之純文字檔的格式，透過文獻參考書目的匯總，可以進行書目計量學的常見研究型式，如引文分析、文獻老化、作者生產力等相關研究。

(3). 文獻主題詞：PubMed/MEDLINE 資料庫將其收錄之文獻，皆會從文獻之標題、摘要等內容，給予適當的醫學主題詞，而醫學主題詞的編撰，皆由相關領域之專家學者所編製，每年皆發布新的版本，透過醫學主題詞，研究者若對特定醫學領域有其偏好，可透過主題詞檢索相關之參考文獻，若結合圖書館

電子資源之使用者分析，從微觀的角度，可探索使用者之下載文獻之偏好，若從鉅觀的角度，可從科系、學院、單位、團體為區隔，分析不同單位等對於某特定醫學領域之下載偏好。或從期刊為出發點，得知不同使用者或科系、單位對該期刊之下載狀況，對圖書館管理階層，可做為館藏評鑑之參考，或為未來是否續訂評估之有效證據，本研究從 PubMed/MEDLINE 資料庫提供的醫學主題表，自行建置 MeSH Tree 資料庫，根據每一文獻中的主題詞，剖析每一主題詞的主題路徑，並建構文獻主題詞對應資料庫。

(4). 使用者與下載文獻明細：從圖書館本身建置之伺服器取得使用紀錄，從紀錄檔之原始資料，配合圖書館自動化系統，從讀者檔內可解析出紀錄檔中使用者代碼的意義，另從下載網址之回溯，可獲取前述的文獻相關資訊，將兩者資料結合分析，即可針對使用者之下載進行分析，但仍須注意個人隱私與相關法律規範。

(5). 使用者下載電子資源使用統計：以往圖書館在進行紙本期刊的使用統計，僅能從上架統計、閱讀調查表單、觀察、訪談或問卷調查等方法搜集資訊，對於電子期刊的使用統計，從資料庫廠商或出版社提供之期刊使用數據平台下載使用數據報告，其數據來自讀者透過圖書館之電子資源管理系統或校園網路等途徑，連結至電子資料庫進行文獻的檢索與下載等行為，最後成功下載文獻之統計數據，其限制是並無法區隔機關單位內的子機構與實際使用者，其報表僅顯能顯示統計數字，若要採取更細微的統計模式，如以文獻為單元，分析由那些使用者所下載的統計報表，就無法進行分析。

7. 資料範例說明：以代理伺服器網際網路使用紀錄為例，如下所列以說明資料轉換過程。

1392966826.714 991 60.251.55.209 TCP_MISS/206 33250 GET http://onlinelibrary.wiley.com/store/10.1002/jbm.a.34466/asset/34466_ftp.pdf?v=1&t=hrx4h5vc&s=aab0ec248a3a488ab3231e4fd2bcc1f4e77f2b39 A123456789 DIRECT/onlinelibrary.wiley.com application/pdf

(1). 在時間戳記部份，「1392966826.714」為 Unix 系統的時間戳記，在經過轉換之後，可得知為「2014/2/21 15:13:46」。

(2). 檔案類型為 PDF，其識別方式為「application/pdf」

(3). 在下載網址部份，原始資料為「http://onlinelibrary.wiley.com/store/10.1002/jbm.a.34466/asset/34466_ftp.pdf?v=1&t=hrx4h5vc&s=aab0ec248a3a488ab3231e4fd2bcc1f4e77f2b39」，在進行資料判讀時，可得知網址中包含了期刊的 DOI，在經由 DOI 網址轉換後即可得知為「Journal of Biomedical Materials Research Part A」。

(4). 進一步從網址中的 DOI，另可對應至文獻篇名「Efficacy of mesoporous silica nanoparticles in delivering BMP-2 plasmid DNA for in vitro osteogenic stimulation of mesenchymal stem cells」。

(5). 紀錄檔中另著記使用者識別碼，因個資保護因素，雖可與圖書館原有自動化系統讀者檔進行比對，僅可針對於未能足以識別個人資訊之項目進行統計分析，如學院系所單位等。

(6). 依據上述轉換後之資料再匯入自行建置之資料庫。

綜上所述，本研究可依據圖書館自動化系統的讀者檔與網際網路使用紀錄檔進行資料比對與區隔，找出使用者與文獻之間的關聯性，進而在圖書館進行使用者分析時，能具備更多的區隔變數，如從點擊(含下載)時間區間、使用者

本身、與其所屬的學院、科系與文獻等變數等，藉以產生更多的加值應用分析。

第伍章　醫學電子資源大數據分析模式之應用

根據醫學電子資源大數據分析模式的內容，本研究再分別進行(1)電子資源使用分析；(2)醫學領域主題分析與(3)醫學領域書目計量等分項應用研究項目，內容分述如下。

第一節　電子資源使用分析

在電子資源使用分析層面，本研究採用某醫學大學圖書館代理伺服器，篩選 Wiley 電子資料庫使用數據為研究個案，原始紀錄檔案為 1.04GB，共 4,181,850 筆紀錄，記錄時間由 2009 年 8 月 17 日至 2014 年 4 月 21 日，剔除無法完整辨識之紀錄，並以 2013 年之期刊清單進行條件篩選，在未進一步分析文獻書目之詳細資料，共有 1,216,156 篇全文 PDF 點擊(含下載)次數。

本研究個案結果說明共有六項，分別為(1)電子資料庫歷年點擊(含下載)次數分析;(2)電子資料庫實際使用分析；(3)年度點擊(含下載)電子資料庫次數分析；(4)跨年度點擊(含下載)期刊之分析-以醫學院為例；(5)單一年度期刊與學科領域科系之點擊(含下載)分析－以公共衛生相關科系為例；(6)特定期刊文獻之點擊(含下載)分析。前二項之統計分析亦可從 COUNTER 統計報表得知，而後面四項則為 COUNTER 統計報表所無法提供，以下分別說明之。

一、電子資料庫歷年點擊(含下載)次數分析(2010 年 8 月至 2014 年 4 月)：

從歷年點擊(含下載)次數分析，2011 年至 2013 年之點擊(含下載)次數皆有超過 30 萬次，若以各年度為區隔，以 2012 年總點擊(含下載)次數為最高，有 404,111 篇文獻點擊(含下載)，其次為 2013 年有 349,358 篇文獻，與 2011 年的 302,760 篇文獻點擊(含下載)次數；2014 年因資料收錄時間，自 1 月 1 日至 4 月 21 日為止，其平均每月下載次數為 29,106 次，與前一年度 2013 年之平均每月下載次數數據相異不大，故平均每月下載次數除 2010 年偏低外，每月點擊(含下載)次數均超過 2 萬 5 千次以上，顯示 Wiley 資料庫在每日的文獻點擊(含下載)量是具有近千篇文獻次數以上的規模。

進一步探析 2010 年之點擊下載次數與平均每月下載次數較少的原因，其可能性研判如下：(1)因 Wiley 資料庫變更電子文獻下載網址，在經由網址樣式(URL Pattern)判斷過程與去重覆之驗證流程後，資料筆數減少；(2)期刊需求之差異，因採取 2013 年的期刊清單做為篩選依據，另收錄時間為 2010 年 8 月 9 日至 12 月 31 日止，故點擊下載次數較其他年度為少。

表 5 歷年點擊下載次數

年份	文獻點擊下載次數	平均每月下載次數
2010	43,496	8,699
2011	302,760	25,230
2012	404,111	33,676
2013	349,358	29,113
2014	116,431	29,106

資料來源：本研究自行整理

說明：2010 年與 2014 年點擊下載紀錄時間區間為分別為：(1) 2010 年 8 月 9 日至 12 月 31 日；(2) 2014 年 1 月 1 日至 4 月 21 日。

二、電子資料庫實際使用分析

以訂購 Wiley 電子資料庫之 2013 年的期刊清單所載之期刊為基準，全數 2,331 種期刊中，僅有 833 種期刊是在 2010-2014 年間至少點擊下載 1 次，而其他的 1,498 種期刊，在橫跨 5 個年度的時間區間，則是未有任何的點擊下載紀錄，亦只有 35.74％的期刊有其讀者需求，而多達 64.26%的期刊則乏人問津，顯示在訂購期刊的方式，採取包裹方式訂購電子資料庫，則可再尋求合理的議價策略；或是採取其他推廣策略，讓讀者能增加點擊下載次數，參見表 6。

表 6 期刊跨年度點擊下載次數至少 1 次以上之比例－以 2013 年訂購之期刊為例

期刊點擊次數(N)	期刊數(J)	比例(J/2,331)%
N>=1	833	35.74%
N=0	1,498	64.26%
總計	2,331	100.00%

資料來源：本研究自行整理

在上述在表 5與表 6的結果中，若從COUNTER的報表類型-Journal Report 1(JR1)為例，亦得知依據月份及期刊名稱統計單本期刊內文章全文的有效需求數，由於代理伺服器網際網路紀錄檔之時間戳記欄位，可將系統記錄之時間，轉換至一般常見的年月日，而 JR1 報表類型僅能依據到月份，故可依據需求的不同，將報表的時間區隔可從「年」、「月」、「日」亦可區分。

三、年度點擊(含下載)電子資料庫次數分析－以 2013 年為例

若以 2013 全年度為時間區隔變數，計有 349,358 篇文獻點擊下載紀錄，分

散於 562 種期刊，期刊之平均點擊下載次數為 621.6 次，另點擊下載次數大於平均值的期刊種數計有 88 種，占 2013 年總期刊數之 24.11%，且該期刊群之點擊下載次數合計為 304,270 次，占 2013 全年度總點擊下載次數之 87.09%，亦可言之，該 88 種期刊為 2013 全年度較為重要之期刊群，參見表 7。

表 7 單一年度期刊點擊下載各分列項目與次數－以 2013 年為例

項目	次數／比例
期刊總數(A)	2,331
2013 年點擊(含下載)期刊數(B)	562
2013 年點擊(含下載)總次數(C)	349,358
平均點擊(含下載)次數(D=C/B)	621.6
點擊(含下載)次數大於平均值(C)的期刊數(E)	88
大於平均值的期刊，其點擊(含下載)次數(F)	304,270
大於平均值的期刊，其占期刊總數之比例(G=D/A)	24.11%
大於平均值的期刊，其點擊(含下載)次數占 2013 年點擊(含下載)總次數的比例(H=F/C)	87.09%

資料來源：本研究自行整理

從上列 2013 年點擊下載期刊再進行分析，以點擊次數多寡為排序的依據，顯示 *Hepatology* 期刊是點擊次數較多的期刊，在 38,941 點擊次數內，分布於 239 人，若以學院數為分類區隔，可得知此 239 人分布於 8 個學院層級的單位，平均每人點擊 *Hepatology* 期刊的文獻次數為 162.9 次；其次如 *The International Journal of Cell Cloning* 期刊，點擊次數為僅次於 *Hepatology* 的期刊，其點擊人數達 180 人，分散於 9 個學院層級的單位，其每人平均點擊下載次數雖不如 *Hepatology*，仍達超過百次之規模，顯示該兩種期刊，其學院層級單位分布較為平均，且平均點擊下載量也較高，顯示該兩種期刊為多數學院層級單位共同皆有閱讀需求；而如以排名第三之期刊-*Journal of Comparative Neurology* 為例，其點擊次數達 15,537 次，但點擊人數僅有 68 人，其平均每人點擊下載次數雖

高於前述的 *Hepatology* 與 *The International Journal of Cell Cloning* 兩種期刊，但僅分布於 5 個學院層級單位，顯示該類期刊可能為特定學院或科系認定是重要期刊，參見表 8。

表 8　2013 年期刊點擊下載次數之前十名期刊

期刊標題	ISSN	點擊(含下載)各項次數統計			
		次數 (A)	人數 (B)	平均次數 (C=A/B)	學院/單位數 (D)
Hepatology	1527-3350	38,941	239	162.9	8
The International Journal of Cell Cloning	1549-4918	19,058	180	105.9	9
Journal of Comparative Neurology	1096-9861	15,537	68	228.5	5
Journal of Comparative Neurology	1550-7130	15,537	68	228.5	5
Journal of Comparative Neurology and Psychology	1550-7149	15,537	68	228.5	5
International Journal of Cancer	1097-0215	9,194	264	34.8	8
Journal of Cellular and Comparative Physiology	1553-0809	8,918	245	36.4	8
Journal of Cellular Physiology	1097-4652	8,918	245	36.4	8
Journal of Cellular Biochemistry	1097-4644	6,591	263	25.1	8
Journal of Supramolecular Structure	1547-9366	6,591	263	25.1	8
Journal of Supramolecular Structure and Cellular Biochemistry	1547-1748	6,591	263	25.1	8
The Journal of Pathology and Bacteriology	1555-2039	6,529	150	43.5	8
The Journal of Pathology	1096-9896	6,529	150	43.5	8
EMBO Molecular Medicine	1757-4684	6,205	65	95.5	6
Journal of Biomedical Materials Research	1097-4636	5,879	86	68.4	7
Journal of Biomedical Materials Research Part A	1552-4965	5,879	86	68.4	7
Journal of Biomedical Materials Research Part B: Applied Biomaterials	1552-4981	5,879	86	68.4	7
Glia	1098-1136	5,637	83	67.9	6

資料來源：本研究自行整理

說明：部份期刊因名稱或 ISSN 改變，事實上仍是同一期刊，如 Journal of Comparative Neurology、Journal of Comparative Neurology and Psychology 為同一期刊。

四、跨年度點擊(含下載)期刊之分析－以醫學院為例

以醫學院與期刊 ISSN 為區隔變數，從 2010 年 8 月至 2014 年 4 月為時間區間，共有 295,161 篇文獻下載紀錄，分散於 561 種期刊，初步以點擊下載次數排序之前十種期刊，參見表 9。

表 9 跨年度期刊點擊下載次數－以醫學院為例之前十名期刊

期刊標題	ISSN	點擊（含下載）次數
Hepatology	1527-3350	30,075
The International Journal of Cell Cloning	1549-4918	16,692
Journal of Cellular and Comparative Physiology	1553-0809	16,422
Journal of Cellular Physiology	1097-4652	16,422
International Journal of Cancer	1097-0215	8,469
Glia	1098-1136	8,419
Journal of Cellular Biochemistry	1097-4644	6,889
Journal of Supramolecular Structure	1547-9366	6,889
Journal of Supramolecular Structure and Cellular Biochemistry	1547-1748	6,889
The Journal of Pathology and Bacteriology	1555-2039	5,724
The Journal of Pathology	1096-9896	5,724
Annals of Neurology	1531-8249	4,721
British Journal of Pharmacology	1476-5381	4,708
Journal of Neuroscience Research	1097-4547	4,623

資料來源：本研究自行整理

由表 9，初步分析自 2010 年 8 月始至 2014 年 4 月期間，醫學院各系所之點擊(含下載)合計以 *Hepatology* 期刊之次數為最高，計有 30,075 次，其次為 *The*

International Journal of Cell Cloning，計有 16,692 次。其他依期刊 ISSN 為區隔，其他期刊按下載次數由高至低為：*Journal of Cellular and Comparative Physiology* (16,422)、*Journal of Cellular Physiology* (16,422)、*International Journal of Cancer* (8,469)、*Glia* (8,419)、*Journal of Cellular Biochemistry* (6,889)、*Journal of Supramolecular Structure* (6,889)、*Journal of Supramolecular Structure and Cellular* (6,889)、*The Journal of Pathology and Bacteriology* (5,724)、*The Journal of Pathology* (5,724)、*Annals of Neurology* (4,721)、*British Journal of Pharmacology* (4,708)、*Journal of Neuroscience Research* (4,623)。

進一步對於點擊(含下載)次數相同的期刊進行分析，部份點擊(含下載)次數相同的期刊，在進行書目資料比對後發現，如 *Journal of Cellular Biochemistry*、*Journal of Supramolecular Structure*、*Journal of Supramolecular Structure and Cellular* 等名稱與 ISSN 不同的期刊，在 Wiley 資料庫進行檢索驗證時，皆為同一份期刊，但在期刊清單內，若無特別註記，在資料處理時易視為各自為獨立的期刊，但卻實際上則為同一份期刊，在資料分析時亦容易造成處理上的困難。

