國家圖書館

自行研究計畫成果報告

計畫編號：NCL-109-001

年　　度：109年度

執行期限：109年1-12月

計畫名稱：大數據與圖書館及其研究初探

大數據與圖書館及其研究初探

【摘　要】

「大數據」顛覆了各行各業原有的生產與服務模式，數據的增長也將繼續改變並影響許多行業的營運生態，以知識保存、利用與服務為任務的圖書館亦不例外。圖書館面臨將數據轉化為資訊，以建立新服務模式的挑戰；圖書館的角色也已從傳統的人類知識的資訊保存場所發展為資訊或知識服務中心，國外圖書館已開始向數據中心轉型，國外圖書資訊領域對大數據應用、研究，已自理論萌芽、技術探索進入實踐應用階段，反觀國內圖書資訊領域對大數據有關議題的討論及應用亟待開展。鑑於國內圖書資訊領域大數據概念基礎文獻較為缺乏，本文試圖自國內外有關圖書館大數據之主題專書及現況研究文獻，綜整大數據的基本概念、建立圖書館與大數據的連結，並掌握圖書館大數據的研究概況，除為瞭解圖書資訊領域中的大數據，也為未來深入探討圖書館數據管理與應用建立基礎知識背景。

【關鍵字】

圖書館、大數據、數據管理、數據分析

Library; Big Data; Data Management; Data Analysis

壹、前言

大數據一詞由來已久，普遍認為這個概念最早出現在 1980 年出版的《第三次浪潮》中，作者 Alvin Toffler 稱大數據為「第三次浪潮的華彩樂章」（劉文江、胡志丹、賈鳳玲，2019）。大數據，或稱巨量資料、海量資料，是一個許多領域與技術的集合體，其起源最早可能可以追溯到歐洲粒子物理研究中心（Conseil européen pour la recherche nucléaire, CERN）開始處理大量資料的問題（臺灣資料科學協會，無年代）；2008 年 9 月 *Nature* 雜誌出版「Big Data」專刊，探討了大數據（Big Data）對互聯網、網路經濟學、生物醫學等多個領域的影響（王春華、李維、文庭孝，2015）；處理大數據的方法及大數據的領域範圍，一直以來都是各有說法，直到 IBM 在 2010 年將「Big Data」列為專業用語，以及 Gartner 公司首席資料分析師 Doug Laney 在 2012 年給了它一個比較完整的定義後，其內涵與界限才開始明朗。

麥肯錫全球研究院（McKinsey Global Institute）在 2011 年 5 月發布了份研究報告，指出美國醫療、政府、零售、製造和地理這五個行業的大數據，約有 15 個或 17 個公司的數據存儲量都比美國國會圖書館的總存儲量還要多。麥肯錫還估計，數據量每年以 40%的速度增長，到 2009 年至 2020 年期間將增加 44 倍（Affelt, 2015）。美國政府也預測到大數據的策略價值，2012 年 3 月 29 日歐巴馬政府宣布將挹注逾兩億美元推動「大數據研究和發展計畫」（Big Data Research and Development Initiative）並著重於科學、公共衛生、教育、國防，以及美國地理和地質調查領域，旨在推展和改善聯邦政府的數據收集、組織和分析工具及技術，以提高從大量的、複雜的數據集中獲取知識和洞見的能力，把大數據上升到了國家戰略的高度（樊偉紅、李晨暉、張興旺、秦曉珠、郭自寬，2012；Marzullo, 2016；Weiss, 2018）。

被譽為「大數據時代的預言家」的 Viktor Mayer-Schönberger 於 2013 年出版了 *Big Data: A Revolution That Will Transform How We Live, Work, and Think* 一書，作者認為大數據促進了人們思維方式和商業管理等各方面的變革，要盡

可能的收集數據、挖掘其價值，並將其運用於各行各業；書中主要概述了大數據對政治、經濟、社會和專業發展的影響，從思維、商業和管理三個角度分析了大數據引起的社會變革（劉文江、胡志丹、賈鳳玲，2019；蘇玲、婁策群，2019；Blummer & Kentonb, 2019）。有中國學者將 2013 年稱為「大數據元年」，研究發現，在大數據元年之後，大數據的內容日益豐富，文獻量在 2013 年之後也呈現爆發式增長趨勢（劉文江、胡志丹、賈鳳玲，2019）。以「講述英語語言的歷史」為目標的《牛津英語詞典》（Oxford English Dictionary）也在 2013 年 6 月加入了「大數據」一詞，新增的內容引起了對大數據討論的激增，因為牛津詞典的更新總是很有新聞價值（Affelt, 2015）。

「大數據」顛覆了各行各業原有的生產與服務模式，數據的增長也將繼續改變並影響許多行業的營運生態。最典型的大數據應用案例是大數據技術在美國大型超市沃爾瑪（Walmart）的應用，透過對消費者的購物行為數據進行分析，創造了「啤酒與尿布」的經典商業案例並成為最瞭解顧客購物習慣的零售商（楊海燕，2012）。大數據在健康服務、交通運輸、運動及娛樂、大氣科學……的應用對民眾的生活都已有顯著影響，新聞媒體、社群媒體、Amzon、Google 等應用大數據推播個人化的資訊或廣告，相信大家也已習慣且認為這就是資訊服務的日常。大數據被認為是一種資訊資產（De Mauro 等，2016），自數據的取得、存儲、組織、分析與決策之全生命週期皆須有效管理；大數據亦是創造價值的資源（IDC，2012），組織需能從各種數據來源收集並組織數據，從數據中產生有價值的洞察力，提供決策的制定，方能彰顯大數據的價值（Du & Khan, 2020）。大數據已在各個領域得到管理，例如業務決策、預測新的醫療保健趨勢、評估客戶的服務滿意度等， 這些實踐驗證了大數據作為推動組織發展「動力」的價值（Zhan & Wan, 2017）。

在大數據時代裡圖書館似乎尚未融入數據的管理與應用，Noh（2015）認為圖書館正在轉變為 Library 4.0，正是一個可以分析資訊並向讀者呈現發現結果的智慧圖書館，「大數據」被認為是未來圖書館發展的重要相關概念

（Zhan & Widén, 2017）。圖書館館員在大數據環境中需要具備數據素養精通數據」（data savvy），還應能從圖書館數據中獲得及時和有洞察力的管理資訊，以能主動回應讀者不斷變化的需求和符應服務不斷轉型的需要（IFLA,2018）。圖書館已經從傳統的人類知識的資訊保存場所發展為資訊或知識服務中心，國外圖書館已開始向數據中心轉型（Du & Khan, 2020），國外圖書資訊領域對大數據應用、研究，已自理論萌芽、技術探索進入實踐應用階段，反觀國內圖書資訊領域在大數據有關議題的討論及應用上亟待開發。鑑於國內圖書資訊領域大數據概念基礎文獻較為缺乏，本文以關鍵字包含「圖書館」(library)與「大數據」(big data)查檢並閱讀國內外之主題專書及現況研究文獻，綜整大數據的基本概念、建立圖書館與大數據的連結，並掌握圖書館與大數據的研究概況，除為瞭解圖書資訊領域中的大數據，也為未來深入探討圖書館數據管理與應用建立基礎知識背景。

貳、大數據的基本概念

進入大數據的定義討論之前，有必要先就我們已經熟悉的數據（Data）資訊（Information）二個詞彙重新定位，並清楚認識「數據」的內涵。「數據」（或稱資料）是一個非常廣泛的術語，用於各個學科和組織；不同的人、組織、企業和學科數據可以有不同的含義；從廣義上來看，數據因形式（量化或質性）、結構（結構化、半結構化或非結構化）、生產者（原始、二手、三手）、類型（索引的、屬性的、詮釋資料）而不同（IFLA, 2018）；數據可以以多種形式存在，例如它是印刷品上的文字或數位資源裡的內容、它是一個人腦海中的事實、或者它是存儲在電子記憶體中的位元和位元組（Du & Khan, 2020）。

數據本身不帶有任何意義，要使數據成為資訊，就必須對其進行解釋並賦予其意義。Ackoff 於 1989 年提出的數據—資訊—知識—智慧（DIKW）層次結構，說明數據、資訊、知識與智慧這四者的關聯，亦即數據轉化為資

訊，資訊轉化為知識，知識轉化為智慧（Du & Khan, 2020）；與圖 1 相仿，也有學者認為數據存在於層次結構的最底層，是原始狀態的數據，它的存在支援著資訊本身的創造（Weiss, 2018）。