五、單一年度期刊與學科領域科系之點擊(含下載)分析－以公共衛生相關科系為例

以公共衛生相關領域科系為區隔，該領域共有 6 個系所，以 2013 年為例，在 2,331 種期刊中，有 117 種期刊是曾為公衛相關領域科系點擊(含下載)之數據，共點擊 2,198 次，可做為公衛領域相關期刊的初步區隔，而每一期刊之平

均點擊次數為 18.78 次。將 117 種期刊做進一步區隔，其分析步驟如下：

(1) 以期刊平均點擊次數為基準，共有 28 種期刊高於平均點擊次數。

(2) 將 28 種期刊以曾點擊之系所個數為區隔變數，超過半數系所(N>=3)曾點擊(含下載)該期刊，則視為重要程度較高之期刊群(A)，計有 8 種期刊。

(3) 由前一步驟，點擊(含下載)次數超過平均次數且點擊該期刊之系所未達總數系所之一半(N<3)，則視為次重要之期刊群(B)，計有 20 種期刊。

(4) 低於期刊平均點擊次數，則視為重要性較低之期刊群(C)，計有 89 種期刊。

表 10 較為重要之期刊群(A)-以公衛領域科系為例

期刊標題	ISSN	科系/點擊(含下載)次數						點擊(含下載)次數
		H1	H2	H3	H4	H5	H6	
Health Economics	1099-1050	110	188	54	1			353
Arthritis & Rheumatism	1529-0131		92	54	30			176
Statistics in Medicine	1097-0258	66	21	11	5	4	4	111
Arthritis Care & Research	2151-4658		46	27	15			88
International Journal of Cancer	1097-0215	29	33	4	3			69
Addiction	1360-0443	8			16		20	44
Pharmacoepidemiology and Drug Safety	1099-1557	1	7	23	2			33
Arthritis & Rheumatology	2326-5205		12	9		1		22

資料來源：本研究自行整理

透過上述步驟，可區隔出以點擊(含下載)次數與系所為基準的 3 種期刊群。參見表 10 可知，2013 年對於公共衛生領域科系，點擊(含下載)次數皆超過百次的三種期刊，分別為 *Health Economics* (353)、*Arthritis & Rheumatism* (176)、*Statistics in Medicine* (111)三種期刊，且達一半的科系曾點擊(含下載)，其中

Health Economics 為點擊(含下載)次數最多，但 *Statistics in Medicine* 則為全部 6 類系所皆曾點擊(含下載)，其次低於百次但高於平均點擊(含下載)次數的期刊群共有 5 種期刊，分別為 *Arthritis Care & Research* (88)、*International Journal of Cancer* (69)、*Pharmacoepidemiology and Drug Safety* (33) 與 *Arthritis & Rheumatology* (22)，等五種，故這兩種分類之期刊群，可做為公共衛生科系領域之核心期刊之參考。

六、特定期刊文獻之點擊(含下載)分析

以 *Hepatology* 期刊的「Comprehensive analysis of the independent effect of twist and snail in promoting metastasis of hepatocellular carcinoma」文獻為例(Yang et al., 2009)，其 DOI 為「10.1002/hep.23221」，篩選出自 2010-2014 年的點擊(含下載)紀錄，共有 2,144 次，該文章在 2011 年與 2012 年之點擊次數接近，約 450 至 480 次，至 2013 年則達 1,073 次，跨年度之點擊讀者之所屬學院/單位分布，以醫學院(5 次)為最多，其次為生命科學院(4 次)、生物醫學工程學院(3 次)、牙醫學院(3 次)、行政支援單位(2 次)，最少則是護理學院與藥學院各出現 1 次，顯示該文獻為醫學院與生命科學院之科系較具有點擊(含下載)需求，參見表 11。

表 11 特定文獻跨年度點擊次數與讀者之所屬學院/單位之分析

年份	點擊讀者之所屬學院/單位分布	點擊(含下載)次數
2010	醫學院、生命科學院	64
2011	醫學院、生物醫學工程學院、牙醫學院、護理學院	457

2012	醫學院、生物醫學工程學院、生命科學院、牙醫學院、行政支援單位	486
2013	醫學院、生物醫學工程學院、生命科學院、牙醫學院、藥學院、行政支援單位	1,073
2014	醫學院、生命科學院、	64

資料來源：本研究自行整理

第二節　醫學領域主題分析

本節分為兩個部份，首先為醫學期刊之主題分析－以 Oncogene 期刊為例，其次為癌症研究領域期刊之主題分析，分述如下：

一、醫學期刊之主題分析－以 Oncogene 期刊為例

本研究以 Oncogene 期刊為例，透過使用者實際點擊下載紀錄檔案，在進行資料清理與整理的過程後，獲取書目資料，再進行統計分析，分為(1)文獻出版年代與作者數之統計分析；(2)MeSH 主題詞個數分布；(3)文獻之重要主題詞次數分布；(4)MeSH 主標目文獻篇數統計等方面進行說明，藉此得知使用者點擊下載的 4,501 篇文獻中的文獻特性。

(一)、文獻出版年代與作者數之統計分布

從使用者點擊下載 Oncogene 期刊的 4,501 篇文獻中，根據出版年代分布，點擊下載最高的是 2010 年的文獻(494 篇)，其次是 2008 年(392 篇)，第三則是 2004 年(343 篇)，最低則為 1997 年(78 篇)，另文獻出版年代會因為出版社的刊登政策而有所差異，分屬 2013 年的 88 篇文獻，但在 2012 年已刊出電子版，故年代也會因為版本的不同，PubMed/MEDLINE 也會修正其書目資訊，參見圖 11。

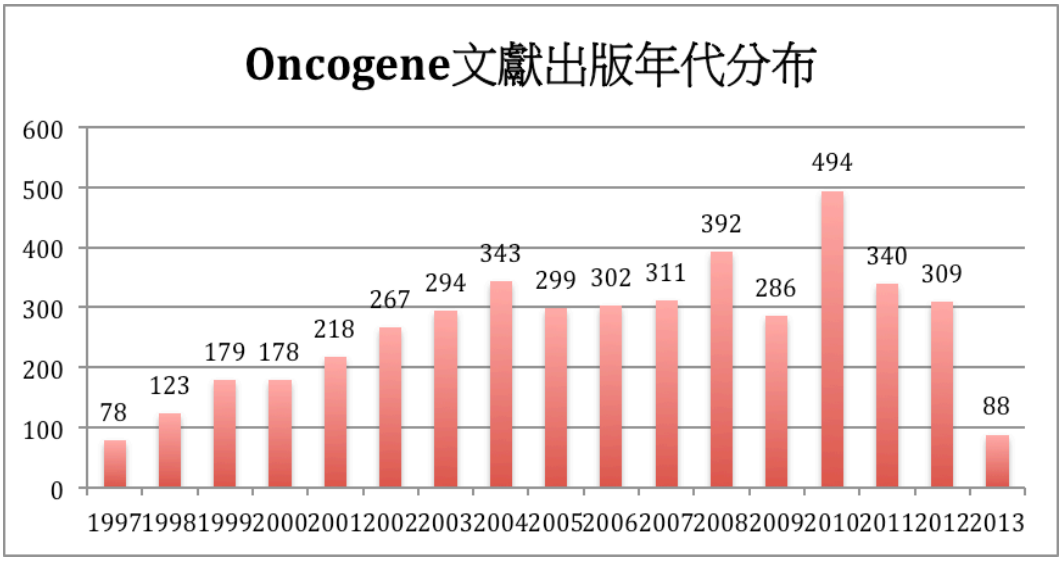


圖 11 Oncogene 文獻出版年代分布

資料來源：本研究整理

從文獻作者數量角度而言，Oncogene 文獻的作者數，多數為 2 位作者(523 篇)，其次為 5 位作者(494 篇)，再其次為 4 位作者(483 篇)與 6 位作者(480 篇)和 7 位作者(416 篇)，顯示多數編撰文獻的作者數約在 2-7 位作者，占文獻總數的 63%，10 位作者以上者的文獻則相對少數，占文獻總數的 19%，參見圖 12。

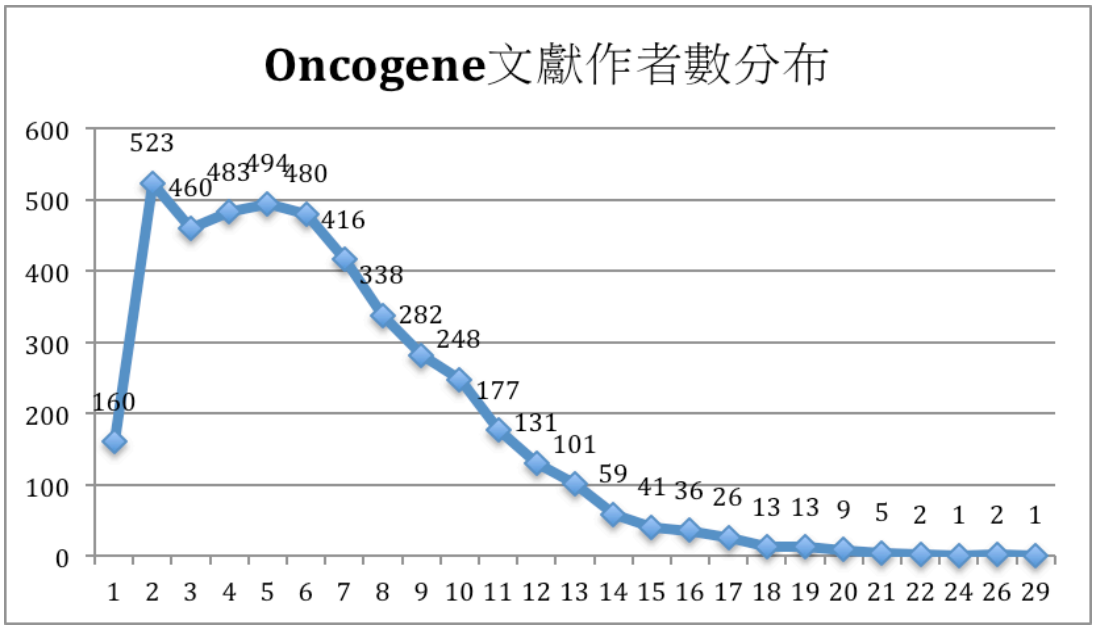


圖 12 Oncogene 文獻作者數分布

資料來源：本研究整理

(二)、MeSH 主題詞個數分布

將依據 PubMed 資料庫的書目資料進行統計，Oncogene 期刊文獻的 MeSH 主題詞個數，最多為在 10-19 個，有 2094 篇，其次為 20-29 個(1432 篇)，而有 0-9 個的 458 篇文獻中，有 72 篇文獻是 0 個主題詞，原因在於 PubMed/MEDLINE 尚未給予 MeSH 主題詞，另有 2 篇文獻有超過 60 個以上的 MeSH 主題詞，參見圖 13。

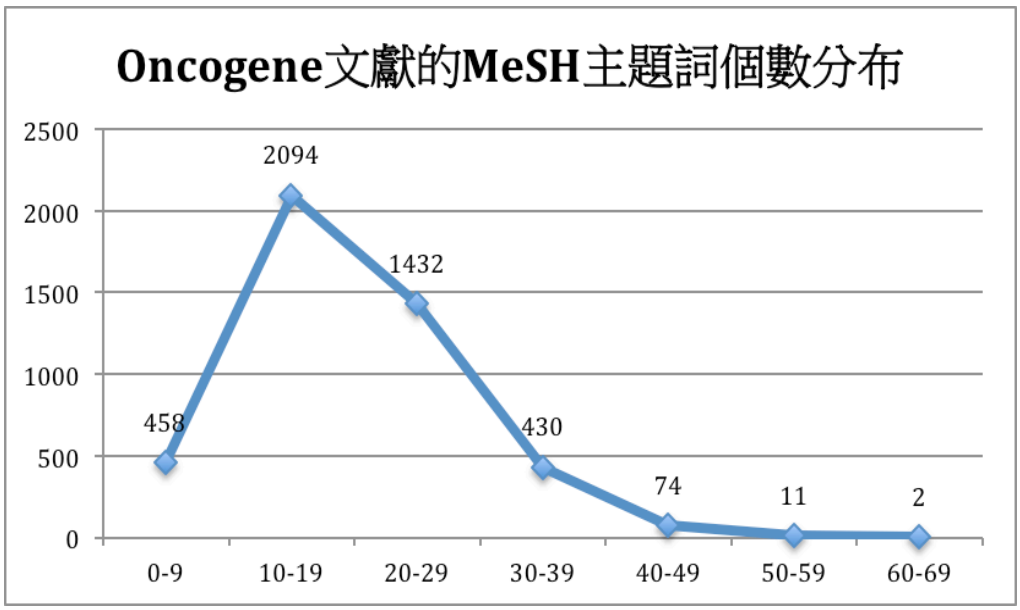


圖 13 Oncogene 文獻的 MeSH 主題詞個數分布

資料來源：本研究整理

(三)、文獻之重要主題詞次數分布

以下載文獻的重要主題(Major Headings)進行分析，首先描述重要主題詞之次數分布，再將其細項分析為(1)主標目次數前十名；(2)以主標目為主之細項分析；(3)依 MeSH Tree 主題階層次數統計等項目，分述如下：

在本研究個案中，重要主題個數為 5 個的文獻篇數為最多(952 篇)，其次為 4 個(909 篇)，其他依序為 6 個(715 篇)、3 個(628 篇)、7 個(457 篇)、2 個(301 篇)、8 個(223 篇)、9 個(106 篇)、10 個(45 篇)、1 個(34 個)、11 個(23 篇)、12 個(8 篇)、13(8 篇)、15 個(2 篇)，顯示 PubMed/MEDLINE 資料庫在描述給予文獻的主要探討主題時，領域專家會賦予 4 至 5 個重要主題詞將文獻進行分類，參見表 12。

表 12 文獻的重要主題詞(Major Headings)個數統計

單篇文獻重要主題個數	文獻篇數

單篇文獻重要主題個數	文獻篇數
1	34
2	301
3	628
4	909
5	952
6	715
7	457
8	223
9	106
10	45
11	23
12	8
13	8
15	2

資料來源：本研究整理

若將重要主題詞的主標目進行統計，在考慮有副標目的情況下，出現次數前十名以 Apoptosis 主標目的文獻篇數為最多(533 篇)，其次為 Neoplasms(462 篇)，再其次為 Proteins(448 篇)、Breast Neoplasms(448 篇)、Transcription Factors(403 篇)、Signal Transduction(389 篇)、Tumor Suppressor Protein p53(377 篇)、Gene Expression Regulation, Neoplastic (374 篇)、Proto-Oncogene Proteins(302 篇)、Cell Transformation, Neoplastic(276 篇)，參見表 13。

1. 重要主題詞次數統計-主標目次數前十名

但如果針對進一步分析，主標目 Apoptosis 雖出現在 533 篇，但僅有 159 篇著錄時限定在主標目，其他 374 篇文獻則是加上副標目，顯示在 Apoptosis

主題詞的相關文獻須其他的副標目，限定所探討的主題內容，另在 DNA-Binding Proteins 的所屬文獻篇數有 1 篇的主副標目，同時都被專家學者認為是重要主題的情況，顯示如果需要針對重要主題詞的主副標目的不同組合進行分析，應該更精確的找出醫學主題文獻的不同面向與組合。

表 13 重要主題詞次數統計-主標目出現次數前十名

主標目	(A)	(B)	(C)	(D)	(E)	(F)
Apoptosis	533	159	374	33	341	0
Neoplasms	462	0	462	3	459	0
DNA-Binding Proteins	448	59	389	4	385	1
Breast Neoplasms	424	0	424	5	419	0
Transcription Factors	403	27	376	0	376	0
Signal Transduction	389	184	205	24	181	0
Tumor Suppressor Protein p53	377	2	375	5	370	0
Gene Expression Regulation, Neoplastic	374	260	114	21	93	0
Proto-Oncogene Proteins	302	41	261	0	261	0
Cell Transformation, Neoplastic	276	107	169	15	154	0