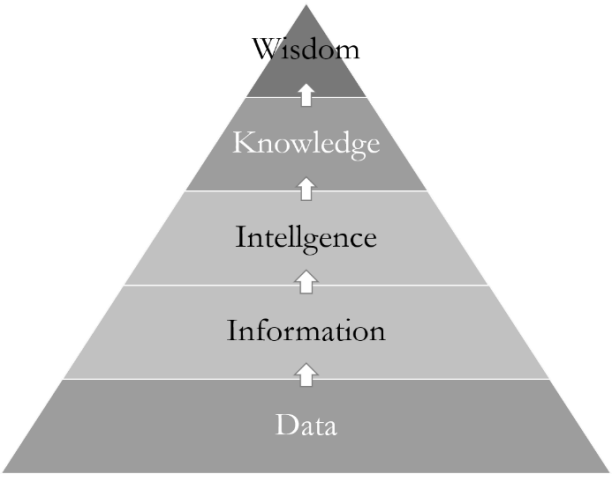

圖 1 資訊層次結構樣本（類似馬斯洛的層次結構），以數據為最原始的資訊形式，並逐級上升到智慧。取自 "What is Data?," Weiss，2018，*Big Data Shocks，頁 9*。

鄭江宇（2019）解釋，「數據」是指未經過處理的原始紀錄，人類是重要的數據生產者，時時刻刻從生活中產出新的數據，而這些數據會經過一連串的演進：處理數據轉化為資訊、資訊進行比較與歸納後轉為知識，再經判斷與分析做出的決策成為富有價值的智慧。隨著資訊和通信技術的發展，我們已經進入了大數據時代，大多數智慧設備、感測器和連網機器都在自我生成數據，數據的產生速度前所未有，而大數據的概念正可用來描述在我們日常生活中呈指數級增長的數據（Du & Khan, 2020；Zhan & Widen，2017）。

一、大數據的定義與價值

前一節提到「大數據」一詞在 2012 年 Doug Laney 給了它一個比較完整的定義之後，其內涵與界限才開始明朗。Laney 指出「大數據是指具有數量大、速度快和種類多等特徵的資訊資產，需要成本效益高、創新形式的資訊處理，以增強洞察力和決策能力。」（Beyer & Laney, 2012）。大數據為近年興起的熱門概念，由於關注點不同，不同研究者有不同的定義，迄今「大

數據」尚無標準的定義。

Zhan 與 Widén（2017）自 39 篇文獻中所採用的 35 個大數據定義，將歸類為數據導向定義與能力導向定義。數據導向的定義，大數據被視為具有某些特徵（例如數量大、快速增長）的數據；能力導向的大數據則被定義為處理數據的技術，係指用於處理大量數據的工具、流程、技術或思想。

Laney 為大數據所給的定義，即為數據導向的定義，其他還包括被引用次數較多的「大數據是巨量、成長速度快、多樣的資訊資產，需要經濟高效的創新資訊處理形式，以增強洞察力和決策能力」、「大數據如同數據集， 其大小超出了典型數據庫軟體工具可獲取、存儲、管理和分析能力」；「大數據一詞主要用於描述數據的數量，多樣和速度，此類數據通常包含大量的半結構化和非結構化數據格式，這些格式很難用傳統的數據庫技術進行存儲，處理和分析。」、「全球資訊網和電子系統隨著新科技迅速發展，正在產生大量的結構化和非結構化數據，這種數據生成的數量之大和速度之快， 以至於傳統的數據庫和資訊系統技術無法適當地管理和處理。」等（Zhan &Widen，2017）。數據導向的定義，將大數據稱為具有數量（volume）、速度（velocity）、多樣（variety）特徵的數據，然而，大數據不僅涵括了巨量的數據，對於數據的分析以及解讀並取得當中之線索、趨勢、商機以及戰略價值，才是最核心的概念（羅濟威，2015），因此，價值（value）及準確性（veracity）二項特徵亦被定義於其中。

「大數據是一個術語，用於描述使用一系列技術（包括但不限於 NoSQL，MapReduce 和機器學習）對大型和/或複雜數據集（dataset）進行存儲和分析。」、「大數據是新一代技術和體系結構，旨在透過大量的各種數據，實現高速獲取、發現和分析，並從中提取價值。」、「應用程式中的數據集和分析技術非常大（從 TB 到 EB）而複雜（從傳感器到社群媒體數據），因此需要高級和獨特的數據存儲、管理、分析和視覺化技術。」等，是能力導向的大數據定義。從能力導向的定義中，大數據與其說是數據的數量，不如

說是關於數據的思考，數據的處理以及透過數據的眼光來應對挑戰和機遇，即便我們擁有最先進技術，缺少數據思維，數據還是數據，沒任何價值；大數據不僅是技術，更是思維方式、發展戰略和運營模式(江雲，2015) 。Hilbert（2016）認為「大數據的全稱其實應該是大數據分析」，人們需要從「具體的peta-、exa-或 zettabytes 規模」中獨立考慮數據，而更加關注數據導向進行決策的過程；他斷言，大數據最終的目標是「為智慧決策而進行分析」（Weiss, 2018）。

Weiss（2018）、Du 與 Khan（2020）透過數據的 5V 特徵來描述、定義或識別大數據的標準：

（一）數量（volume）：生成和存儲數據的數量；可觀的數據量、存儲的數據量；數據量是指科學和教育、商業和人際交往記錄所產生的大量數據集；數據量在存儲和處理中起著重要作用；大量的數據為分析提供了重要的資訊來源。

（二）速度（velocity）：數據創建和處理的速度；數據在時間上的創建頻率（可能是即時的、交互的，也可能是批次的）、存儲資訊的速度、數據集演變的速度；數據創建速度既能提供洞察力，也可能阻礙準確分析。

（三）多樣性（variety）：指不同類型的可用數據、數據的多樣性、數據的來源多種多樣；數據多樣性也指數據的組織程度，非結構化數據缺乏足夠的組織程度，而結構化數據具有較高的組織程度；簡單來說，就是數據類型的多樣性。瞭解數據的多樣性，將決定如何將其應用於分析。

（四）真實性（veracity）：數據真實性主要是指數據的品質和準確性，定義了當需要依據數據做出重要決策時，數據的可信度有多高。在一個數據集內，品質可能會有所不同，亦可能影響分析並妨礙其有用性。

（五）價值（value）：數據價值是指其在決策中的有用性。

IFLA（2018）總結了大數據的 5V 特徵如圖 1，然而大數據的定義還在不斷

發展，這從大數據特徵中不斷增加的V（如 variability、validity）數量，即可看出。

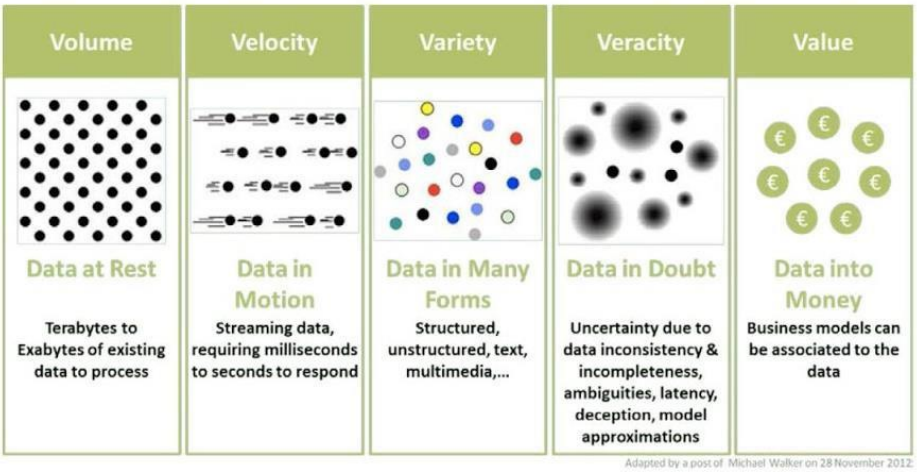

圖 1　大數據的特徵。取自 "A concept data science framework for libraries," IFLA Big Data Special Group, 2018. 檢自 https://www.ifla.org/files/assets/big-data/publications/a_concept_data_science_framework_for_libraries.pdf