資料來源：本研究整理

註：A: 文獻篇數(A=B+C)；B:僅有主標目；C:主標目/副標目(C=D+E)；D: *主標目/副標目 E: 主標目/*副標目；F: *主標目/*副標目

2. 重要主題詞次數統計－以主標目為主

如進一步針對下載文獻所屬重要主題的主標目統計，在不考慮副標目的情況，以前十名為，第一以 Gene Expression Regulation, Neoplastic 為最多(260 篇)，其次是 Signal Transduction(184 篇)，再其次為 Apoptosis(159 篇)，其他依序為 Cell Transformation, Neoplastic(107 篇)、Genes, Tumor Suppressor(84 篇)、Mutation(83 篇)、Transcription, Genetic(75 篇)、Gene Expression Regulation(72 篇)、DNA Methylation(67 篇)、DNA Damage(67 篇)，參見表 14。

表 14 下載文獻所屬重要主標目統計-僅針對主標目，不包括含副標目在內

主標目	篇數
*Gene Expression Regulation, Neoplastic	260
*Signal Transduction	184
*Apoptosis	159
*Cell Transformation, Neoplastic	107
*Genes, Tumor Suppressor	84
*Mutation	83
*Transcription, Genetic	75
*Gene Expression Regulation	72
*DNA Methylation	67
*DNA Damage	67

資料來源：本研究整理

3. 重要主題詞文獻篇數統計－依 MeSH 主題階層

如從 MeSH 主題詞階層節點為準則，依據點擊下載 Oncogene 期刊的文獻進行分類與統計，在考慮每一篇文獻著錄的重要主題，包括上述的主標目與副標目組合的情況，則 Chemicals and Drugs 主題的文獻篇數(27,306 篇)為最高，其次為 Diseases(11,458 篇)，第三則為 Phenomena and Processes(8,895 篇)，第四則為 Anatomy(1,799 篇)，顯示所下載的文獻屬性，偏重在這四項大類，而 Anthropology,Education,Sociology and Social Phenomena 與 Humanities 兩類主題的文獻篇數最少，都僅有 2 篇，參見表 15。

表 15 Oncogene 文獻重要主題詞統計-依 MeSH 主題階層

MeSH 主題詞分類	代表字母	文獻篇數
Anatomy	A	1799
Organisms	B	424
Diseases	C	11458

MeSH 主題詞分類	代表字母	文獻篇數
Chemicals and Drugs	D	27306
Analytical,Diagnostic and Therapeutic Techniques and Equipment	E	592
Psychiatry and Psychology	F	6
Phenomena and Processes	G	8895
Disciplines and Occupations	H	57
Anthropology,Education,Sociology and Social Phenomena	I	2
Technology,Industry,Agriculture	J	32
Humanities	K	2
Information Science	L	14
Named Groups	M	1
Health Care	N	36

資料來源：本研究整理

（四）、MeSH 主標目文獻篇數統計-以 Chemicals and Drugs(D 類)為例

從上述可知 Chemicals and Drugs(D 類)會是主題分布中最多的一類，下面針對 Chemicals and Drugs(D 類)進行細項分析：

1. Chemicals and Drugs 之細部主題分佈

進一步針對 Chemicals and Drugs(D 類)進行分析，從 Chemicals and Drugs 延申的 16 個主題中，以 Amino Acids, Peptides, and Proteins(D12)主題的文獻數為最多，共有 18,775 篇，其次為 Enzymes and Coenzymes(D08)(3,672 篇)，再其次為 Biological Factors(D23)(1,797 篇)，顯示下載文獻中，在 Chemicals and Drugs 的分類，明顯集中在 Amino Acids, Peptides, and Proteins(D12)主題內，參見表 16。

表 16 MeSH 主標目文獻篇數統計-以 Chemicals and Drugs(D 類)為例

MeSH 主標目	類號	文獻篇數
Inorganic Chemicals	D01	244
Organic Chemicals	D02	503
Heterocyclic Compounds	D03	309
Polycyclic Compounds	D04	110
Macromolecular Substances	D05	168
Hormones, Hormone Substitutes, and Hormone Antagonists	D06	98
Enzymes and Coenzymes	D08	3672
Carbohydrates	D09	89
Lipids	D10	72
Amino Acids, Peptides, and Proteins	D12	18775
Nucleic Acids, Nucleotides, and Nucleosides	D13	779
Complex Mixtures	D20	9
Biological Factors	D23	1797
Biomedical and Dental Materials	D25	1
Pharmaceutical Preparations	D26	6
Chemical Actions and Uses	D27	674

資料來源：本研究整理

2. MeSH 主標目文獻篇數統計-以 Amino Acids, Peptides, and Proteins (D12 類)

進一步將 Amino Acids, Peptides, and Proteins 主題進行分析，在 18775 篇文獻內，分為 Amino Acids(D12.125)、Peptides(D12.644)與 Proteins(D12.776)三種分類，其中以 Proteins 占最多，有 16,021 篇，與其他 Peptides(2,637 篇)與 Amino Acids(117 篇)兩類包含的文獻數量相距甚大，顯示文獻主要密集程度在 Proteins 分類，參見表 17。

表 17 MeSH 主標目文獻篇數統計-以 Amino Acids, Peptides, and Proteins (D12 類)為例

MeSH 主標目	類號	文獻篇數

Amino Acids	D12.125	117
Peptides	D12.644	2,637
Proteins	D12.776	16,021

資料來源：本研究整理

3. MeSH 主標目文獻篇數統計-以 Proteins(D12.776)類，其文獻超過 1000 篇為例

再將 Proteins(D12.776)主標目向下延伸分析，共有 61 個主標目，所屬文獻超過 1000 篇以上有 7 個主標目，所屬文獻最多的主標目為 Neoplasm Proteins(2,203 篇)、其他依序為 DNA-Binding Proteins(2,081 篇)、Intracellular Signaling Peptides and Proteins(1,733 篇)、Transcription Factors(1,708 篇)、Membrane Proteins(1,615 篇)、Nuclear Proteins(1,188 篇)與 Intercellular Signaling Peptides and Proteins(1,023 篇)，上述所屬文獻總和為 11,551 篇，占 Proteins(D12.776)文獻篇數的 72%，顯示在 Proteins(D12.776)一類中，該七項主標目為其主要的大類，參見表 18。

表 18 MeSH 主標目文獻篇數統計-以 Proteins(D12.776)類，其文獻超過 1000 篇為例

MeSH 主標目	類號	文獻篇數
Neoplasm Proteins	D12.776.624	2203
DNA-Binding Proteins	D12.776.260	2081
Intracellular Signaling Peptides and Proteins	D12.776.476	1733
Transcription Factors	D12.776.930	1708
Membrane Proteins	D12.776.543	1615
Nuclear Proteins	D12.776.660	1188
Intercellular Signaling Peptides and Proteins	D12.776.467	1023

資料來源：本研究整理

4. MeSH 主標目文獻篇數統計-以 Proteins(D12.776)類，其文獻超過 1000 篇為例

如依據在 MeSH 主題樹深度階層最高者為例，在本研究統計結果，階層為第 12 層為最高，而其中主標目 Tretinoin (D02.455.426.392.368.367.379.249.700.860.500) ，其文獻篇數有 16 篇，為該階層文獻最多的主標目，文獻分布在 1997-2011 年，除 1999 年、2004 年、2009 年未有文獻，其他分散在各年度，其中 2001 年的文獻為最多(3 篇)，參見表 19。

表 19 MeSH 主標目階層最高者所屬文獻年代分布-以 Tretinoin 為例

文獻年代	篇數
1997	1
1998	1
2001	3
2002	2
2003	1
2005	1
2006	1
2007	2
2008	2
2010	1
2011	1

資料來源：本研究整理

由上述項目可知，單一期刊之使用分析，過去僅能針對單一期刊之點擊下載次數統計，並未能針對所點擊下載之文獻進行細部分析與主題分析，透過本研究提出分析模式之框架，能進行較為深度與細部的分析，以利於得知電子資源使用者之使用行為。

二、癌症研究領域期刊之主題分析

本研究個案依據 SJR 期刊評比指標資料庫之資料，主題分類選擇癌症研究(Cancer Research)，該主題於 2012 年公布的期刊資訊，收錄的相關期刊計有 196 種，本研究取得期刊之書目資料做為關鍵詞，至 PubMed 進行檢索，於 2013 年 5 月 3 日進行檢索後，自行建置期刊文獻書目資料庫後，撰寫本個案的數據與分析結果，分為(1)癌症研究於 SJR 期刊與 PubMed 均收錄之文獻，其排名數量比例；(2)期刊評比等級 Q1 中癌症研究文獻篇數前十名之期刊；(3)癌症研究領域期刊的文獻出版年代分布-以期刊評比等級 Q1 為例；(4) 癌症研究領域文獻之前 20 項核心主題詞；(5)特定重要主題詞之次數分布-以 Breast Neoplasms 為例。以下分述之。

(一)、癌症研究於 SJR 期刊與 PubMed 均收錄之文獻，其排名數量比例

根據 2012 年 SJR 期刊評比資料庫所公布的資訊，在癌症研究領域中，共計有 196 種期刊，但將其期刊標題與 ISSN 設為檢索詞，連結至 PubMed 資料庫進行檢索，則計有 102 種期刊，同時篩選 PubMed 已給予主題詞的文獻，其文獻筆數共有 402,495 篇，期刊的平均篇數計有 3,946 篇，顯示在癌症研究領域方面，約超過半數以上的期刊(102/196=52%)，為 SJR 期刊評比資料庫和 PubMed 生物醫學文獻資料庫同時收錄，參見表 20。

表 20 癌症研究於 SJR 期刊與 PubMed 同時收錄文獻之排名數量比例

期刊排名 (Rank)	文獻數 (PQn)	比例(%) (PQn/P)	期刊數 (JQn)	比例 (%) (JQn/J)
Q1	251,357	62.45%	39	38.24%

期刊排名 (Rank)	文獻數 (PQn)	比例(%) (PQn/P)	期刊數 (JQn)	比例 (%) (JQn/J)
Q2	62,592	15.55%	30	29.41%
Q3	70,317	17.47%	29	28.43%
Q4	18,229	4.53%	4	3.92%

資料來源：本研究整理

（P=402,495, J=102, P/J= 3,946）

進一步計算 102 種期刊中，期刊排名(Rank)為第一區(Q1)的期刊數計有 39 種，占總數的 38%，其所出版的文獻計有 251,357 篇，占文獻總篇數的 62.45%，顯示排名 Q1 的期刊，多數亦會被 PubMed 資料庫所收錄；而排名為第二區(Q2)的期刊計有 30 種，文獻篇數計有 62,592 篇，占文獻總篇數的 15.55%；排名第三區(Q3)的期刊計有 29 種，文獻篇數計有 70,317 篇，文獻總篇數的 17.47%，數據顯示，Q2 與 Q3 的期刊數與所收錄的文獻篇數相距不大，而排名第四區(Q4)的期刊，相對被 PubMed 資料庫收錄最少，計有 4 種期刊，僅有 18,229 篇文獻被收錄，文獻被收錄的比例也僅占 4.53%，但同時被 SJR 與 PubMed 資料庫同時收錄的期刊數，Q1 的期刊數有 39 種，占總數的 38.24%，而 Q2 與 Q3 的期刊數被同時收錄的種數接近，分別為 30 與 29 種，故所占比例也分別近總數的 30%，顯示期刊評比排名較前段的 Q1 與 Q2，其所占比例就已達近七成。

（二）、 期刊評比等級 Q1 中癌症研究文獻篇數前十名之期刊

從期刊評比等級 Q1 的 39 種期刊進一步分析，其出版的文獻篇數為排序基準，在前十名方面，文獻出版篇數最高者為 *Cancer research*，計有 47,099 篇文獻，其次為 *Cancer*，計有 37,055 篇，其餘期刊的文獻篇數皆在 2 萬篇以下，在出版國家方面，除 *Cancer Letters* 期刊為愛爾蘭出版外，均以英、美兩國為多數，

此外，出版文獻篇數較多的期刊，未必其排名較前，以 *Cancer research* 期刊為例，文獻篇數雖為 Q1 之第一，但於 2012 之排名為第八名，而排名第十名的 *European journal of cancer (Oxford, England : 1990)*，其出版文獻數也未超過萬篇(9,653 篇)，顯示期刊評比的排名雖主要與文獻引用、作者聲望等因素有關，但在文獻出版數仍須達至一定的規模，參見表 21。

表 21 期刊評比等級 Q1 中癌症研究文獻篇數前十名之期刊

期刊名稱	ISSN	文獻篇數	Q1 排名 (2012)	出版國家
Cancer research	0008-5472	47,099	8	美國
Cancer	1097-0142	37,055	20	美國
Journal of the National Cancer Institute	0027-8874	19,707	6	英國
British journal of cancer	1532-1827	19,558	24	英國
Journal of clinical oncology : official journal of the American Society of Clinical Oncology	1527-7755	18,603	3	美國
Oncogene	0950-9232	15,669	11	英國
Clinical cancer research :an official journal of the American Association for Cancer Research	1078-0432	12,594	10	美國
Carcinogenesis	1460-2180	10,809	21	英國
Cancer letters	1872-7980	10,539	44	愛爾蘭
European journal of cancer (Oxford, England : 1990)	0959-8049	9,653	22	英國

資料來源：本研究整理

在原始資料處理時，期刊名稱與 ISSN 雖為較具有識別能力的欄位，但實

際在進行 SJR 與 PubMed 兩資料庫進行比對時，仍會有難以互相對應的狀況，因資料庫在著錄期刊書目資料時，雖會將期刊名稱與 ISSN 著錄，但 PubMed 資料庫會將期刊的出版社、地區、發行年份、副刊名等資訊，一併著錄在刊名欄位也一併著錄，如 *European journal of cancer (Oxford, England : 1990)*期刊，在 SJR 資料庫僅會顯示「*European journal of cancer*」，但 PubMed 資料庫則著錄為「*European journal of cancer (Oxford, England : 1990)*」，若再加上 ISSN 資訊有所變更，則在資料清理與整理階段時，造成資料不一致的狀況產生。

（三）、癌症研究領域期刊的文獻出版年代分布-以期刊評比等級 Q1 為例

從文獻的年代分布層面來看，同樣以期刊評比等級為 Q1 的 39 種期刊出版的 251,357 篇文獻為統計範圍，可以得知在 PubMed 資料庫收錄的文獻範圍中，最早的是出版年代為 1945 年，若以十年做為一個區間，可看出癌症研究文獻的成長趨勢，1945-1949 年為癌症研究的萌芽期，該時間區段的文獻出版數，僅有 593 篇，但從 1950 年代開始蓬勃發展，癌症研究的文獻陸續出版，至 1960 年代，從 4,618 篇成長至 7,706 篇，具有近一倍成長量，從下一個階段-1970 年代起始，則文獻於十年間的出版量開始破萬篇，1970 年代與 1960 年代相比，其成長量則有近 12,000 篇的增量，1970 年代至 1990 年代，其二十年之間的文獻出版量，從 19,726 篇，成長至 62,388 篇，而至 2000-2009 年期間的文獻出版量，為歷年出版量最高，計有 91,339 篇的文獻出版，而 2010-2013 年則有 30,168 篇文獻出版，也已相近於 1980-1989 年十年間的文獻出版量(34,819 篇)，未來極有可能超越 1990-1999 年與 2000-2009 年的文獻出版量，值得注意的是，自 1990-2009 年的出版文獻量，已為全部出版文獻總數的

38.19%(153,727/402,495)，代表有近四成的研究貢獻，出現於近 20 年的研究文獻裡，顯示癌症研究為醫學研究領域中，文獻出版量相當穩定且為研究者重視的研究領域，參見圖 14。

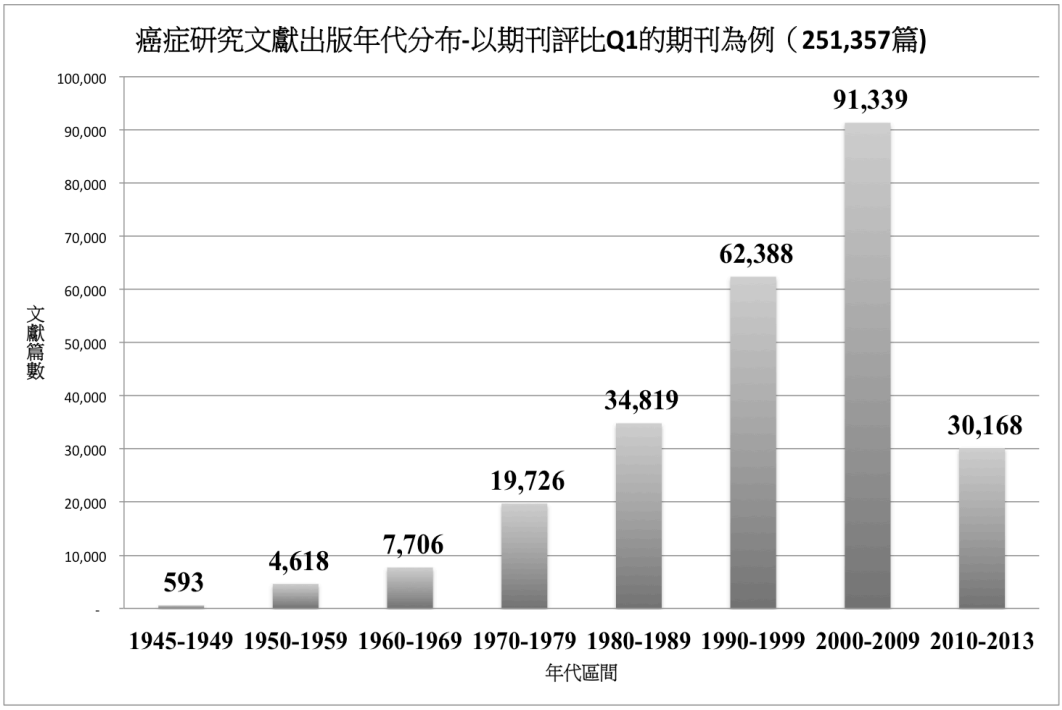


圖 14 癌症研究領域期刊的文獻出版年代分布-以期刊評比等級 Q1 為例

資料來源：本研究整理

（四）、癌症研究領域文獻之前 20 項核心主題詞

從主題分布層面探討，由於 PubMed 資料庫，若同時為 MEDLINE 收錄時，依據文獻的特性，給予不定數目的主題詞，同時若認為該主題詞足以識別該文獻的主要範疇，則以星號(*)做為重要主題詞的識別依據。本研究的分析結果，共有 17,061 個重要主題詞，分布於 402,495 篇文獻，其中前 20 項重要主題詞，為詞頻次數超過 1 萬次以上，依頻率次數再分類為三類，該三類主題詞群可視為癌症研究核心群中，極為重要的核心主題詞組，參見表 22。

表 22 癌症研究領域文獻之前 20 項核心主題詞

依出現次數排序	重要主題詞 (Major Topic)	出現次數 (Total)
1	Breast Neoplasms	63,767
2	Neoplasms	50,680
3	Antineoplastic Agents	36,570
4	Antineoplastic Combined Chemotherapy Protocols	33,743
5	Lung Neoplasms	30,632
6	Adenocarcinoma	21,150
7	Prostatic Neoplasms	17,465
8	Carcinoma, Squamous Cell	16,558
9	Liver Neoplasms	16,171
10	Colorectal Neoplasms	14,972
11	Brain Neoplasms	14,890
12	Ovarian Neoplasms	14,274
13	Melanoma	13,528
14	Stomach Neoplasms	13,389
15	Tumor Markers, Biological	13,213
16	Apoptosis	12,042
17	Carcinoma	11,987
18	Colonic Neoplasms	11,413
19	Skin Neoplasms	11,072
20	Carcinoma, Non-Small-Cell Lung	10,899

資料來源：本研究整理

從詞頻分布層面說明，以詞頻出現次數為排序依據，在 17,061 個重要主題詞中，詞頻出現次數超過一萬次以上的，共有 20 個重要主題詞，分為三類：(1)詞頻出現次數超過 5 萬次，以乳腺腫瘤(Breast Neoplasms)出現次數為最多，計有 63,767 次，其次則為腫瘤(Neoplasms)，出現 50,680 次，前兩者為出現頻率

最高，可視為在癌症研究中，代表性最高且最普遍的重要主題詞；(2)詞頻出現次數達 2~3 萬次間：抗腫瘤藥(Antineoplastic agents)、抗腫瘤聯合化療方案(Antineoplastic Combined Chemotherapy Protocols)、肺腫瘤(Lung Neoplasms)、腺癌(Adenocarcinoma)等四項，可視為次一級的重要主題詞群；(3)詞頻出現次數 1 萬次以上但未達 2 萬次：該詞群計有 14 項，分別為前列腺腫瘤(Prostatic Neoplasms)、鱗狀細胞癌(Carcinoma, Squamous Cell)、肝腫瘤(Liver Neoplasms)、大腸癌(Colorectal Neoplasms)、腦癌（Brain Neoplasms)、卵巢癌（Ovarian Neoplasms）、黑色素瘤(Melanoma)、胃腫瘤(Stomach Neoplasms)、腫瘤生物標記(Tumor Markers, Biological)、細胞凋亡(Apoptosis)、肺癌(Carcinoma)、結腸癌(Colonic Neoplasms)、皮膚癌(Skin Neoplasms)、非小細胞肺癌(Carcinoma, Non-Small-Cell Lung)等 14 項重要主題詞。上述 20 項重要主題詞，即為在進行癌症研究相關領域中，足以代表該領域的主題分類。

（五）、特定重要主題詞之次數分布-以 Breast Neoplasms 為例

醫學文獻的主題詞分布，從廣義和狹義範圍分析探討，可以分為兩種：(1)文獻的主題詞欄位本身著錄 Breast Neoplasms 主題詞；(2)從 MeSH Tree 的主題詞主題詞編號得知，其主題編號之組成含前項主題詞的主題編號，以 Breast Neoplasms 為例，其主題路徑則有 2 條，分別為「C04.588.180」與「C17.800.090.500」，亦言之，當有其他醫學文獻所著錄的主題詞，其主題編號包括上述兩項，皆可以視為該文獻與 Breast Neoplasms 相關，其主題結構與樹狀編號路徑對應，參見圖 15。

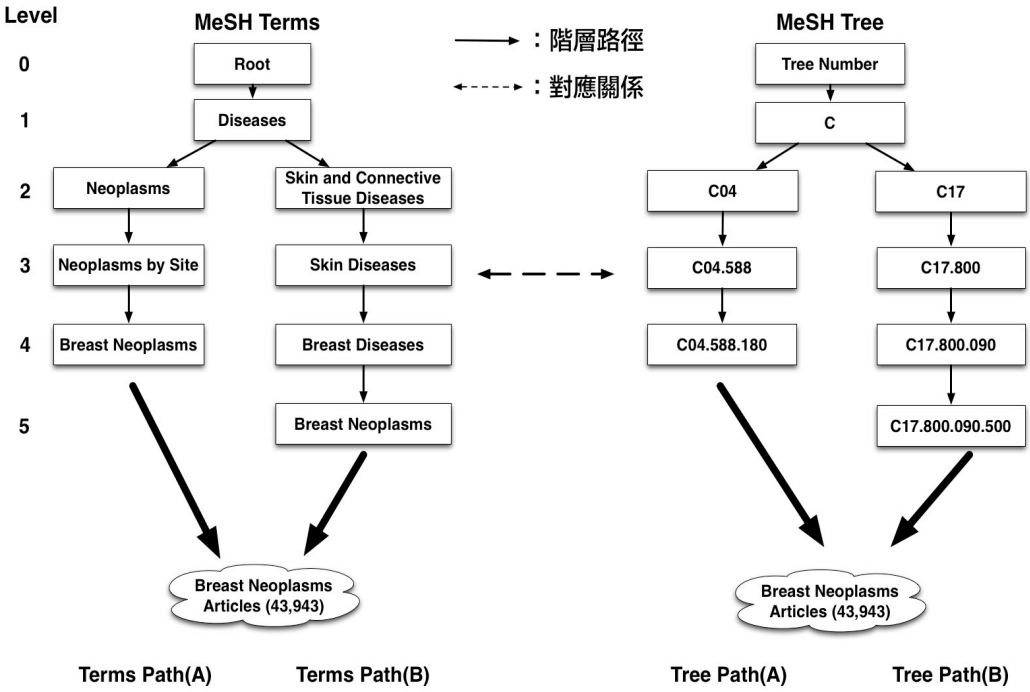


圖 15 醫學主題詞 Breast Neoplasms 之主題詞路徑與樹狀編號路徑對應圖

資料來源：本研究整理

從圖 15 得知，從主題路徑與樹狀編號路徑，將每一個節點由上至下的通過，至 Breast Neoplasms 主題詞，經本研究統計，Breast Neoplasms 主題詞一共出現 63,767 次，其中直接著錄 Breast Neoplasms 主題詞的癌症研究文獻有 43,836 篇，而加上主題詞編號包括「C04.588.180」與「C17.800.090.500」兩條路徑的文獻 107 篇，合計共有 43,943 篇，亦可言之，在檢索 Breast Neoplasms 的相關醫學文獻時，43,943 篇的醫學文獻可為檢索的求全率(Recall)為 100%的結果，但求準率(Precision)則仍須依據使用者需求的差異而有所區別。

1. 以 SJR 資料庫期刊排名(Rank)為區隔之次數統計

若以 SJR 資料庫之期刊排名為區隔，有 99 種期刊含有 Breast Neoplasms 主題詞的文獻共 43,943 篇(P=43,943, J=99, P/J=443.86)，以所占比例而論，以期刊排名 Q1 的文獻數占最多，在 38 種期刊中，出版篇數 30,839 篇，占總出版篇

數 70.18%(30,839/43,943)，其平均篇數也高達 811.55 篇，顯示 Breast Neoplasms 一詞，在期刊排名 Q1 的 38 種期刊中，是相當重要且常見的研究主題；而 Q2 與 Q3 的期刊數分屬 29 與 28 種，出版篇數相對較為接近，分別有 4,685 篇與 6,829 篇，且 Q3 的期刊較 Q2 的期刊對於該主題是較感興趣的，Q3 的平均篇數也較 Q2 為高；而 Q4 由於被 PubMed 資料庫收錄較少，故文獻數與平均篇數相較 Q1、Q2、Q3 少。

2. 以出版年代為區隔之次數統計

以出版年代作為區隔變數進行分析，主題詞 Breast Neoplasms 首篇文獻出版於 1945 年，自此之後研究有關該主題的文獻之出版趨勢明顯成長，若以十年為一區間探討，自 1990 年起至 2009 年的二十年間，為研究 Breast Neoplasms 主題並作為研究成果發表文獻達 30,008 篇，達文獻數量的 68.29%，其中以 2000 年至 2009 年期間的出版文獻為最多，計有 19,645 篇，而 1990 年至 1999 年間的出版文獻，計有 10,363 篇；同時 1990 年代至 2000 年代期間的其文獻成長量為最高，達 9,282 篇，參見圖 16。

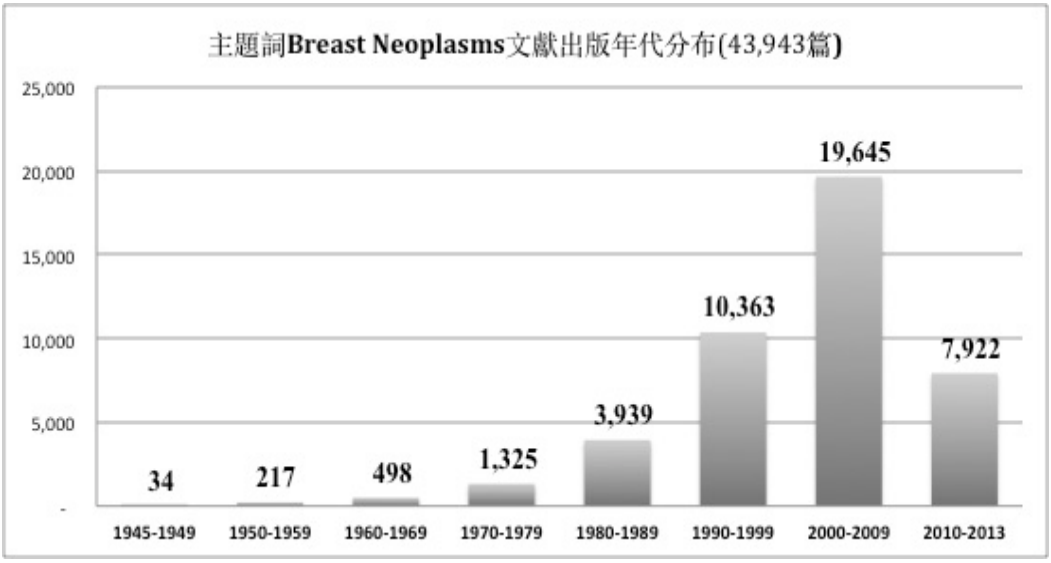


圖 16 主題詞 Breast Neoplasms 之次數統計(以出版年代為區隔)

說明: 1. 主題詞 Breast Neoplams 首篇文獻出版於 1945 年；2. 2010 年代統計僅至 2013 年

資料來源：本研究整理

第三節　醫學領域書目計量

在醫學領域書目計量方面，本研究之結果分析的內容包括：(1)癌症研究文獻成長與出版品類型分布；(2)布萊德福定律驗證；(3)布萊德福-齊夫定律驗證；(4)洛卡定律驗證與作者特性，分述如下。

一、癌症研究文獻成長與出版品類型分佈

(一)、文獻數量成長

本研究共計檢索出 608,830 篇文獻書目，將 PubMed 預先著錄的 2016 年文獻 40 篇省略不計，共 608,790 篇，出版年介於 1945 年~2015 年間，首篇出版

文獻自 1945 年，由各年出版量的成長趨勢來看，從圖 17 觀察 70 年來的研究文獻成長，癌症研究呈向上成長趨勢，自 1945 年至 1991 年，每年的文獻成長量以百篇至數百篇為成長，自 1992 年起，其文獻成長量突破每年千篇，並呈現逐步升高的趨勢，至 2014 年的 33,866 篇達到高峰，文獻成長數量統計詳如表 23。

表 23 癌症研究期刊文獻逐年分布統計表

年代	期刊篇數	累計篇數	年代	期刊篇數	累計篇數
1945	14	14	1981	4,480	56,814
1946	141	155	1982	5,154	61,968
1947	198	353	1983	5,676	67,644
1948	180	533	1984	6,435	74,079
1949	366	899	1985	6,365	80,444
1950	384	1,283	1986	7,036	87,480
1951	477	1,760	1987	7,280	94,760
1952	495	2,255	1988	8,029	102,789
1953	500	2,755	1989	9,003	111,792
1954	517	3,272	1990	9,686	121,478
1955	757	4,029	1991	9,873	131,351
1956	726	4,755	1992	10,815	142,166
1957	770	5,525	1993	11,960	154,126
1958	818	6,343	1994	12,478	166,604
1959	943	7,286	1995	12,698	179,302
1960	916	8,202	1996	13,263	192,565
1961	918	9,120	1997	13,993	206,558
1962	866	9,986	1998	14,358	220,916
1963	1,029	11,015	1999	15,364	236,280
1964	1,153	12,168	2000	15,536	251,816
1965	1,108	13,276	2001	15,624	267,440
1966	1,358	14,634	2002	16,641	284,081
1967	1,477	16,111	2003	18,122	302,203

年代	期刊篇數	累計篇數	年代	期刊篇數	累計篇數
1968	1,569	17,680	2004	18,778	320,981
1969	1,660	19,340	2005	20,179	341,160
1970	1,849	21,189	2006	21,027	362,187
1971	1,873	23,062	2007	21,610	383,797
1972	2,150	25,212	2008	22,799	406,596
1973	2,407	27,619	2009	23,971	430,567
1974	2,767	30,386	2010	25,513	456,080
1975	2,847	33,233	2011	26,178	482,258
1976	3,279	36,512	2012	29,072	511,330
1977	3,632	40,144	2013	30,925	542,255
1978	3,661	43,805	2014	33,866	576,121
1979	4,040	47,845	2015	32,669	608,790
1980	4,489	52,334			