學術界將大數據視為數據導向科學研究的第四個範式，簡稱為數據科學（data science）（Du & Khan, 2020）數據導向的研究受到重視，研究人員面臨大量數據創造、處理、利用、分析和提供他人重複使用的新問題（Gray & Szalay, 2007）；隨著數據規模的增加，數據科學有助於從數據中提取新的見解（Gupta & Rani, 2018）。Du 和 Khan（2000）指出數據科學的核心關於從數據中產生有價值的洞察力，以便做出明智的決策；數據科學家（data scientist）必須能夠確定相關的問題，從各種數據來源收集數據，全面組織資訊，並將其結果轉化為可行的解決方案，以及有效地傳達他們的發現。

「數據就像原油一樣，它很有價值，但如果未經提煉，它就不能真正被使用」（Palmer, 2006），大數據的價值在於發現、洞察力，然而大數據的增長速度超過了數據存儲，分析，管理等方面的技術進步，從而給研究人員帶來了巨大挑戰（Gupta & Rani, 2018）。隨著資訊和通信技術（ICT）的發展，我們已經進入了大數據時代，大多數智慧設備、感測器和聯網機器都在自我生成數據，這些可供解讀並作為資訊呈現的大量數據，需要數據科學家對數據進行清理，選擇合適的軟體進行分析，並呈現結果（Du & Khan, 2020）。

二、大數據的基本框架

圖 2 係筆者以孟小峰、慈祥（2013）及楊帆、張紅與薛堯予（2017）所建立的大數據處理基本流程與大數據平臺架構，期能揭示數據自獲取與存儲、組織與管理、利用與分析，至提供決策的過程；並以「大數據的管理」 及「大數據分析與數據導向的決策」說明大數據管理全生命週期的運作框架。

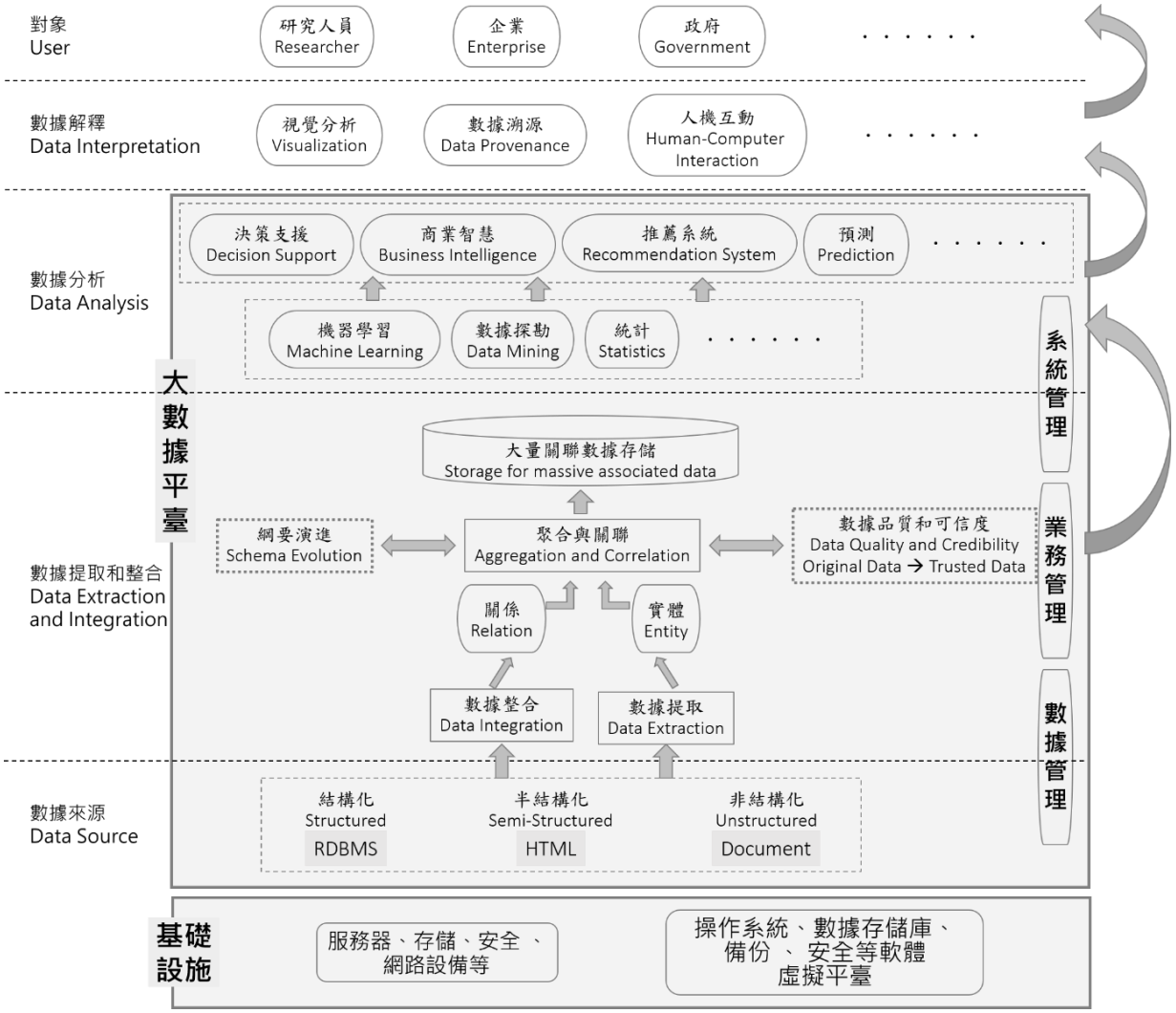

圖 2 大數據運作及平臺框架。改編自「大數據管理：概念、技術與挑戰」，孟小峰、慈祥，2013，*計算機研究與發展*，*50*(1)，頁 151；「基於核心業務系統的圖書館大數據平臺構建策略研究」 楊帆、張紅與薛堯予，2017，*圖書館學研究*，頁 42。

（一）大數據的管理

數據的管理重點主要在數據的生命週期中成功保存和管理數據以備使用和再利用的各個階段。圖 3 揭示了數據自創建或接收（create or receive）、評估與選擇（appraise & select）、攝取（ingest）、保存行動（preservation action）、存儲（store）、檢索，利用與重複使用（access, use & reuse），至轉置（transform）的全生命週期（full lifecycle actions）。

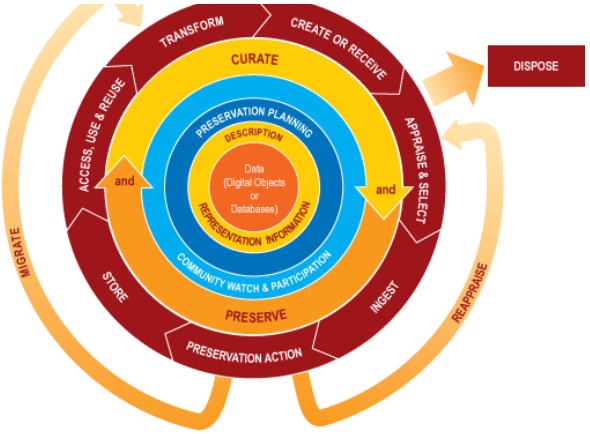

圖 3 數據生命週期，成功管理，整理和保存數據所需的階段。取自 "The DCC curation lifecycle model," Digital Curation Centre, 2008.檢自 https://www.dcc.ac.uk/sites/default/files/documents/publications/DCCLifecycle.pdf

創建或接收（create or receive）實質上是根據收集政策從數據創建者、資料庫、存儲庫或數據中心接收數據，同時為接收到的數據創建適當的詮釋資料；評估與選擇（appraise & select）是指評估數據的長期度用和保存；攝取（ingest）是指將數據轉移到新的資料庫、存儲庫或數據中心；保存行動（preservation action）則為數據進行清理，以確保數據保持真實、可靠和可用的品質；存儲庫（store）是保護數據的方式，亦為數據共用（分享）的工具；檢索，利用與重複使用（access, use & reuse）確保數據在日常的基礎上可以被使用者取用；最後一個順序動作是轉置（transform），

它指的將檔案轉換為可長期保存的格式（Du & Khan, 2020）。

數據的管理，是為了提供有意義的和持久的數據取用，而採取的包含工作和行動（Johnston etc., 2018），是對數據的整個生命週期進行管理以實現長期可用性和再使用性的過程（Lee & Stvilla, 2017）。從廣義上講數據管理可以是為當前和未來的使用者在整個生命週期中對數據進行積極和持續的管理（Yoon & Donaldson，2019）以提高數據的可訪問性、用性和可發現性。Du 與 Khan（2020）指出大數據環境下，巨量的數據、數據產生的速度與數據格式的多樣化等，讓數據管理工作面臨許多挑戰，也帶來相關的問題，包括需要採集和存儲不同形式的數據，這些數據往往是來自多個來源的、即時的，和連續格式化的數據，如視頻、關聯式數據庫、推文、金融轉行動、檔案格式的學術紀錄等；我們收集數據的能力迅速提高，但在過濾、支援和管理數據的能力卻沒有同樣匹配，例如基礎設施無法跟上，沒有足夠的結構來管理數據，沒有足夠的有用或有意義的應用；此外巨量的非結構化數據在攝取與保存、大數據存儲庫的建置、維護等所費費用，以及大數據的安全管理等等，都較傳統的數據管理來得複雜且具挑戰性。