資料來源：本研究整理

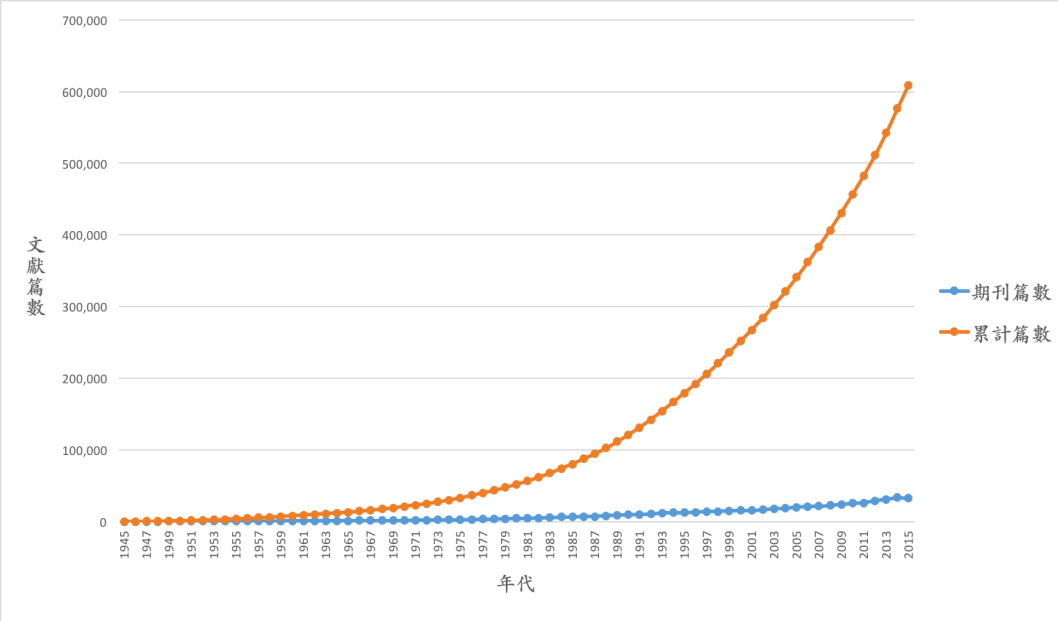
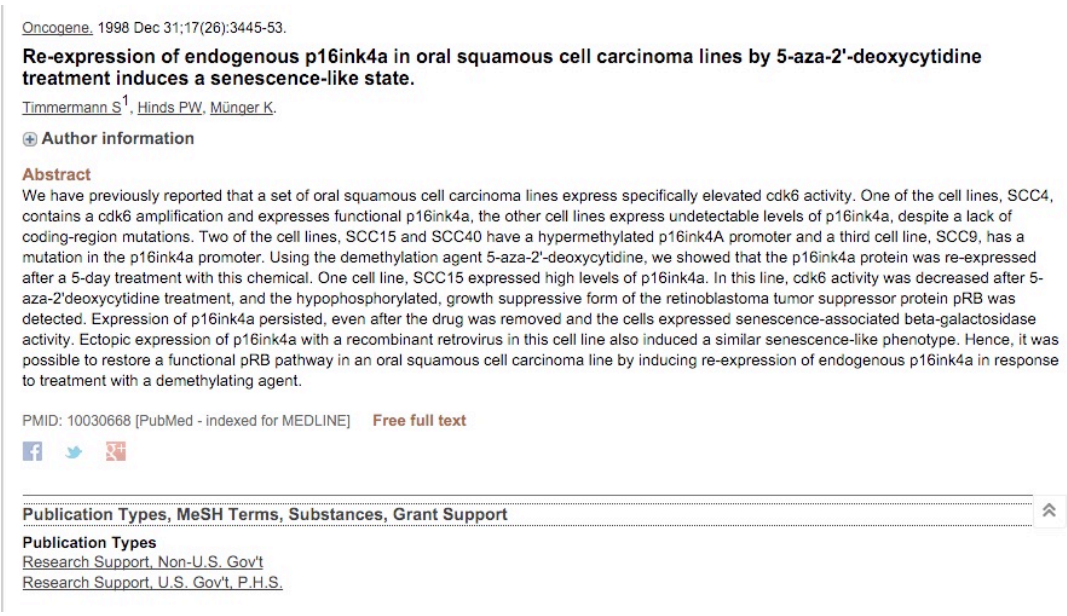


圖 17 癌症研究期刊文獻數量成長曲線圖

資料來源：本研究整理

（二）、出版品類型分佈

本研究共得 608,830 筆文獻，在出版品類型(Publication Type)分類方面，採用 PubMed 資料庫著錄之出版品類型，如圖 18 所示，在 PubMed 資料庫的「Publication Types, MeSH Terms, Substances, Grant Support」一項中，顯示該文獻之出版品類型、主題詞等資訊，本研究經資料整理過程，將其文獻之出版品類型區隔為三種(1)按資料型式區分；(2)按內容型式區分；(3)研究報告類，分述如下：。


Oncogene. 1998 Dec 31;17(26):3445-53.

Re-expression of endogenous p16ink4a in oral squamous cell carcinoma lines by 5-aza-2'-deoxycytidine treatment induces a senescence-like state.

Timmermann S[1], Hinds PW, Münger K.

Author information

Abstract

We have previously reported that a set of oral squamous cell carcinoma lines express specifically elevated cdk6 activity. One of the cell lines, SCC4, contains a cdk6 amplification and expresses functional p16ink4a, the other cell lines express undetectable levels of p16ink4a, despite a lack of coding-region mutations. Two of the cell lines, SCC15 and SCC40 have a hypermethylated p16ink4A promoter and a third cell line, SCC9, has a mutation in the p16ink4a promoter. Using the demethylation agent 5-aza-2'-deoxycytidine, we showed that the p16ink4a protein was re-expressed after a 5-day treatment with this chemical. One cell line, SCC15 expressed high levels of p16ink4a. In this line, cdk6 activity was decreased after 5-aza-2'deoxycytidine treatment, and the hypophosphorylated, growth suppressive form of the retinoblastoma tumor suppressor protein pRB was detected. Expression of p16ink4a persisted, even after the drug was removed and the cells expressed senescence-associated beta-galactosidase activity. Ectopic expression of p16ink4a with a recombinant retrovirus in this cell line also induced a similar senescence-like phenotype. Hence, it was possible to restore a functional pRB pathway in an oral squamous cell carcinoma line by inducing re-expression of endogenous p16ink4a in response to treatment with a demethylating agent.

PMID: 10030668 [PubMed - indexed for MEDLINE] Free full text

Publication Types, MeSH Terms, Substances, Grant Support

Publication Types

Research Support, Non-U.S. Gov't

Research Support, U.S. Gov't, P.H.S.

圖 18 PubMed 資料庫著錄之出版品類型

資料來源：Timmermann, S., Hinds, P. W., & Munger, K. (1998, Dec 31). Re-expression of endogenous p16ink4a in oral squamous cell carcinoma lines by 5-aza-2'-deoxycytidine treatment induces a senescence-like state. Oncogene. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/10030668

（三）、按資料型式區分

按資料型式區分的種類分布，如表 24 所示，以期刊文獻為最多(Journal Article , 574,924 篇)，其次依序為文獻回顧(含系統性回顧)(Review , 61,463 篇)、英文文摘(English Abstract , 27,403 篇)、評論(Comment , 17,603 篇)、信件(Letter ,

17,017 篇) ，該五類為出版篇數超過萬篇的種類；其次在出版篇數低於萬篇但高於千篇(1000<N<10000)方面，共有四類，分別為編輯的話(Editorial , 8,605 篇)、新聞(News , 4,147 篇)、醫學會議(Congresses , 2,212 篇)、歷史文章(Historical Article , 1,912 篇)。

第三區為出版篇數低於千篇(N<1000)，該區類型較為分散，共有 30 種，分別為：傳記(Biography , 910 篇)、專輯(Overall , 858 篇)、介紹性期刊文章(Introductory Journal Article , 742 篇)、實務指南(Practice Guideline , 548 篇)、書目(Bibliography , 493 篇)、肖像(Portraits , 363 篇)、指南(Guideline , 354 篇)、已出版的勘誤(Published Erratum , 309 篇)、發展共識研討會(Consensus Development Conference , 308 篇)、出版品撤回(作者的撤回聲明)(Retraction of Publication , 205 篇)、撤回的出版品(Retracted Publication , 198 篇)、目錄(Directory , 196 篇)、訪談(Interview , 145 篇)演講(Lectures , 140 篇)、地址(Addresses , 47 篇)、經典文章(適用於再版的研討會文章)(Classical Article , 40 篇)、發展共識研討會(NIH)(Consensus Development Conference, NIH , 39 篇)、已修正與再版文章(Corrected and Republished Article , 39 篇)、紀念論文集(Festschrift , 27 篇)、自傳(Autobiography , 26 篇)、重複出版品(Duplicate Publication , 22 篇)、病患教育傳單(Patient Education Handout , 21 篇)、網路廣播(Webcasts , 20 篇)、期刊索引(Periodical Index , 15 篇)、個人生平(Personal Narratives , 7 篇)、影音媒體(Video-Audio Media , 7 篇)、法規(Legislation , 3 篇)、法律案例(Legal Cases , 2 篇)、字典(Dictionary , 1 篇)。

表 24 癌症研究領域出版品類型分佈-按資料型式區分

出版篇數(N)	資料型式

出版篇數(N)	資料型式
(N>=10000)	期刊文獻(574,924)、文獻回顧(含系統性回顧)(61,463)、英文文摘(27,403)、評論(17,603)、信件(17,017)
(1000<N<10000)	編輯的話(8,605)、新聞(4,147)、醫學會議(2,212)、歷史文章(1,912)
(N<1000)	傳記(910)、專輯(858)、介紹性期刊文章(742)、實務指南(548)、書目(493)、肖像(363)、指南(354)、已出版的勘誤(309)、發展共識研討會(308)、作者的撤回聲明(205)、撤回的出版品(198)、目錄(196)、訪談(145)、演講(140)、地址(47)、NIH 內部發展共識研討會(39)、已修正與再版文章(39)、紀念論文集(27)、自傳(26)、重複出版品(22)、病患教育傳單(21)、網路廣播(20)、期刊索引(15)、個人生平(7)、影音媒體(7)、法規(3)、法律案例(2)、字典(1)

資料來源：本研究整理

（四）、按內容型式區分

本研究以內容型式區分，參見同樣以出版篇數區隔，篇數為萬篇以上的共有六類，以對比研究(Comparative Study)為最多，計有 46,698 篇；其次為案例報告(Case Reports，27,677 篇)、臨床試驗(Clinical Trial，25,648 篇)、跨中心研究(Multicenter Study，13,859 篇)、隨機控制試驗(Randomized Controlled Trial，13,396 篇)、臨床試驗第二期(Clinical Trial, Phase II，10,471 篇)；其次篇數在千篇以上，但不滿萬篇的種類，共有六類，以臨床試驗相關研究為最多，分別為臨床試驗第一期(Clinical Trial, Phase I ，6,199 篇)，其次為評估研究(Evaluation Studies，4,908 篇)、統合分析(Meta-Analysis，3,912 篇)、臨床試驗第三期(Clinical Trial, Phase III，2,595 篇)、對照臨床試驗(Controlled Clinical Trial，1,757 篇)、驗證研究(Validation Studies，1,380 篇)。

篇數相對較少的文獻內容型式同樣共有六類，分別為觀測研究

(Observational Study，212 篇)、適用於雙胞胎的研究(Twin Study，72 篇)、臨床案例研討會(Clinical Conference，59 篇)、臨床試驗第四期(Clinical Trial, Phase IV，49 篇)、技術報告(Technical Report，17 篇)、務實臨床試驗(Pragmatic Clinical Trial，1 篇)。

表 25 癌症研究領域出版品類型分佈－按內容型式區分

出版篇數(N)	內容型式
(N>=10000)	對比研究(46,698)、案例報告(27,677)、臨床試驗(25,648)、跨中心研究(13,859)、隨機控制試驗(13,396)、臨床試驗第二期(10,471)
(1000<N<10000)	臨床試驗第一期(6,199)、評估研究(4,908)、統合分析(3,912)、臨床試驗第三期(2,595)、對照臨床試驗(1,757)、驗證研究(1,380)
(N<1000)	觀測研究(212)、適用於雙胞胎的研究(72)、臨床案例研討會(59)、臨床試驗第四期(49)、技術報告(17)、務實臨床試驗(1)

資料來源：本研究整理

（五）、研究報告類

在研究報告方面，以研究計畫的計畫經費來源分類共有六種，以非美國政府資助的研究報告為最多，計有 251,024 篇，(Research Support, Non-U.S. Gov't , 251,024 篇)，其次為美國政府公共衛生署資助(Research Support, U.S. Gov't, P.H.S. , 64,728 篇)、美國國家衛生研究院院外研究資助(Research Support, N.I.H., Extramural , 45,460 篇)、非美國政府公共衛生署資助 (Research Support, U.S. Gov't, Non-P.H.S. , 18,021 篇、美國國家衛生研究院院內研究資助(Research Support, N.I.H., Intramural , 3,118 篇)，最後則為美國經濟復甦暨再投資法(Research Support, American Recovery and Reinvestment Act , 61 篇)，參見表 26。

表 26 癌症研究領域出版品類型分佈－研究報告類

出版品類型	篇數
Research Support, Non-U.S. Gov't	251,024
Research Support, U.S. Gov't, P.H.S.	64,728
Research Support, N.I.H., Extramural	45,460
Research Support, U.S. Gov't, Non-P.H.S.	18,021
Research Support, N.I.H., Intramural	3,118
Research Support, American Recovery and Reinvestment Act	61

資料來源：本研究整理

二、 布萊德福定律驗證

本研究蒐集 608,830 筆文獻分布於 172 種期刊。根據以布萊德福定律 1：n：n^2：n^3 之計算公式進行分區，統計每一種期刊所刊登之文獻篇數，並依照篇數多寡遞減排序、就期刊數(A)、文獻數(B)、累積期刊數(C)累積文獻數(D)、累積期刊數的常用對數值(E)，製成癌症研究文獻分布統計表。本研究取 608,830 的 1/4 約 152,207.5 作為分界線，將期刊分布依累積文獻數分為 4 區，可得前 4 種期刊即刊登了佔文獻總數的 25%(131,773 篇，佔 21.64%)，其餘約 75%的文獻則分布於 168 種期刊。

本研究將 172 種期刊分成四區，各區含約等量的文章篇數，如表 27 所示，第一區有 4 種期刊，共出版 131,773 篇文獻，每種期刊刊載 21,542 至 49,063 篇文獻；第二區有 10 種期刊，共出版 164,587 篇文獻，每種期刊刊載 11,583 至 21,258 篇文獻；第三區有 25 種期刊，共出版 158,407 篇文獻，每種期刊刊載 4,100 至 10,719 篇文獻；第四區有 133 種期刊，共出版 154,063 篇文獻，每種期刊刊

載 1 至 3,998 篇文獻。各區期刊種數的比例為 4：10：25：133，若以布萊德福定律 $1:n:n^2:n^3$ 之計算公式，求得 n 值為 2.5，亦即各區之比大約為 $1*4:2.5*4:2.5^2*4:2.5^3*4$，等於 4：10：25：62.5。前三區的比值符合布萊德福定律，但到第四區時，因癌症研究期刊文獻趨於分散，比值約為布萊德福定律期刊數的 2 倍。

表 27 癌症研究領域期刊文獻分區表

分區	期刊總數	文獻總數	出版之文獻數
一	4	131,773	21,542 至 49,063 篇
二	10	164,587	11,583 至 21,258 篇
三	25	158,407	4,100 至 10,719 篇
四	133	154,063	1 至 3,998 篇
總計	172	608,830	

資料來源：本研究整理

三、 布萊德福－齊夫定律驗證

本研究依布萊德福－齊夫定律之規則，繪製癌症研究文獻半對數分佈圖，如圖 19 所示。圖 19 顯示大約前 14 種期刊是曲線分佈，形成所謂核心期刊，連接的相關區則呈現線性成長，大概排名第 15 名到 100 名的曲線，最後邊緣區的圖形曲線之尾端略為向右下偏垂，整體圖形與布萊德福－齊夫標準的 S 型曲線頗為相似，但是 S 型並不明顯。

由上述方式之驗證，可歸納出癌症研究領域之文獻和布萊德福定律相似，但並不完全符合，而少數的期刊刊載文獻佔總體文獻相當大的比例，前 15 種期刊即刊載了半數以上(307,079 篇，佔 50.43%)，至於其他 49.57%的文獻則分佈於 157 種期刊，其中有 72 種期刊的刊載數小於 1000 篇，顯示在 1945~2015 年

的癌症研究文獻分佈相當分散。

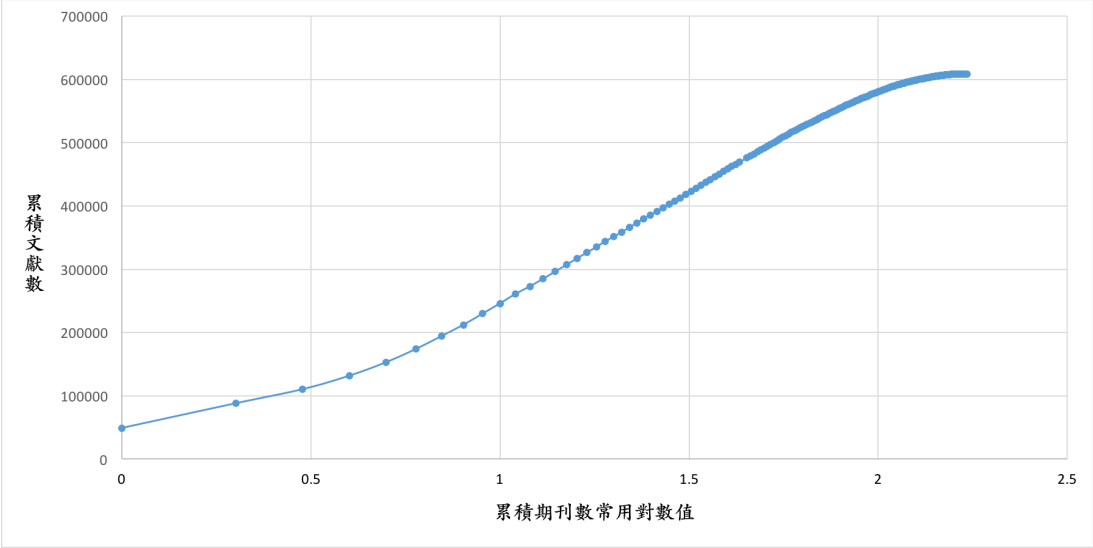


圖 19 癌症研究領域文獻半對數分布圖

資料來源：本研究整理

四、 洛卡定律驗證與作者特性

(一)、洛卡定律

洛卡定律(Lotka's Law) 其公式為 $f(x)=c/x^n$，本研究藉統計方法，分析作者之分佈情形。並以最小平方法進行著作量與作者分析，由於洛卡定律計算之原則為 senior author，在過去相關研究中，如蔡明月與鄭琚媛（2005）採取第一作者計算，而楊燕枝（2010）認為文獻名列其上的作者，無論是第一作者或合著作者，皆視為同等重要，故採用全部作者進行計算，故本研究兩種計算方式皆採用，故採取(1)第一作者；(2)全部作者；分別計算，詳如下述。

1. 第一作者

本研究之 608,830 篇文獻，扣除未著錄作者之文獻紀錄，共有 603,630 篇文獻，計有 293,216 位作者，根據洛卡定律的定義，發表一篇文章的作者人數應為全部作者數的 60.79%，換言之，應為 293,216 * 0.6 = 175,929 人，而實際觀

察人數為 190,912 人，約為全體的 65.10%，由此可見，發表一篇癌症研究的相關文獻作者佔作者人數的百分比略大於洛卡定律的 60.79%。

其次，為求得符合洛卡定律之 n 值與 c 值，本研究擬以 Pao 的公式，先將文獻量與相對的作者數轉換為自然對數 ln(X)與 ln(Y)加以計算。採用最小平方法帶入 n 值，再帶入 Pao 的演算公式以計算 c 值。將上述計算值分別帶入最小平方法的公式以估計 n 值。

$$n = \frac{K\sum \ln(X)\ln(Y) - \sum \ln(X)\ln(Y)}{K\sum \ln(X)^2 - \sum(\ln(X))^2}$$

得 n=96*(-2053.0019)-(350.09)*(-491.943)/96*(1369.0678)-(350.09)^2 = -2.804

由 n=-2.084，證明此為一由左上向右下走的直線。再分別帶入(1)洛卡定律公式與(2)Pao 的計算公式，分別計算 C 值：

(1)將 n 值代入公式$C = 1/\sum \frac{1}{x^n}$ 得 C=0.8031，即 C=80.31%。

(2) 以 Pao 的公式計算：

$$C = 1/[\sum_1^{p-1} \frac{1}{x^n} + \frac{1}{(n-1)(p^n-1)} + \frac{1}{2p^n} + n/24(p-1)^{n+1}]$$

P＝20，則 C=0.8022，即 C=80.22%。表示僅發表一篇的癌症研究領域文獻的作者人數，約占作者總數的 80%。且 C 值採取洛卡與 Pao 的公式分別計算，C 值差異不大，同時驗證當觀察樣本(P)設定為 20，C 值的計算結果亦相當接近。但本研究所得之 n 值為-2.084，該值與洛卡計算之-1.888 值並不十分吻合；常數值 80.22 亦與洛卡估計的 60.79 與觀察值的 65.10 相差甚多，依此數據看來洛卡

定律似乎不適用於癌症研究領域。

由於洛卡定律又稱為科學生產力的倒平方律，其內涵為:「發表 x 篇文獻的作者約為發表一篇文獻的作者總數除以 x^2」，洛卡根據計算推導得出公式為 $f(x)=c/x^2$，f(x)為發表 x 篇的作者佔全部作者總數的百分比，c 為常數即為僅發表一篇文獻的作者所佔的百分比，經洛卡計算 c 值為 0.6079，故一篇文獻的作者約為全部作者的 60.79%，其公式亦可改寫為 $f(x)=0.6079/x^2$。

從洛卡定律與相關研究可知其斜率(n)非固定為 2，在資料量龐大時，其計算結果會有較分散之現象發生，故須修正觀察值(K)與資料樣本數，始能計算其斜率值(n)與常數(C)，並驗算其發表文獻總數，因此本研究以作者總人數的 98%為樣本數，約為 287,929 人，對應觀察值為 10(K=10)，設定斜率值(n)為 1.975(n=1.975)，代入洛卡定律計算公式，其計算結果，其常數(C)為 0.6387，顯示發表一篇文獻的作者所佔的百分比為 63.87%，再計算相對應的文獻總數，如表 28 所示。

從表 28 可知，當作者人數為 287,929 人，在 n=1.975 時，發表一篇文獻的作者約為全部作者的 63.88%，而發表二篇文獻的作者佔全部作者的 16.25%，而發表三篇文獻的作者則為 7.30%，較符合洛卡定律的推導結果，且所需文獻總數為 572,286 篇，此項推導結果適用於分析癌症研究領域文獻的作者生產力，亦可將公式改為 $f(x)=0.6387/x^{1.975}$。

表 28 洛卡定律之計算數據-癌症研究領域期刊(第一作者)(n=1.975 c= 0.6387)

著作數量		作者人數	文獻觀察值	作者比例	作者數	文獻數	作者比例	N 值
種數(K)	數量(X)	(P)	(A1)=(X)*(P)	(Y1)	(P2)	(A2)	(Y2)	(Z)
1	1	190,912	190,912	66.31	190,912	190,912	63.88	1.000
2	2	48,572	97,144	16.87	48,562	97,125	16.25	0.254
3	3	20,330	60,990	7.06	21,803	65,409	7.30	0.114

4	4	10,544	42,176	3.66	12,353	49,411	4.13	0.065
5	5	6,262	31,310	2.17	7,950	39,750	2.66	0.042
6	6	4,015	24,090	1.39	5,546	33,276	1.86	0.029
7	7	2,755	19,285	0.96	4,090	28,633	1.37	0.021
8	8	1,985	15,880	0.69	3,142	25,137	1.05	0.016
9	9	1,453	13,077	0.50	2,490	22,410	0.83	0.013
10	10	1,101	11,010	0.38	2,022	20,222	0.68	0.011
總和	55	287,929	505,874		298,871	572,286		1.565

資料來源：本研究整理

註. Y1=(作者人數(P)/總人數(287,929)*100;C=1/1.565 = 0.6387;Z=1/$(X)^{1.975}$

作者樣本數比例: 287,929 / 293,216 = 0.9819

文獻樣本數比例：505,874 / 603,680 = 0.8379

2. 全部作者

本研究再以 603,630 篇文獻的全部作者計算洛卡定律之計算，計有 1,045,486 作者，同樣以洛卡定律的定義，發表一篇文章的作者人數應為全部作者數的 60.79%，則應為 1,045,486 * 0.6 = 627,291 人，而實際觀察值為 587,662 人，約為全體的 56.20%，若採以文獻中全部著錄的作者計算，發表一篇癌症研究的相關文獻作者佔作者人數的百分比小於洛卡定律的 60.79%。

其次，同樣以 Pao 的公式，先將文獻量與相對的作者數轉換為自然對數 ln(X) 與 ln(Y)加以計算。採最小平方法帶入 n 值，再採用 Pao 的演算公式以計算 c 值。將上述計算值分別帶入最小平方法的公式以估計 n 值。

$$n = \frac{K\sum \ln(X)\ln(Y) - \sum \ln(X)\ln(Y)}{K\sum \ln(X)^2 - \sum(\ln(X))^2}$$

得 n=294*(-9834)-(1391)*(-1914) / 294*(6878)-(1391)^2 = -2.639

由 n=-2.639，同樣可以證明此為一由左上向右下走的直線，再將 n 值分別帶入(1)洛卡定律公式與(2)Pao 的計算公式，分別計算 C 值：

(1) 將 n 值代入公式$C = 1/\sum \frac{1}{x^n}$ ，得 C=0.7736，即 C=77.36%

(2) 以 Pao 的公式計算：$C = 1/[\sum_1^{p-1} \frac{1}{x^n} + \frac{1}{(n-1)(p^n-1)} + \frac{1}{2p^n} + n/24(p-1)^{n+1}]$

P=20，則 C=0.7687，即 C=76.87%，表示僅發表一篇的癌症研究領域文獻的作者人數，約占作者總數的 76%，C 值採取洛卡與 Pao 的公式分別計算，同時驗證當觀察樣本(P)設定為 20，其計算結果亦相當接近。但本研究所得之 n 值為-2.639，同樣也與洛卡計算之-1.888 值並不十分吻合；常數值 77.36 與洛卡估計的 60.79 和觀察值的 56.20 略有差異，依此數據同樣似乎不適用於癌症研究領域。

由於在資料量龐大時，其計算結果會較分散之現象發生，故本研究修正觀察值(K)與資料樣本數，始能計算其斜率值(n)與常數(C)，並驗算其發表文獻總數，若每一作者的著作貢獻皆視為相同，則文獻總數為：文獻數量(X)*作者人數(P)的總和，為 3,561,107 篇，故本研究以作者總人數的 94%為樣本數，約為 985,429 人，對應觀察值為 10(K=10)，設定斜率值(n)為 1.975，代入其常數(C)為 0.6387，顯示一篇文獻的作者所佔的百分比為 63.87%，再計算相對應的文獻總數，如表 29 所示。

從表 29 可知，當作者人數為 985,429 人，在 n=1.975 時，發表一篇文獻的作者約為全部作者的 63.88%，而發表二篇文獻的作者佔全部作者的 16.25%，而發表三篇文獻的作者則為 7.3%，較符合洛卡定律的推導結果，且所需文獻總數為 1,761,602 篇，此項推導結果適用於分析癌症研究領域文獻的作者生產力。

表 29 洛卡定律之計算數據-癌症研究領域期刊(全部作者)(n=1.975 c= 0.6387)

著作數量		作者人數	文獻觀察值	作者比例	作者數	文獻數	作者比例	N
種數(K)	數量(X)	(P)	(A1)=(X)*(P)	(Y1)	(P2)	(A2)	(Y2)	1.975

著作數量		作者人數	文獻觀察值	作者比例	作者數	文獻數	作者比例	N
種數(K)	數量(X)	(P)	(A1)=(X)*(P)	(Y1)	(P2)	(A2)	(Y2)	1.975
1	1	587,662	587662	59.64	587,662	587,662	63.88	1.0000
2	2	169,815	339630	17.23	149,484	298,967	16.25	0.2544
3	3	81,180	243540	8.24	67,114	201,342	7.30	0.1142
4	4	47,439	189756	4.81	38,024	152,096	4.13	0.0647
5	5	31,136	155680	3.16	24,472	122,358	2.66	0.0416
6	6	21,675	130050	2.20	17,072	102,431	1.86	0.0291
7	7	16,253	113771	1.65	12,591	88,137	1.37	0.0214
8	8	12,508	100064	1.27	9,672	77,378	1.05	0.0165
9	9	9,954	89586	1.01	7,665	68,983	0.83	0.0130
10	10	7,807	78070	0.79	6,225	62,248	0.68	0.0106
總和	55	985,429	2,027,809		919,980	1,761,602		1.5655

資料來源：本研究整理

註.Y=(作者人數(P)/總人數(985,429)*100;C=1/1.5655=0.6387

作者樣本數比例：985,429/1,045,486=0.9425

文獻樣本數比例：2,027,809/3,561,107=0.5694

驗算文獻樣本數比例：=1,761,602/3,561,107=0.4946

由上述可知，由於洛卡定律所規定的「Senior Author」，並未明確指明為第一作者(First Author)或為最後作者(Last Author)，為求符合癌症研究領域之洛卡定律的公式值，本研究採以第一作者與全部作者分別進行計算，原始結果分別為 $f(x)=0.8031/x^{2.804}$ 與 $f(x)=0.7736/x^{2.639}$ ，由於資料量龐大易導致過度分散的現象發表，故調整斜率值(n)與常數(C)，公式修正為 $f(x)=0.6387/x^{1.975}$，可符合洛卡定律。

(二)、作者特性分布

在總數 608,830 篇癌症研究領域文獻中，總計有 5,150 篇無作者標示，扣除

無作者文獻，共計有 603,680 篇文獻，進行作者數量與合著情形及高產量者分析。

1. 文獻作者數量與合著情形

在總計 603,680 篇文獻，合計有 3,561,105 筆作者紀錄，在不考慮同名同姓與英文縮寫的情況，經去重複後的計算結果，總計有 1,045,486 位作者，平均一篇文獻約有 5 位作者共同撰寫(3,561,105/603,680=5.89)。

在癌症研究領域之 603,680 篇文獻，總計 1,045,486 位作者中，有 53,320 篇文獻是單一作者，佔總數的 8.83%；11.22%的文獻為二人合著(67,715 篇)；11.34%的文獻為三人合著(68,431 篇)；11.27%的文獻為四人合著(68,017 篇)；10.88%的文獻為五人合著(65,699 篇)；10.12%的文獻為六人合著(61,086 篇)，詳如表 30。顯示癌症研究領域之文獻多數為合著作者型式，個人作者的發表偏低，顯見合著作者的型式以 2 至 6 位作者合著為最多，其比例佔總數的五成(54.822%)；七人以上合著的比例占總數的 36.35%(219,412 篇)。

表 30 癌症研究領域文獻作者合著數量統計表

合著作者數	文獻篇數	百分比(%)
一人作者	53,320	8.83%
二人作者	67,715	11.22%
三人作者	68,431	11.34%
四人作者	68,017	11.27%
五人作者	65,699	10.88%
六人作者	61,086	10.12%
七人以上作者合計	219,412	36.35%
總計	603,680	100.01%