（二）大數據分析與數據導向的決策

良好的數據管理本身並不是目標，而是導致知識發現和創新，以及後續數據和知識在數據發布後被社會整合和再利用（Wilkinson et al., 2016）。擁有巨量的數據，並不一定意味著數據價值增加，孟小峰與慈祥（2013）指出，數據分析是整個大數據處理流程的核心，因為大數據的價值產生於從大量資料中識別有效的、新穎的及潛在有用的資訊的分析過程。而大數據挑戰是對數據進行大規模分析，並為未來的行動提取廣泛的資訊或知識，這比以往任何時候都更加困難（Gupta & Rani, 2018）。

傳統意義上的數據分析，主要是以數據庫裡存儲的結構化數據為分析標的，透過線上分析處理（online analytical processing），從數據中多維

分析、提取出更深層次的需求，進而促使了數據挖掘技術的產生，並發明了類聚、關聯分析等已形成一套行之有效的分析體系（孟小峰、慈祥，2013）。傳統數據分析方法，大多數都是透過對原始數據集進行抽樣或者過濾，然後對數據樣本進行分析，尋找特徵和規律，其最大的特點是透過複雜的演算法從有限的樣本空間中獲取盡可能多的資訊。然而隨著計算能力和存儲能力的提升，大數據分析的分析標的是全體數據，而不是數據樣本，其最大的特點在於不追求演算法的複雜性和精確性，而追求可以高效地對整個數據集的分析（張引、陳敏、廖小飛，2013）。為處理巨量的半結構化或非結構化數據、即時性數據，大數據分析技術和方法不斷升級，且正在影響、融合並試圖革新傳統的決策支援、知識發現、資訊管理、知識管理理論與方法；數據環境和智慧環境下，數據挖掘、文本分析、社會網路分析、情感分析等方法，人工智慧、物聯網、雲端運算、機器學習、深度學習等技術也呈現融合應用的態勢（周耀林、柴昊、與趙躍，2019）。

數據分析的目的主要是為從不同的資訊源中提取見解，為組織提供寶貴的洞察力分析報告和建議，如何傳達分析結果、如何向組織解釋發現的內容是一個挑戰，而數據視覺化提供了一種機制來解決這一難題。數據視覺化在數據分析過程中起著至關重要的作用，有助於揭示有用的見解（Gupta & Rani, 2018）。視覺化工具將繁雜的數據簡化成有效的圖表，讓者可從易於理解的資訊中發現模式（Du & Khan, 2020）。大多數視覺化工具以處理數位數據為主，另一種經常需要進行視覺化分析的數據是文本，文本視覺化則從文檔中提取意義，對文檔進行分組或比較，或尋求兩組文檔之間的相關性。除引入視覺化技術解釋數據分析結果外，張引、陳敏與廖小飛（2013）也建議讓需求者在一定程度上參與具體的分析過程，藉由人機互動技術引導需求者，讓需求者在獲取分析結果的同時能更加理解分析結果的由來。

囿於篇幅，本文未能巨細靡遺地指出大數據基本框架中各個重要的細部內容，茲以 Gupta 與 Rani（2018）為大數據的來源、特徵、存儲類型、

分析階段、應用軟體、技術等分類為基礎，並整理圖書館大數據專書與相關文獻內容製表如表 1，期讓讀者能一覽未來進一步認識大數據內涵的議題或範圍。

表1
大數據分類

來源(Sources)			
○社群媒體(Social media)		○物聯網(Internet of Things)	
○Web 2.0時代的產物 (如Wiki、部落格)		○交易數據(Transactional data)	
○軟體紀錄(Software logs)		○企業軟體(Enterprise applications)	
○科學及研究(Science and research)		○健康照護(Healthcare)	
○天文(Astronomy)		○經濟活動數據(Economic data)	
○基因學(Genomics)		○零售業與製造業(Retail and manufacturing)	
○通訊(Telecommunication)		○公部門(Public sector)	
特徵 (Characteristics)			
○巨量(Volume)	○速度(Velocity)	○多樣(Variety)	○變異性(Variability)
○真實(Veracity)	○價值(Value)	○有效性(Validity)	
存儲類型 (Storage types)			
○寬欄存儲(Wide-column store)		○文檔存儲(Document store)	
○鍵值存儲(Key-value store)		○圖表存儲(Graph store)	
分析階段 (Analytics Phases)			
○獲取(Acquistion)	○提取(Extraction)	○清洗(Cleaning)	○過濾(Filtering)
○驗證(Validation)	○集成(Intergrate)	○分析(Analysis)	○視覺化(Visualization)
分析型態 (Analytics types)			
○描述性(Descriptive)		○診斷性(Diagnostic)	
○預測性(Predictive)		○描述性(Prescriptive)	
軟體工具 (Open-source tools)			
○Apache Hadoop	○Apache Hadoop Ecosystem		○Apache Spark
○Apache Storm	○Apache Samza	○Apache Flink	○Apache Apex
技術 (Techniques)			
○機器學習(Machine learning)	○數據探勘(Data mining)		○統計學(Statistics)
○深度學習(Deep learning)	○眾包(Crowdsourcing)		○優化(Optimization)
○自然語言處理(Natural language processing)		○信號處理(Signal processing)	
○優化方法(Optimization methods)		○視覺化(Visualization)	
○人工智慧(Artificial intelligence)		○認知計算(Cognitive computing)	

參、圖書館與大數據

一、大數據環境下的圖書館

作為傳統知識資訊服務中心，圖書館在文件存儲和數據管理中發揮著重要作用，因此圖書館不可避免地會受到大數據的影響（Li, Lu, Dou and Wang, 2017）；「大數據」正在重塑圖書館於執行任務時所擁有和使用的模式（Affelt，2015）。圖書館受到網路科技發展、使用者行為改變等影響，面臨許多發展困境，在大數據時代，很多圖書館員認為圖書館可以利用大數據幫助圖書館擺脫目前的困境，因為他們認為大數據是圖書館難得的資產（Hashem etc., 2015; Kim & Choi, 2016），圖書館可以利用大數據處理技術對大數據進行提取和分析，挖掘讀者的潛在需求，然後採取相應的措施吸引和留住讀者（Provost & Fawcett, 2013; Narendra, 2016）。

楊海燕（2012）指出圖書館蒐藏了種類繁多、格式不一，來自購置或自製的數據資源，數據量日益龐大且每天都在迅速增長，並且隨著多元化的資訊服務累積了可觀的、每時每刻都在遞增的讀者數據，以及尚未進入大數據系統的圖書館書目紀錄、讀者紀錄及其他日常業務統計數據等，圖書館已具有大數據特徵。美國圖書館協會（American Library Association, ALA）於 2013 年發表了一份大數據報告，「您的圖書館可能正在收集大數據進行分析，以幫助做出數據導向的決策。您可以使用哪些類型的大數據來對館藏發展、改造公共空間，或透過學習管理系統跟蹤圖書館資料的使用情況等，做出更好的決策？或者您可以成為您所在機構大數據館藏的思想領袖，為存儲和提供可取用的大數據集提供指導。現在是你的圖書館瞭解大數據為研究人員、管理部門和你所在機構的圖書館員帶來哪些問題和機會的機會」（Bieraugel, 2013; Weiss, 2018），建議圖書館建立並發展數據導向的決策管理策略。美國學與研究圖書館協會（Association of College & Research Libraries, ACRL）2015 年環境掃描報告也指出了對更先進、更複雜的數據度用服務的要求，

並敦促圖書館員對此一領域研究與實踐有深入的認識和理解，以便能夠幫助專業研究人員進行數據共用、管理和保存（Du & Khan, 2020）。

以數據為導向的營運策略，能在關注圖書館裡每一個具體的結構化資 訊資源需求的同時，也可使非結構化數據分析變得可行和經濟高效，從而實 現知識橫向擴展以滿足急劇擴張的知識服務需求。作為一個新的尚未開發 的資訊源，非結構化數據分析可揭露之前很難或無法確定的重要相互關係。 圖書館以數據為導向的營運策略，旨在獲得更加豐富、深入和更加準確的使 用者、知識運營者以及知識服務洞察，並最終提高圖書館的核心競爭力，與 以往相比，大數據應用可更加快速地做出時間敏感的決策、監控最新知識服 務趨勢、快速調整方向並抓住新的知識服務機遇（樊偉紅、李晨暉、 張興旺、秦曉珠、郭自寬，2012）。建立圖書館大數據應用與服務框架，可以從 整體業務與數據的宏觀角度掌握圖書館的業務水準，實現業務監測、決策分 析、統籌規劃。從狹義角度看，圖書館大數據可「促進業務融合，增強數據 的統一管理」、「實現海量數據的有效存儲和系統數據的高效利用」、「提 高服務水準，增強用戶黏性」、「統一規劃資源部署，提供輔助決策支援」，有助於圖書館提高資源利用能力、資源利用效率與數據品質，驅動業務創新、快速準確做出決策，可以更深入地瞭解用戶並準確定位，提高服務水準（楊 帆、張紅、薛堯予，2017）。

二、圖書館的大數據

圖書館在大數據中扮演著「大數據的使用者或受益者」、「大數據的提供者或開發者」、「大數據的營運者或維護者」三個角色（樊偉紅、李晨暉、張興旺、秦曉珠、郭自寬，2012），圖書館不僅管理大數據資源，也創造大數據；不僅保存著大數據，也提供大數據服務。數據資源在圖書館廣泛被運用，在參考服務被用來回答讀者所提出的問題；來自流通服務的數據提供了用於規劃資訊資源採購、衡量主題興趣和量化圖書館使用量的指標。技術服務部門的館員將文字轉化為數據，提供讀者快速查找和檢索所需資料；資訊

系統人員在管理資訊系統時需要處理大量的數據—從圖書館網站或內網的點擊量，包括不同類型內容的點擊量的區分，到每天不同時間的個人登錄次數，都屬於他們的工作範圍；行政管理者將數據用於從預算編制和策略規劃到目標設定和員工績效評估等各個方面（Affelt, 2015）。過去圖書館員所處理的數據，大部分是儲存在關聯式數據庫中，而且主要是傳統格式。例如， 在參考服務查找歷史事實或事件，在流通服務和資訊服務，我們追蹤讀者過去的使用情況；技術服務人員根據以前確定的編碼做出分類決定，行政管理者則根據歷史數據來規劃未來（Affelt, 2015）。

以下提供圖書館大數據應用案例，讀者可概觀圖書館的大數據資源及大數據資源的應用方式。

（一）數位館藏的應用

美國國會圖書館（Library of Congress）於 2012 年開啟了圖書館大數據應用計畫，包括「Digging into Data」計畫，將 500 萬頁的報紙數位化，並將影像 OCR 為文本，進行數據挖掘；分析網站典藏庫中的 50 億份不同類型的檔案；向 Twitter 提出研究請求取得超過 500 億條推文，進行語言分析、分析新聞報導的地理傳播等（Showers, 2015）。2014 年，大英圖書館（British Library）宣布與倫敦大學學院電腦科學系和數位人文科學系（ University College London’s Computer Science and Digital Humanities departments）合作，開放大英圖書館的數位館藏作為大數據實驗的一部分，讓藝術和人文科學研究人員受益（Showers, 2015）。

（二）讀者使用資源及服務紀錄的應用

Wollongong 大學圖書館（University of Wollongong Library）2009 年與學校的績效指標部門（Performance Indicators Unit）合作，建立「Library Cube」數據庫，將圖書館的使用數據與學生的人口統計和學業成績數據建立關聯，以研究學生使用圖書館資源與成績之間的關係。圖書館的兩個數據源包括學生的館藏資源借閱紀錄和電子資源使用紀錄，電子資源使用

紀錄來自學生使用 EZproxy 的日誌，包括透過圖書館目錄或學校的學習管理系統提供的連結取用的資料庫、電子書和電子期刊等。同年，Huddersfield 大學（University of Huddersfield）以圖書借閱紀錄（使用圖書館管理系統的數據）、電子資源利用紀錄（電子資源系統的點擊）、進館紀錄（利用門禁系統的統計數據）調查不同族群對圖書館資源的非/低使用率情形（Showers, 2015）。

明尼蘇達大學圖書館（University of Minnesota Libraries）2011 年與大學的機構研究辦公室合作，蒐集相關數據衡量學生使用圖書館服務和資源的頻率和方式，並確定圖書館服務資源的使用對學生的學業成功有什麼樣的影響。研究團隊研究了如何將圖書館服務的使用情況與個人使用者帳戶建立關聯，同時保留讀者的隱私，以確定誰使用和誰不使用圖書館。研究使用了 5 類、13 個不同的圖書館數據，包含資訊檢索紀錄（資料庫、電子書、電子期刊、網站的登入紀錄）、流通紀錄（資料借閱及館際互借紀錄）、共用電腦使用紀錄（使用者必須透過名為 CybraryNTM 的共用電腦管理服務系統登錄，登錄數據包括互聯網 ID，每學期末從 CybraryNTM 資料庫中提取。）、圖書館利用指導紀錄（參加講習會、圖書館融入學科利用指導課程或圖書館研究工作坊之紀錄），以及利用參考服務之紀錄（Showers, 2015）。

Huddersfield 大學（University of Huddersfield）2010 年底與布拉德福大學（University of Bradford）、德蒙福特大學（De Montfort University）、埃克塞特大學（University of Exeter）、林肯大學（University of Lincoln）、利物浦約翰摩爾斯大學（Liverpool John Moores University）、索爾福德大學和提賽德大學（University of Salford and Teesside University）等合作，以「圖館影響數據專案計畫」（Library Impact Data Project, LIDP）取得了Jisc資訊環境計畫（Jisc Information Environment Programme 2009-11）的補助。計畫初期利用 Huddersfield 非/低使用率研究的原始框架成功地證明了圖書借閱和電子資源的使用與最終的學位成績之間存在著正相關關係，「因

此，圖書或電子資源的利用率越高；學生獲得更高一級學位成績的可能性就越大」，計畫的第二階段加入了人口統計學、學科和其他數據來源研究了學科與圖書館使用情況之間的關係。此計畫為一成功的合作個案，在過程中，不同的機構可以在很短的時間內檢索和共用數據（Showers, 2015）。

此外，哈佛法學院圖書館（Harvard Law School Library）的哈佛圖書館創新實驗室（Harvard Library Innovation Lab）開發了一個名為 Haystacks的圖書館分析工具。透過逾 1,200 萬筆哈佛大學圖書館（Harvard University Libraries）空間利用、參考文獻互動、採購、流通和電子資源的使用數據；Haystacks 重點是將館藏數據視覺化，館藏管理館員能夠瞭解哈佛圖書館館藏的使用情況及其如何隨著時間的推移而變化；Haystacks 能對圖書館的行動提供見解，使圖書館工作人員在日常工作中做出更明智的決策（Showers, 2015）。

（三）館藏書目資料的應用

英國研究圖書館（Research Libraries UK, RLUK）於 2011 年發起，和里茲大學（Universities of Leeds）、雪菲爾大學（Universities of Sheffield）及約克大學（Universities of York）與 Copac 團隊合作開發 Copac 館藏管理工具計畫（Copac Collection Management Tools Project），該計畫整合了各圖書館的館藏書目資料、館藏典藏及狀態資料及使用紀錄等，旨在提供館藏管理支援服務，為圖書館工作人員提供工具（簡稱CCM 工具），使他們能夠圍繞與館藏有關的數據管理活動做出更明智的決定，如資料處置、保存或數位化、館藏評估和發展等（Showers, 2015）。

CCM 工具應用的範圍廣泛，包含書庫的管理，圖書館可將自己館的館藏與其他圖書館的館藏進行比較，並確定那些在全國範圍內罕見或獨 特的書目，CCM 工具提供了一個基準，圖書館可以根據這個基準作出圖書撤架或館藏淘汰的決定。CCM 工具亦提供館藏分析功能，圖書館可評 估自己館的館藏在數據存儲庫中是否具有區域或國家等之代表性之意義，