註：百分比取到小數第三位四捨五入

資料來源：本研究整理

2. 高產量作者分析

本研究將癌症研究領域文獻中全數1,045,486位作者進行分析，以Takahashi, T發表文獻為最多，計有559篇；其次是La Vecchia, C，計有543篇；第三以下依序是Takahashi, M，計有514篇；Zhang, Wei，514篇；Sugimachi, K，447篇；Wang, Y，435篇；Tanaka, T，423篇；Wang, Wei，412篇；Hortobagyi, Gabriel N，412篇；Franceschi, S，397篇，如表 31所示。

表 31 前十名高產量作者與發表篇數

作者縮寫	發表篇數
Takahashi, T	559
La Vecchia, C	543
Takahashi, M	514
Zhang, Wei	514
Sugimachi, K	447
Wang, Y	435
Tanaka, T	423
Wang, Wei	412
Hortobagyi, Gabriel N	412
Franceschi, S	397

資料來源：本研究整理

第陸章　研究結論與建議

本研究為解決電子資源使用分析上的困境，採取實證研究方法，建構醫學電子資源大數據分析模式，並針對(1)電子資源使用分析；(2)醫學領域主題分析；(3)醫學領域書目計量等不同項目進行探討，以下提出本研究之結論與未來建議。

第一節　研究結論

一、提出圖書館電子資源大數據分析模式。

本研究依據現行的困境與實際需求，提出電子資源大數據分析模式，再以醫學電子資源做為研究個案進行模式驗證，模式中如何分析原始使用數據，建立書目及使用紀錄明細資料庫；以作為圖書館分析探討電子資源使用分析之參考。未來並可將使用資訊整合出如書目計量學、書目探勘與電子資源使用者行為等圖書資訊學相關研究的資料來源，提供更多的整合應用。

二、電子資源使用分析

(一)、使用數據分析顯示集中在少數期刊：

本研究以 2013 年所訂購 Wiley 資料庫為基準，並以 2,331 期刊為研究範圍，在橫跨多個年度的點擊(含下載)次數分析，僅有 833 種期刊之點擊(含下載)次數為至少 1 次，而若以 2013 年單一年度作為時間區隔因素，僅有 562 種期刊曾為

使用者點擊(含下載)，顯示圖書館所訂購之電子期刊，使用者在點擊(含下載)的數據集中在特定少數期刊，未來可協助圖書館在訂購政策上的參考，同時亦可針對使用者的類型與文獻再進行深入探討。

(二)、重要期刊的選擇－透過學院、特定領域科系與期刊交叉分析：

本研究分別以醫學院與公共衛生相關科系為個案說明，顯示醫學院之重要期刊群為 Hepatology、The International Journal of Cell Cloning、Journal of Cellular and Comparative Physiology，其跨年度下載次數皆超過萬次以上；而公共衛生相關科系之重要期刊群為 Health Economics、Arthritis & Rheumatism、Statistics in Medicine、Arthritis Care & Research、International Journal of Cancer、Addiction、Pharmacoepidemiology and Drug Safety、Arthritis & Rheumatology 等 8 種期刊。

(三)、學院單位對特定文獻的需求分析：

本研究以 Hepatology 期刊的「Comprehensive analysis of the independent effect of twist and snail in promoting metastasis of hepatocellular carcinoma」文獻為例，若以跨年度區隔分析，指出該文獻為醫學院與生命科學院之科系較具有點擊(含下載)需求，顯示日後若配合課程或研究需求，圖書館可提供相關建議給該學院或科系，以滿足在研究或課程準備之需求。

三、醫學領域主題分析

(一)、學科電子資源主題分析之架構

過去在進行主題分析之相關研究，研究者須自行依據文獻主題與內容之特

性，透過現有的主題分類基礎，擬定新的分類體系，由於美國國家醫學圖書館所公布的學科主題詞表，為該領域專家學者制訂且每年固定公布新的版本，故本研究採用該分類體系，自行開發網路爬蟲程式下載書目資料，並建置學科主題詞樹、文獻書目與文獻學科主題詞路徑等資料庫，以利於透過自動化方式，在最短時間獲取研究所需之原始資料，並可自動產生統計數據，利於分析大量文獻之主題趨勢。

(二)、採用第三方學科領域資料庫期刊清單，利於區隔研究主題範圍

PubMed 生物醫學文獻資料庫為免費提供生物醫學文獻之搜尋引擎，並未提供期刊評比與領域分類之功能，本研究為以醫學學科主題為個案驗證，採用 SJR 透過期刊評比資料庫所公布期刊清單，並以癌症研究做為驗證個案，至 PubMed 生物醫學文獻資料庫進行檢索，研究結果顯示在癌症研究領域的相關期刊，以 2012 年公布之版本為，共有 196 種期刊為 SJR 期刊評比資料庫收錄，其中達半數以上的期刊為 SJR 期刊評比資料庫和 PubMed 生物醫學文獻資料庫同時收錄。

四、醫學領域書目計量

文獻資料收集的完整性，為進行書目計量學相關研究的困難點之一，亦影響計量研究之分析結果，本研究採以期刊評比資料庫-SJR 資料庫，癌症研究相關領域期刊為個案，並以 PubMed 資料庫做為文獻資料收集來源，以書目計量學三大定律進行驗證，研究結果顯示：(1)文獻成長量突破每年千篇，並呈現逐步升高的趨勢；(2)文獻出版類型以期刊文獻(Journal Article)為最多；(3)癌症研究領域之文獻多數為合著作者型式，個人作者的發表偏低，顯見合著作者的型

式以 2 至 6 位作者合著為最多；(4)採用布萊德福定律驗證，結果顯示前三區的比值符合布萊德福定律，但到第四區時，因癌症研究期刊文獻趨於分散，比值約為布萊德福定律期刊數的 2 倍；(5)針對作者生產力分析，由於洛卡定律所規定的「Senior Author」，並未明確指明為第一作者(First Author)或為最後作者(Last Author)，本研究採以第一作者與全部作者分別進行計算，由於資料量龐大易導致過度分散的現象發表，故調整斜率值(n)與常數(C)，公式修正為 $f(x)=0.6387/x^{1.975}$，可符合洛卡定律。

第二節　研究建議

本研究之研究建議分為(1)建議圖書館自行建置代理伺服器；(2)自行建置讀者檔之備援資料庫；(3)建立期刊資料之沿革歷史檔；(4)透過雲端技術儲存與分析巨量點擊下載資訊；(5)應用領域宜延伸至其他學科領域；(6)配合相關專業領域，執行專家訪談；(7)應針對不同學科領域訂定索引典；(8)圖書館在電子資源主題分析扮演之角色等項目，分述如下：

一、建議圖書館自行建置代理伺服器

在身份驗證層面，本研究提出之建議有二：(1)圖書館自行建置代理伺服器；(2)限縮 IP 驗證範圍，僅將代理伺服器做為對外連至電子資料庫惟一合法 IP。實務上，圖書館為考量讀者的便利性，在訂購電子資料庫與電子期刊時，會將屬於學校或機構內的 IP 範圍通知廠商，協助設定納入合法使用的網路 IP 範圍內，亦指若讀者在校園內或機構內的網路連線至電子資料庫，僅驗證是否為合法網路 IP 範圍，無須透過身份驗證即可檢索與下載文獻，但從圖書館管理

的角度，讀者若未透過圖書館提供的代理伺服器進行身份驗證，圖書館則無法得知讀者的點擊下載狀況，故建議圖書館自行建置代理伺服器，無論讀者是否使用校園網路或在家連線，皆須連線至圖書館的代理伺服器，取得合理身份驗證之後再連至電子資料庫，藉以取得合法讀者之點擊下載紀錄。

二、自行建置讀者檔之備援資料庫

圖書館自動化系統內存的讀者檔，因讀者資料隨著時間、情境等因素等進行更改讀者的狀態，如實務上大學圖書館在為學生辦理離校流程時，會將讀者身份類型由「學生」更改為「校友」類型，雖可明確辨識該學生為校友身份，且限制使用者權限，但在進行資料清理之程序時，須多重回溯程序始能得知學生的原始科系，故本研究建議，除原本圖書館自動化系統的原始讀者檔，另行建構讀者檔之備援資料庫，以利資料回溯。

三、建立期刊資料之沿革歷史檔

本研究在進行資料比對過程中，因期刊會有因為改換名稱或是 ISSN 的狀況產生，在電子資料庫廠商提供之年度期刊清單中，雖列出年度訂購之期刊資料，但無法確知期刊之間的關聯性，必須自行依據網址的樣式重新回溯與整理，故建議期刊資料之沿革歷史檔，以供期刊歷史資料之比對，

四、透過雲端技術儲存與分析巨量點擊下載資訊

電子資源使用資訊，其點擊下載資料為即時產生，因資訊量之巨，無法純以人力進行資料處理，但資訊清理之過程中會因大量讀寫而影響效能，亦無法以單一主機或伺服器處理，建議透過雲端主機進行儲存、更新、處理、分析等

動作，一則減少實體主機之效能損耗，另可透過分散式架構進行資料處理，增加資料處理與檢索之效率。

五、應用領域宜延伸至其他學科領域

本研究雖以醫學圖書館為研究個案，仍可適用於其他學科領域，如 IEEE 等，雖其困難點在於最初的原始資料收集、儲存、清理、整合等過程，與圖書館自動化系統與期刊書目等資訊之縱向與橫向整合，透過本研究建構之分析模式，可將巨量之原始使用資料，透過系統化的資料處理與分析方法，減少人工作業，依據圖書館之內部需求，增加分析類型與輸出所需資料，亦可針對其他學科領域進行期刊分類與分析，對於圖書館之管理層面，達到有效管理之目的，在學術研究方面，其輸出資料可應用在資訊計量學、書目探勘、資訊行為等圖書資訊學之研究領域。

六、配合相關專業領域，執行專家訪談

本研究之研究結果為依據代理伺服器網際網路紀錄檔進行分析與統計，對於使用者本身在點擊與下載動機，或進一步的資訊尋求行為並無法直接得知，若能執行專家訪談，能更進一步得知使用者之相關資訊需求，但因個人資料保護法的法律規範，在執行層面上，必須事先取得原始使用者之同意，始能進行。

七、應針對不同學科領域訂定索引典

索引典為學科領域詞彙的一種實作方式，透過特定知識領域詞彙的收集，並以特定的結構排列，進而表達學科領域詞彙之間的關係，須由學科專家制定與從大量文獻進行集合分析，然學科領域未必有專業索引典，故在進行特定學

科領域之主題分析，須先制定該學科領域之索引典，未來在進行相關研究時，始能減少研究者在主題分類的個人主觀性，與確保主題概念表達的一致性。

八、圖書館在電子資源主題分析扮演之角色

由於讀者對於電子資源的仰賴程度與日遽增，亦影響圖書館對於電子資源的重視程度，無論館藏政策、服務推廣等議題皆有其發展的空間，建議圖書館除進行學科領域的電子資源主題分析，亦可結合讀者的電子資源使用分析，以增進參考業務之服務品質。

參考文獻

Jiawei, H.（2008）。 **資料探勘: 概念與方法（王派洲，譯）**。台北市：滄海書局。

Mckibbon, KA.（2008）。 實證醫學原理與應用快速入門 (楊宗翰，譯)。 台北市：合記圖書館出版社。

牛道立、甄俊杰、何芬 （2010）。 近五年 PubMed 收錄的中國放射腫瘤學文獻的分析[Chinese Literatures of Radiation Oncology Covered by PubMed over the Past Five Years]。**癌症， 29**（4），頁 495-499。

王明玲、杜立中、曾彩娥 （2009）。 臺灣出版五四運動文獻之書目計量研究[A Bibliometrics Study on the May Fourth Movement Documents Published in Taiwan]。**國家圖書館館刊**（98 年 1 月），頁 93-113。

王敏、李揚、張燕舞、張玢、阿麗塔、許培揚等人 （2010）。 國際人體肝移植研究的文獻计量分析[Bibliometric Analysis of International Publications on Human Liver Transplantation]。**中華醫學圖書情報雜志， 19**（1），頁 67-73。

王梅玲（2006）。 電子計量應用在電子期刊評鑑[The Application of E-Metrics to Electronic Journal Evaluation]。**圖書與資訊學刊**（57），頁 1-18。

王梅玲 （2008）。 電子期刊網路服務與評鑑模式之研究[A Study of Electronic Journals' Network Service and Evaluation Models]。**圖書資訊學研究， 3**（1），頁 75-102。

何光國（1994）。**文獻計量學導論**。：三民書局。

吳政叡 （2010）。 輔仁大學圖書館 RapidILL 使用分析與經濟效益評估[The Study of RapidILL Usage Analysis and Economic Assessment at Fu-Jen University Library]。**大學圖書館， 14**（2），頁 48-61。 doi: 10.6146/univj.2010.14-2.03

李美燕 （1994）。 線上公用目錄使用行為初探：交易處理記錄分析[A Primary Study of the Use of OPAC: Transaction Log Analysis]。**圖書與資訊學刊**（8），頁 19-36。

肖燕秋、徐奎、李世莉、陳克俠 （2011）。 PubMed 收錄甲型 H1N1 流感文獻分析[Bibliometrics Analysis of Documents on Influenza a H1N1 Influenza in PubMed]。**蚌埠醫學院學報， 36**（1），頁 88-90。

林巧敏（2009）。 臺灣地區資訊組織文獻書目計量分析[A Bibliometric Analysis

on the Literature of Information Organization in Taiwan]。**圖書資訊學刊， 7**（1&2），頁 101-123。 doi: 10.6182/jlis.2009.7(1.2).101

林巧敏、范蔚敏（2010）。臺灣地區檔案學文獻計量分析[A Bibliometric Analysis of the Literature on Archival Studies in Taiwan]。**圖書與資訊學刊**（72），頁 16-38。

林信成、洪銘禪 （2010）。 應用 Google Analytics 於數位典藏網站計量分析[A Web Metrics Study on Taiwan Baseball Wiki Using Google Analytics]。**教育資料與圖書館學， 47**（3），頁 343-369。

林偉翔、林雯瑤 （2012）。 北市醫學雜誌之書目計量分析[A Bibliometric Study of Taipei City Medical Journal]。**北市醫學雜誌， 9**（1），頁 45-60。 doi: 10.6200/tcmj.2012.9.1.05

林頌堅 （2000）。 從網站工作記錄觀察大學圖書館網站使用情形[The Usage of a University Library Site from Observation of Web Log Files]。**大學圖書館， 4**（2），頁 2-22。

林頌堅 （2003）。 基於詞語抽取的圖書與資訊學刊研究主題分析。**圖書與資訊學刊**（47），頁 15-35。

林頌堅 （2008）。 以網站流量資料探討使用者的網路使用行爲[Using Traffic Data of Websites to Explore Users' Behavior in Network Use]。**圖書資訊學研究， 2**（2），頁 39-60。

邱均平、文庭孝、宋艷輝（2014）。**知識計量學**。北京：科學出版社。

洪榮照、鄧英蘭、蘇雅敏 （2010）。 臺中教育大學研究生對電子資源館藏使用情形之研究。**臺中教育大學學報: 教育類， 24**（2），頁 151-172。

紀征、任淑敏、朱啟貞 （2010）。 基於 PubMed 數據庫人禽流感研究的文獻計量分析[PubMed-covered Papers on Human Avian Influenza: A Bibliometric Analysis]。**中華醫學圖書情報雜誌， 19**（6），頁 70-73+79。

徐路 （2007）。 MEDLINE 收錄 1996-2005 年貴陽醫學院論文的統計。**貴陽醫學院學報， 32**（3），頁 315-317。

國家教育研究院（2012）。 圖書館學與資訊科學大辭典-領域分析(Domain Analysis)。取自 http://terms.naer.edu.tw/detail/1678820/

張晗、劉鵬年、崔雷 （2007）。 國外消化性潰瘍文獻研究熱點的分析[Analysis of Foreign Literatures on Peptic Ulcer Research]。**世界華人消化雜誌， 15**（10），頁 1150-1155。

張慧銖（2011）。**圖書館電子資源組織——從書架到網路**。台北市：華藝。

張慧銖、黃婷鈺 （1996）。 統一醫學語言系統簡介[Introduction to the UMLS]。

圖書館學刊（11），頁 165-204。

許昌泰、王明華、徐禮鮮 （2010）。 PubMed 中麻醉學研究文獻計量學分析[Bibliometric Analysis and Developing of Biomedical Publications on Literature Anaesthesia in PubMed]。**現代情報， 30**（5），頁 3-8。