或可評估某一學科主題或研究領域的館藏強弱（Showers, 2015）。

（四）讀者行為軌跡數據的應用

Open 大學圖書館（Open University Library）的「RISE 推薦改善體驗」（Recommendations Improve the Search Experience）計畫，研究用戶對推薦價值的看法，並使用 Google Analytics 來跟蹤工具和推薦的使用情況，以驗證「推薦系統可以提高學生在新一代電子資源發現服務中的體驗」。該計畫利用了 EZproxy 日誌檔中的數據，內容包括日期/時間戳記資料、會話資料、推薦人和請求數據，以及利用公用電腦使用者用戶 ID；在MySQL 中設計一個數據庫來存儲日誌檔數據，並研究如何解析日誌檔；透過用戶 ID，確定用戶的類型以及學生選修的課程；日誌檔中的請求，被用來作為獲取更完整書目數據的依據；日誌檔數據還提 供了會話 ID 和日期/時間戳記，這些數據有助於形成其他類型的建議，如「看了資源 A的人也看了資源 B」；使用 EBSCO Discovery API 設計了一個簡單的使用者介面，作為向使用者展示推薦資訊的基礎，建立了一個以 Google Gadget為基礎的檢索系統，向使用者提供 (1)對課程的建議—我的課程的用戶訪問了這些資源，(2)對查詢的建議—使用這個或類似檢索詞的用戶訪問了這些資源，(3)對擴展檢索的建議—這些資源可能與你最近瀏覽過的其他資源有關（Showers, 2015）。

從上述四類應用案例可以得知圖書館大數據的幾個來源，包括圖書館的數位館藏、館藏書目紀錄、館藏目錄使用紀錄、讀者使用圖書館各項資源及服務的紀錄等。在大數據環境下，圖書館還有許多數據尚未被有制度地蒐集整理與分析應用，例如 (1)RFID 數據：RFID 嵌入到圖書館欄資源中，資源被利用情形的跟蹤及分析，將會是大數據的主要來源之一； (2)感測器數據：透過分布在圖書館不同位置或環境中的感測器對所處環境和資源進行的感知，不斷生成的數據，由於長時間積累所產生的數據量也非常巨大；(3)社群媒體用戶的互動數據：社群媒體所產生的數據量遠

遠超過以往任何一個資訊傳播媒介，毫無疑問，它將會成為未來很長一段時間內，大數據最為主要的來源之一；(4)行動數據：圖書館用戶透過行動載具利用服務及資源的行為和需求等資訊，如能進行即時分析亦能協助圖書館開展有效的智慧輔助決策（樊偉紅、李晨暉、 張興旺、秦曉珠、郭自寬，2012）。

圖書館大數據應用的案例，還展現在數據整理與分析工具的開發上。賓夕法尼亞大學（University of Pennsylvania）開發了 Metridoc，它被稱為「一個可擴展的框架，支援圖書館評估和分析，使用從異構來源收集的各種活動數據」。該服務的架構意味著它能夠從整個校園的不同系統中收集數據，並將其匯總到一個大型的數據庫中。當新系統在本地實施時，它也可以進行擴展以採集數據，從而確保它不受機構當前系統和流程的束縛。Metridoc 旨在實現對整個大學（從管理系統到圖書館）的數據進行類似於大數據的分析，其架構是專門為發揮全校所有那些小型、離散數據集的潛力而設計的（Showers, 2015）。

另外，由 Huddersfield 大學（University of Huddersfield）領導的圖書館影響數據專案計畫（LIDP）既關注圖書館的影響和價值，也關注個性化推薦服務的專案，特別的是也滿足了各校對共用數據分析服務的需求。第二階段（2013 年 1 月到 2015 年 5 月）推展的圖書館分析和測量計畫（Library Analytics and Metrics Project, LAMP）為英國學術圖書館開發一個圖書館共用的分析服務模型，提供一個數據儀錶板，使圖書館能夠利用他們在日常服務及工作中所獲取的多種類型數據，去支援改進和發展新的服務，並以新的方式展示整個機構的價值和影響。LAMP 計畫的用意包含分析（如何將數據有意義地展示給用戶）、社群（扮演圖書館社群開發、塑造和實施 LAMP 方面的角色）以及數據（攝取和分析不同機構的數據集）三大部分，它整合了 LAMP 合作大學與另六個機構的數據，也使用API 傳送數據至各機構的儀錶板；LAMP 是一個全國性的數據集，這有助於開發基準和績效衡量等應用（Showers, 2015）。

肆、圖書館與大數據基礎研究現況

本節就圖書資訊領域大數據研究現況的相關文獻進行探討，說明圖書館與大數據的研究開端、研究方法、研究主題及熱點等研究結果。

一、圖書館與大數據研究的開端

2012 年，*Library Journal* 刊載了第一篇圖書館大數據文獻〈How do libraries need “big data”?〉、《圖書與情報》也於同年刊登了第一篇圖書館大數據文獻〈大數據時代的圖書館服務淺析〉，之後圖資領域大數據相關文獻每年都在增加；圖書館大數據文獻的作者主要來自中國和美國（王倩、李天柱、劉小琴，2017；Ann & Mannan, 2020）。促使中國與美國有較多的研究產出，或許與這二個國家將大數據列入國家發展的策略中有關。2012 年歐巴馬政府公布「大數據的研究和發展計畫」（Big Data Research and Development Initiative），使得國外大數據的研究正式步入軌道（賈玉文、黃小淋、王康，2019），也引起世界各國對大數據的重視；中國則於 2017 年 1 月，頒布了《大數據產業發展規劃（2016-2020 年）》，將大數據產業作為關乎國家核心競爭力的戰略制高點（賈玉文、黃小淋、王康，2019），同時中國教育部門將大數據研究列入人文社會科學重點研究基地重大專案—「大數據資源規劃與統籌發展研究」(周耀林、柴昊、與趙躍，2019），亦加速了大數據的應用及學術研究。

二、圖書館與大數據研究現況

筆者確定以「圖書館與大數據」為研究方向後，以關鍵詞包含「圖書館」（library）及「大數據」（或「巨量資料」）（big data）查詢中文及西文圖書及期刊文獻，並篩選出圖書資訊學領域的基礎文獻與現況研究文獻，期瞭解圖書資訊領域在圖書館大數據的研究概況。基礎文獻與研究主要以文獻分析探究大數據的基本概念、大數據環境下的圖書館等，並以主題詞、關鍵詞分析圖書資訊領域的研究熱點。

有關圖書館與大數據的基礎文獻及研究，在臺灣亟待發展。2020 年 11 月以此二關鍵詞在「臺灣博碩士論文加值系統」僅查得 1 篇於 2015 年完成的「圖書借閱行動服務模式之建立研究」碩士論文，「臺灣期刊論文索引系統」中亦僅有 4 篇，但都非屬基礎文獻或研究；有關研究現況分析的文獻，中國產出最為豐富，不論是針對其國內或國際發表之文獻，自2014 年起即有學者開始相關的討論，表 2 為本次篩選以蒐集文獻來源一致、分析方法相近，並以圖資領域為範圍具代表性（或核心期刊）之現況研究文獻；其中，僅陳軍營、白如江、王效岳、劉自強（2018）「中外圖情領域大數據近十年（2007-2016）研究現狀與發展趨勢」一文分析之範圍兼含中外。

表 2

圖書館與大數據研究現況文獻分析範圍及方法

作者	分析文獻範圍及規模	分析方法
王春華、李維、文庭孝（2015）	檢索日期：2014.12.19 使用資料庫：中國知網（CNKI） 文獻分類—「圖書情報與數位圖書館」 查詢條件：2008-2014 發表文獻，主題中有「大數據」 查得結果：479 篇	運用詞頻統計和共詞分析方法，借助 SPSS 和 UCINET 軟體進行聚類分析、戰略座標圖分析和核心－邊緣結構分析。
司莉、王雨娃（2018）	檢索日期：2017.6 使用資料庫：中國知網（CNKI） 文獻分類—「圖書情報與數位圖書館」 查詢條件：2011-2017 發表之文獻，篇名中有「大數據」 查得結果：1,693 篇	採用 Citespace 軟體，就文獻的發表時間、發表機構、作者等進行統計分析。
陳軍營、白如江、王效岳、劉自強（2018）	檢索日期：2017.7.15 使用資料庫：中國知網（CNKI）與 Web of Science 核心合集 查詢條件：中文以主題詞為「大數據」或「 big data」檢索 2007-2016 年圖資領域的 18 種核心期刊；西文以主題詞（topic）為「Big Data」檢索 2007-2016 年圖資核心期刊。 查得結果：中文 692 篇、西文 90	利用數據分析軟體 KNIME 挖掘研究熱點主題進行識別分析，然後以視覺化的方法對熱點主題在時間維度上進行演化趨勢呈現，最後進行國內外熱點趨勢對比分析。