郭麗芳 （1997）。 網路電子期刊評估之研究[A Research on the Evaluation of Electronic Journals on the Network]。**大學圖書館， 1**（3），頁 73-105。

陳明來 （2002）。 主題分析在圖書資訊組織之角色探討[A Study of the Role of Subject Analysis in the Information Organization]。**圖書與資訊學刊**（42），頁 69-84。

陳敏珍 （1994）。 主題分析與主題檢索初探。**國立臺灣師範大學圖書館館訊， 14**，頁 4-8。

黃慕萱、董蕙茹、王俊傑 （2013）。 數位典藏網站評估－以網路連結與網站記錄分析[Web Link and Transaction Log Analyses of Digital Archive Websites]。**圖書館學與資訊科學， 39**（2），頁 66-83。 doi: 10.6245/jlis.2013.392/612

黃慕萱、賴麗香 （2009）。 圖書資訊學門國科會專題研究計畫之書目計量分析[A Bibliometric Study of the Funding Research Projects on LIS of the NSC in Taiwan]。**圖書與資訊學刊**（70），頁 1-16。

楊喻翔、釋惠敏 （2011）。 安寧療護文獻之計量研究：1952-2009[A Bibliometric Study of Palliative Care Literature,1952-2009]。**安寧療護雜誌， 16**（1），頁 42-61。

楊燕枝（2010）。 以文獻計量法分析台灣在國際科管期刊的發表現況[Academic Publication of Taiwanese Scholars in Top International Journals of Technology Management Discipline: Biliometric Approach]。**科技管理學刊， 15**（4），頁 61-86。

葉謙歆 （2006）。 從電子資源統計問題看 COUNTER 計畫之發展與應用[Statistical Problems of Electronic Resources and the Development and Application of COUNTER Project]。**教育資料與圖書館學， 43**（4），頁 453-469。

詹麗萍（2005）。**電子資源與圖書館館藏發展**。台北市：麥田。

翟俊霞、劉玉華、安翠平 （2001）。 MEDLINE 數據庫(1982～1998 年)中原癌基因 ras 的主題詞分析。**醫學信息學雜誌**（2001 年 03），頁 46-50。

蔡明月（2003）。**資訊計量學與文獻特性**。台北市：華泰文化。

蔡明月、徐韻婷 （2009）。 臺灣與中國大陸圖書資訊學核心期刊論文之主題分

析研究－以 1997-2006 年為例[An Analysis of Subjects and Citations in Core Journals of Library and Information Science in Taiwan and China from 1997 to 2006]。**圖書與資訊學刊**（70），頁 17-38。

蔡明月、鄭琚媛 （2005）。 台灣地區生命科學國際會議文獻生產力研究[A Productivity Study of International Conference Literature on Life Science in Taiwan]。**圖書與資訊學刊**（53），頁 1-17。

蔡明月、賴芊卉 （2007）。 資訊科學引用與被引用文獻分散現象與主題變化研究：1985-2005[Comparison of Literature Scattering and Subject Changes between Citing and Cited Literature on Information Science: 1985-2005]。**圖書資訊學研究， 1**（2），頁 1-31。

龍莉艷、張桂云、周丹、劉麗華、馮小東、張磊 （2008）。 肺癌研究文獻的循證計量分析[Evidence-based Bibliometric Analysis of Lung Cancer Research Literature]。**中國循證醫學雜誌， 8**（12），頁 1105-1111。

戴玉旻、陳莉君、偌弗．卡克拉爾、柯皓仁 （2001）。 台灣電子期刊使用者行為分析：以 Elsevier SDOS 電子期刊系統為例[Exploring the Behavior of E-Journal Users in Taiwan: An Analysis of Web Server Log of Elsevier Science Direct On Site]。**圖書與資訊學刊**（39），頁 1-27。

謝建成（2010）。**書目探勘**。台北市：文華圖書館管理。

羅思嘉 （2000）。 資訊行為探索-以國立成功大學學生為例: 從資訊搜尋行為看資訊環境對讀者資訊行為之影響， **64**，頁 141-160。

羅思嘉、陳光華、林純如 （2001）。 圖書資訊學學術文獻主題分類體系之研究[The Study of Framework of Subject Classification for Journal Articles in Library and Information Science]。**圖書資訊學刊**（16），頁 185-207。

Albrechtsen, H. (1993). Subject analysis and indexing: from automated indexing to domain analysis. *Indexer, 18*, 219-219.

Atkins, S. (1996). Mining automated systems for collection management. *Library Administration & Management, 10*(1), 16-19.

Banerjee, K. (1998). Is data mining right for your library? *Computers in Libraries, 18*(10), 28-31.

Borrego, Á., & Urbano, C. (2007). Analysis of the behaviour of the users of a package of electronic journals in the field of chemistry. *Journal of documentation, 63*(2), 243-258.

Boyce, P., King, D. W., Montgomery, C., & Tenopir, C. (2004). How electronic journals are changing patterns of use. *The Serials Librarian, 46*(1-2), 121-141.

Breeding, M. (2001). Offering remote access to restricted resources. *Information Today, 18*(5), 52-53.

Breeding, M. (2009). Measuring the Use of Electronic Resources. *Library Technology Reports, 38*(3), 36-52.

Brookes, B. C. (1969). Bradford's law and the bibliography of science. *Nature, 224*(5223), 953.

Chu, C. M., & O'Brien, A. (1993). Subject analysis: the critical first stage in indexing. *Journal of Information Science, 19*(6), 439-454.

COUNTER. (2002). COUNTER: COUNTER: Counting Online Usage of NeTworked Electronicresources. Retrieved from http://www.projectcounter.org/index.html

COUNTER. (2012, May 28). The COUNTER Code of Practice for e-Resources: Release 4. Retrieved from http://www.projectcounter.org/r4/COPR4.pdf

COUNTER. (2014). The COUNTER Code of Practice for Articles: Release 1. Retrieved from http://www.projectcounter.org/documents/counterart_cop_MAR2014.pdf

Davis, P. M. (2002). Patterns in electronic journal usage: Challenging the composition of geographic consortia. *College & Research Libraries, 63*(6), 484-497.

De Groote, S. L., & Dorsch, J. L. (2003). Measuring use patterns of online journals and databases. *Journal of the Medical Library Association, 91*(2), 231.

Dehdarirad, T., Villarroya, A., & Barrios, M. (2015). Research on women in science and higher education: a bibliometric analysis. *Scientometrics, 103*(3), 795-812. doi:10.1007/s11192-015-1574-x

Fox, J. P., Tyler, J. A., Vashi, A. A., Hsia, R. Y., & Saxe, J. M. (2014). A variation in the value of colectomy for cancer across hospitals: Mortality, readmissions, and costs. *Surgery, 156*(4), 849-860.

Hancock-Beaulieu, M., Robertson, S., & Neilson, C. (1990). *Evaluation of online catalogues: an assessment of methods*: British Library Research and Development Department.

Hjørland, B. (2002). Domain analysis in information science: eleven approaches-traditional as well as innovative. *Journal of documentation, 58*(4), 422-462.

Hjørland, B., & Albrechtsen, H. (1995). Toward a new horizon in information

science: domain-analysis. *JASIS, 46*(6), 400-425.

Hsu, C.-L., & Chiang, C.-H. (2015). A bibliometric study of SSME in information systems research. *Scientometrics, 102*(3), 1835-1865. doi:10.1007/s11192-014-1334-3

Huang, Y., Zhou, M., Deng, Q., Zhang, J., Zhou, P., & Shang, X. (2015). Bibliometric analysis for the literature of traditional Chinese medicine in PubMed. *Scientometrics, 105*(1), 557-566.

Jansen, B. J., & Spink, A. (2000). Methodological approach in discovering user search patterns through Web log analysis. *Bulletin of the American Society for Information Science and Technology, 27*(1), 15-17.

Khan, A. M., Zaidi, S. M., & Zaffar Bharati, M. S. (2009). Use of on-line databases by faculty members and research scholars of Jawaharlal Nehru University (JNU) and Jamia Millia Islamia (JMI), New Delhi (India): A survey. *The International Information & Library Review, 41*(2), 71-78.

Kurth, M. (1993). The limits and limitations of transaction log analysis. *Library Hi Tech, 11*(2), 98-104.

Lawrence, S., Giles, L. C., & Bollacker, K. (1999). Digital libraries and autonomous citation indexing. *Computer, 32*(6), 67-71.

Li, B., Pan, H., & Ye, D. (2013). A bibliometric study of literature on SLE research in PubMed (2002–2011). *Lupus, 22*(8), 772-777.

Libraries, A. O. R. Statistics & Assessment. Retrieved from http://www.arl.org/focus-areas/statistics-assessment

Lin, J., & Wilbur, W. J. (2007). PubMed related articles: a probabilistic topic-based model for content similarity. *BMC bioinformatics, 8*(1), 423.

Lipscomb, C. E. (2000). Medical subject headings (MeSH). *Bulletin of the Medical Library Association, 88*(3), 265.

Lotka, A. J. (1926). The frequency distribution of scientific productivity. *Journal of Washington Academy Sciences*.

Mancini, D. D. (1996). Mining your automated system for systemwide decision making. *Library Administration & Management, 10*(1), 11-15.

Mercer, L. S. (2000). Measuring the use and value of electronic journals and books. *Issues in science and technology librarianship, 25*, 1-8.

Mork, J. G., Jimeno-Yepes, A., & Aronson, A. R. (2013). *The NLM Medical Text Indexer System for Indexing Biomedical Literature.* Paper presented at the

BioASQ@ CLEF.

Morse, D. H., & Clintworth, W. A. (2000). Comparing patterns of print and electronic journal use in an academic health science library. *Issues in science and technology librarianship, 28*(26.04), 2008.

National Library of Medicine. (2003). Current Semantic Types. Retrieved from http://www.nlm.nih.gov/research/umls/META3_current_semantic_types.html

National Library of Medicine. (2004). UMLS-About. Retrieved from http://www.nlm.nih.gov/research/umls/about_umls.html - Semantic

National Library of Medicine. (2006). UMLS Semantic Network Fact Sheet. Retrieved from http://www.nlm.nih.gov/pubs/factsheets/umlssemn.html

National Library of Medicine. (2013). Introduction to MeSH - 2014. Retrieved from http://www.nlm.nih.gov/mesh/introduction.html

National Library of Medicine. (2014a). History of MTI. Retrieved from http://ii.nlm.nih.gov/MTI/history.shtml

National Library of Medicine. (2014b). MeSH Tree Structures — 2014. Retrieved from http://www.nlm.nih.gov/mesh/2014/mesh_trees/trees.html

National Library of Medicine. (2014c). MeSH-Preface. Retrieved from http://www.nlm.nih.gov/mesh/intro_preface.html - pref_rem

National Library of Medicine. (2014d). MetaMap- A Tool For Recognizing UMLS Concepts in Text. Retrieved from http://metamap.nim.nih.gov

National Library of Medicine. (2014e). PubMed Help - Computation of Related Citations. Retrieved from http://www.ncbi.nlm.nih.gov/books/NBK3827/ - pubmedhelp.Computation_of_Related_Citati

National Library of Medicine. (2014f). Pubmed Help - MeSH Subheadings. Retrieved from http://www.ncbi.nlm.nih.gov/books/NBK3827/table/pubmedhelp.T41/

National Library of Medicine. (2014g). UMLS Reference Manual. Retrieved from http://www.ncbi.nlm.nih.gov/books/NBK9684/

Neighbors, J. M. (1980). *Software construction using components.* Citeseer.

Nelson, S. J., Johnston, W. D., & Humphreys, B. L. (2001). Relationships in medical subject headings (MeSH) *Relationships in the organization of knowledge* (pp. 171-184): Springer.

Nicholson, S. (2003). The bibliomining process: Data warehousing and data mining for library decision making. *Information Technology and Libraries, 22*(4),

146-151.

Nicholson, S. (2006). The basis for bibliomining: Frameworks for bringing together usage-based data mining and bibliometrics through data warehousing in digital library services. *Information processing & management, 42*(3), 785-804.

NISO. (2013). Standardized Usage Statistics Harvesting Initiative (SUSHI)—National Information Standards Organization. Retrieved from http://www.niso.org/workrooms/sushi

Papatheodorou, C., Kapidakis, S., Sfakakis, M., & Vassiliou, A. (2003). Mining user communities in digital libraries. *Information Technology and Libraries, 22*(4), 152.

Patra, S., & Mishra, S. (2006). Bibliometric study of bioinformatics literature. *Scientometrics, 67*(3), 477-489. doi:10.1556/Scient.67.2006.3.9

Peters, T. (1996). Using transaction log analysis for library management information. *Library Administration & Management, 10*(1), 20-25.

Peters, T. A. (1993). The history and development of transaction log analysis. *Library Hi Tech, 11*(2), 41-66. doi:doi:10.1108/eb047884

Pilsczek, F. H. (2015). RESPIRATORY INFECTIONS RESEARCH IN AFGHANISTAN: BIBLIOMETRIC ANALYSIS WITH THE DATABASE PUBMED. *Journal of Ayub Medical College Abbottabad, 27*(2), 464-466.

Pinto, M., Isabel Escalona, M., Pulgarín, A., & Uribe-Tirado, A. (2015). The scientific production of Ibero-American authors on information literacy (1985–2013). *Scientometrics, 102*(2), 1555-1576. doi:10.1007/s11192-014-1498-x

Prieto-Díaz, R. (1990). Domain analysis: An introduction. *ACM SIGSOFT Software Engineering Notes, 15*(2), 47-54.

Pritchard, A. (1969). Statistical bibliography or bibliometrics. *Journal of documentation, 25*(4), 348-349.

Rajgoli, I. U., & Laxminarsaiah, A. (2015). Authorship pattern and collaborative research in the field of spacecraft technology. *Electronic Library, 33*(4), 625-642. doi:10.1108/EL-12-2013-0210

Ramos, J. M., González-Alcaide, G., & Bolaños-Pizarro, M. (2013). Bibliometric analysis of leishmaniasis research in Medline (1945-2010). *Parasites & vectors, 6*(1), 55.

Rowlands, I., & Nicholas, D. (2007). *The missing link: Journal usage metrics.* Paper

presented at the Aslib Proceedings.

Sachithanantham, S., & Raja, S. (2015). Scientometric analysis of rabies research literature in India: 1950–2014. *Scientometrics, 105*(1), 567-575. doi:10.1007/s11192-015-1694-3

SARG. (2002). Squid Analysis Report Generator. Retrieved from http://sarg.sourceforge.net

Scheckler, W. E. (1982). A realistic journal reading plan: The cornerstone of continuing medical education. *JAMA, 248*(16), 1987-1988. doi:10.1001/jama.1982.03330160035021

Schulman, S. (1998). Data mining: Life after report generators. *Information Today, 15*(3), 52.

Sengupta, I. (1985). Bibliometrics: A bird's eye view. *IASLIC Bulletin, 60*(4), 167-174.

Shepherd, P. T. (2006). COUNTER: usage statistics for performance measurement. *Performance Measurement and Metrics, 7*(3), 142-152.

Siddique, M. A., & Ali, A. (2010). Online Journals in the Jawaharlal Nehru University (JNU): A User Survey. *Trends in Information Management, 6*(1).

Vickery, B. C. (1968). Analysis of information *Subject and Information Analysis* (pp. 6-35).

White, A., & Kamal, E. D. (2006). *E-metrics for library and information professionals : how to use data for managing and evaluating electronic resource collections* New York: Neal-Schuman.

Yang, M. H., Chen, C. L., Chau, G. Y., Chiou, S. H., Su, C. W., Chou, T. Y., . . . Wu, J. C. (2009). Comprehensive analysis of the independent effect of twist and snail in promoting metastasis of hepatocellular carcinoma. *Hepatology, 50*(5), 1464-1474.

Yu, L., & Apps, A. (2000). Studying e-journal user behavior using log files: the experience of superjournal. *Library & Information Science Research, 22*(3), 311-338.