表 2

圖書館與大數據研究現況文獻分析範圍及方法

作者	分析文獻範圍及規模	分析方法
	篇	
趙棟祥，張瑞（2018）	檢索日期：2017.12.3 使用資料庫：Web of Science 核心合集 查詢條件：至 2017 年，圖書資訊領域期刊主題詞（topic）為「big data」 查得結果：443 篇	(1) 以文獻題錄數據分析工具(SATI 3. 2)對數據集中的關鍵字進行提取，選取頻次在一定閾值的關鍵字作為高頻關鍵字，生成高頻關鍵字清單。 (2) 再次利用 SATI 對高頻關鍵字進行共現分析，為下一步聚類分析生成相應的高頻關鍵字相似矩陣和相異矩陣。 (3) 使用 SPSS 對上一步生成的關鍵字相異矩陣進行系統聚類分析，生成聚類分析樹狀圖。 (4) 根據聚類分析結果，結合相關文獻，對挖掘出的熱點主題做更深入的討論和分析。
周耀林、柴昊、趙躍（2019）	檢索日期：2018.10.30 使用資料庫：Web of Science 核心合集、ProQuest Research Library 的 Library Science Database 查詢條件：Web of Science 核心合集中 2011-2018 發表之圖書資訊學文獻主題詞（topic）中有「big data」者；Library Science Database 所有欄位（不含全文）中出現「big data」者 查得結果：WoS 檢得 604 篇 Library Science Database 檢得 8 篇，合併二者去除重複後，共得到 877 篇	(1) 以文獻題錄數據分析工具 SATI 進行年度發文統計、關鍵字詞頻統計。 (2) 用 SATI 匯出關鍵字共現網路矩陣，供分析使用。 (3) 進一步以 Louvain 演算法，對關鍵詞共現網路進行社區劃分，以揭示學科、地區、作者等之間合作而形成的社區或群組現象。 (4) 利用國外新興的數位化基礎設施平臺 CorTexT 分析研究主題隨時間演化態勢。 (5) 利用突發詞檢測演算法（Burst Detection）分析研究前沿，將題錄數據檔導入 Sci2，獲取視覺化圖譜。
劉文江、胡志丹、賈鳳玲（2019）	檢索日期：2019.3.15 使用資料庫：中國知網（CNKI）中的 26 種 CSSCI 期刊 查詢條件：2010-218 年發表之文獻，主題詞中有「大數據」 查得結果：1,680 篇	採用 Citespace、SPSS 和 Netdraw，結合共詞分析和多元統計分析方法，對文獻的作者、機構以及關鍵字詞頻進行梳理。
虞秋雨、徐躍權（2020）	檢索日期：2019.10.9 使用資料庫：第八版北大《中文核心期刊目錄》中的 18 種圖書情報學科期刊	建立了一種以 g 指數為主要基礎的劃分高頻詞的方法，並利用 Excel 軟體進行數據統計及構建共詞矩陣。同時借助 Spss、Pajek 軟體對

表 2

圖書館與大數據研究現況文獻分析範圍及方法

作者	分析文獻範圍及規模	分析方法
	查詢條件：2014-2019 發表之文獻，主題詞有大數據（不含小數據）者 查得結果：815 篇	陣進行視覺化分析、K-core 分析及聚類分析，研究文獻中各關鍵字間的關係。

主題演化分析的科學計量方法主要有詞頻分析法、共引分析法和共詞分析法，自表2 可知大陸學者大多採用此一方法。趙蓉英與余波（2019）指出，主題詞是一篇文章的高度概括和凝練，也是一篇文章的核心和精髓；關鍵字能體現出文章的主題與核心內容，關鍵字的演化也體現出研究主題的變遷（司莉、王雨娃，2018）。詞頻分析法是利用能夠揭示或表達文獻核心內容的關鍵字或主題詞在某一學科領域的研究文獻中出現的詞頻高低，來確定該學科領域的研究熱點和發展動向的文獻計量方法（馬費成、張勤，2006）；共詞分析法以文獻的關鍵字為分析物件，關注詞與詞之間的關聯，可從更微觀的角度揭示學科主題演化規律（唐果媛、張薇， 2015）。此方法在揭示研究領域的主題分布和演化規律等方面具有明顯優勢，已成為描述學科或領域發展現狀與態勢的重要定量分析方法 （周耀林、柴昊、與趙躍，2019）。此外被引論文頻次的高低反映了文獻學術影響力，其經典程度可以透過被引文獻在一定時間內對研究者的認可和傳遞來體現（趙蓉英、余波，2019）。

三、圖書館與大數據研究主題及熱點

陳軍營、白如江、王效岳與劉自強（2018）採用文本數據挖掘技術識別了 2007 年至 2016 年間中國與國際不同時期圖資領域大數據研究的熱點主題，並初步分析了研究主題隨時間推移的演變情況，發現 2007-2009 年中國發文量為 0，國際則以論文數據為研究對象，側重於開放獲取、科學出版、引文分析等技術層面；2010-2012 年中國已經開始有相關的研究 但研究的主題特徵尚不明顯，而國際的研究文獻持續發表，研究主題除

持續前期論文數據的引文分析外，更關注資訊資源獲取和技術分析，開源數據和自然語言處理技術等；2013-2014 年中國已經掀起研究熱潮，數圖書館、科學數據管理、數據挖掘與圖書館研究等議題獲得研究人員的注意；國際的研究在此時則平穩發展，傳統數據分析技術研究演變為細微性更小的語義分析和大數據分析；2015-2016 年中國研究主題與國際接近，研究角度繼續聚焦在數位圖書館、數據視覺化技術、大數據挖掘等，引文分析中社會網絡研究在此時也有進展，語義分析促進了科學計量分析和研究主題分析的研究。

中國於 2008 年至 2014 年期間研究的核心關鍵字有大數據、圖書館圖書館大數據、數據挖掘、高校圖書館、數據分析、數據處理、圖書館服務、資訊服務、雲端運算等，這些關鍵字都有較高的詞頻，在分屬的各聚類中密度和中心度也相對較高，詞間關係相對緊密，反映了相關研究已趨於穩定和成熟；此階段大數據研究的主題包含智慧圖書館與物聯網、數據挖掘與處理、大數據與企業競爭情報、大數據與高校圖書館、大數據與資訊分析、大數據與公共圖書館、大數據與數位圖書館等（王春華、李維、文庭孝，2015）。2014 至 2019 年高詞頻之關鍵詞則達34個，擠身前十者包含大數據、圖書館、高校圖書館、情報學、數位圖書館、知識服務、智慧圖書館、人工智能、數據素養、資訊服務；聚類分析的結果顯示近五年研究主題著重在智慧圖書館及智慧服務、大數據分析與情報研究、雲端運算與數據挖掘、圖書館的創新及服務、智庫與情報服等（虞秋雨、徐躍權，2020）。

周耀林、柴昊、趙耀（2019）分析 2011-2018 年間發表的文獻，數據管理、社會網絡、人工智能、數據挖掘、數據分析、資訊系統、隱私、大數據分析等主題是國際圖資領域大數據研究中關注的熱點；並已形成「商務智能&大數據分析」、「人工智能&資訊科學」、「社會網絡分析&社群媒體」、「數據倫理&個人隱私」、「資訊技術&資訊管理」、「圖書館&數據素養」等六個規模不一的研究方向；自 2013 以來，國際

圖資領域大數據研究共出現 12 個關注度不一的突發性研究主題，其中數據管理、資訊系統、知識管理、機器學習及物聯網五個主題為當前研究的前沿性主題。

此外，Blummer 與 Kenton（2019）利用「big data」 或 「large datasets and libraries」二個檢索詞查找 1981 年至 2018 年間有關圖書館應用和管理數據的文獻，研究大數據的可用性及其對圖書館的影響。作者綜整了76篇文獻歸納出四個主要議題，分別是「大數據管理」、「數據分析服務」、「大數據概述」和「數據館員的培育」。「大數據管理」探討隱私和數據管理、館員保護隱私的技能、管理大數據的挑戰、大數據管理的需求評估、協作與大數據管理、大數據管理專案、館員的數據管理活動等；「數據分析服務」除了討論為讀者提供數據分析服務，也利用大數據對圖書館服務進行評估，證明圖書館在機關裡的價值，並有數篇以書目紀錄作為數據分析標的，研究館藏的異質性；「大數據概述」主要為討論大數據的定義及其對圖書館的機遇與挑戰；「數據館員的培育」主要是就館員在大數據管理的角色和職責、教育計畫與培訓機會進行探討。Ann 和Mannan（2020）分析 Scopus 資料庫中篇名出現「圖書館大數據」的 73篇文獻，利用 VOX Viewer 查看大數據術語與圖書館之間的關係，並且瞭解圖書館大數據發展的概況。研究發現大數據的應用在許多大學或學術圖書館中都可以找到，但很少在公共圖書館中討論；圖書館大數據研究的成果仍然圍繞著大數據的概念、大數據對圖書館的本質、圖書館管理者對大數據的觀點以及大數據在圖書館的實施等議題。作者並以圖書館應用大數據分析的實例有限之現象，認為圖書館利用大數據仍處於早期階段， 同時，圖書館管理者不瞭解大數據對圖書館的重要性、館員缺乏大數據分析技能等，限制了圖書館應用大數據的發展。

伍、結語

本文簡要地說明了國際圖資領域在大數據環境下，有關「圖書館與大數據」議題被討論及應用的概況，期能發揮拋磚引玉之效引起國內圖資機構及圖資研究者對此議題之關注。同時，因囿於篇幅、時間及能力限制，仍有諸多未及討論或交待之處，留待後續跟進。

大數據的核心跨了數據科學、資訊科學、圖書館學和電腦科學領域，為跨學科領域合作提供了多種機會，也對這些學科的核心使命增添了價值（Du & Khan, 2020）。大數據環境下，圖書館日常的數據管理面臨極大的挑戰，尤其是半結構化或非結構化的數據管理；圖書館所典藏的數位資源、書目紀錄及詮釋資料、讀者服務歷程中各資訊系統及服務平臺所產生的大量數據、社群媒體上讀者的使用軌跡等，都是圖書館可以應用的大數據材料；圖書館應及早建立大數據資產管理及分析制度，就圖書館中各種來源數據逐一盤點並建立數據保存計畫，檢核並建立各數據平臺數據提取及整合功能，組織與整理數據並建立數據管理機制、強化數據的存儲及設施、數據館員的知能等，進而積極應用大數據資源，強化以數據為導向的決策管理；而圖書資訊研究除文獻計量基礎研究外，亦應拓展相關的理論與實證研究，為數據管理及大數據分析實務提供有用之模式，也為臺灣的圖書資訊學發展增添一新的核心價值。

參考文獻

王春華、李維、文庭孝（2015）。我國圖書情報領域大數據研究熱點分析。*圖書情報知識*，*4*，82-89。

王倩、李天柱、劉小琴（2017）。全球大數據研究的歷史演進：1993-2016 年。*情報雜誌論壇*，*7*，33-39。

司莉、王雨娃（2018）。國內圖書情報領域大數據研究演化分析。*新世紀圖書館*，*12*，88-94。

江雲（2015）。再論大數據在我國圖書館的應用與推進。*情報科學*，*33*(10)，99-105。

周耀林、柴昊、趙躍（2019）。國際圖情領域大數據研究現狀與趨勢探析。*圖書館雜誌*，*12*，16-44。

孟小峰、慈祥（2013）。大數據管理：概念、技術與挑戰。*計算機研究與發展*，*50*(1)，146-169。

唐果媛、張薇（2015）。基於共詞分析法的學科主題演化研究進展與分析。*圖書情報工作*，*59*(5)，128-136。

馬費成、張勤（2006）。國內外知識管理研究熱點—基於詞頻的統計分析。*情報學報*，*2*，163-171。

張引、陳敏、廖小飛。（2013）。大數據應用的現狀與展望。*計算機研究與發展*，*50*(Suppl.)，216-223。

陳軍營、白如江、王效岳、劉自強（2018）。中外圖情領域大數據近十年（2007-2016）研究現狀與發展趨勢分析。*情報科學*，*7*，104-110

楊帆、張紅、薛堯予（2017）。基於核心業務系統的圖書館大數據平臺構建策略研究。*圖書館學研究*，*6*，38-42，86。

楊海燕（2012）。大數據時代的圖書館服務淺析。*圖書與情報*，*4*，120-122。

虞秋雨、徐躍權（2020）。近 5 年我國圖書情報領域大數據研究熱點分析。*圖書館學研究*，*8*，10-18。

趙棟祥、張瑞（2018）。國際圖情領域大數據研究熱點挖掘與分析。*圖書館學研究*，*14*，10-19。

趙蓉英、余波（2019）。近三年國際圖書情報學研究熱點比較分析。*情報科學*，*37*(4)，3-9, 170。

劉文江、胡志丹、賈鳳玲（2019）。國內圖書情報領域大數據發展態勢研究。*農業圖書情報*，*31*(12)，56-63。

廿、李晨暉、張興旺、秦曉珠、郭自寬（2012）。圖書館需要怎樣的"大數據"。*圖書館雜誌*，*31*(11)，63-68，77。

鄭江宇（2019）。「數大」便是美—大數據與現代生活的連結。*科學月刊*，*50*(11)，檢自，https://www.scimonth.com.tw/tw/article/show.aspx?num=2199&root=2&page=2

羅濟威（2015）。先進國家巨量數據政策分析—以英美日澳為例。國家實驗研究院科技政策研究與資訊中心。檢自 https://portal.stpi.narl.org.tw/index?p=article&id=4b1141427395c699017395c756941e50

蘇玲、婁策群（2019）。我國情報學和傳播學領域大數據研究探析。*情報科學*，*5*，31-37。

Affelt, A. L. (2015). *The accidental data science: Big data application and opportunities for librarians and information professionals.* Medford, NJ: Information Today, INC.

Anna, N. and Mannan, E. (2020). Big data adoption in academic libraries: a literature review. *Library Hi Tech News*, *37*(4), 1-5. DOI: 10.5339/jist.2018.13

Bieraugel, M. (2013). Keeping up with…big data. Association of College and Research Libraries. Retrieved from http://www.ala.org/acrl/publications/keeping_up_with/big_data

Blummer, B., & Kenton, J. M. (2019). Big data and libraries: identifying themes in the literature. *Internet Reference Services Quarterly*, 23(1-2),15-40. DOI: 10.1080/10875301.2018.1524337

De Mauro, A., Greco, M., & Grimaldi, M. (2016). A formal definition of big data based on its essential features. *Library Review*, *65*(3), 122–135. https://doi.org/10.1108/LR-06-2015-0061

Digital Curation Centre. (2008). The DCC curation lifecycle model. Retrieved from https://www.dcc.ac.uk/sites/default/files/documents/publications/DCCLifecycle.pdf

Du, Y. & Khan, H. R. (2020). *Data science for librarians*. Santa Barbara, California: Libraries Unlimited.

Gray, J. & Szalay, A. (2007). eScience – A transformed scientific method. Retrieved from http://research.microsoft.com/en-us/um/people/gray/talks/NRC-CSTB_eScience.ppt

Gupta, D., & Rani, R. (2018). A study of big data evolution and research challenges. *Journal of Information Science*, *2018*, 1-19. DOI: 10.1177/0165551518789880

Hashem, I. A. T., Yaqoob, I., Anuar, N. B., Mokhtar, S., Gani, A., & Khan, S.U. (2015). The rise of "big data" on cloud computing: review and open research issues. *Information Systems*, *47*, 98-115.

Hilbert, M. (2016). Big data for development: A review of promises and challenges. *Development Policy Review*, *34*(1). 135–174. Retrieved from http://www.martinhilbert.net.

IDC (2012) The digital universe in 2020: Big data, bigger digital shadows, and biggest growth in the far east. Retrieved from https://www.emc-technology.com/collateral/analyst- reports/idc-the-digital-universe-in-2020.pdf

IFLA Big Data Special Interest Group. (2018). A concept data science framework for libraries. Retrieved from https://www.ifla.org/files/assets/big-data/publications/a_concept_data_science_framework_for_libraries.pdf

Johnston, L. R., Carlson, J., Hudson-Vitale, C., Imker, H., Kozlowski, W., Olendorf, R., & Stewart, C. (2018). How important are data curation activities to researchers? Gaps and opportunities for academic libraries. *Journal of Librarianship and Scholarly Communication*, 6, eP2198.

Kim, S., & Choi, M. S. (2016). Study on data center and data librarian role for reuse of research data. In Knowledge and Smart Technology (KST), 2016 8th International Conference on IEEE, 303-308.

Lee, D. J., & Stvilia, B. (2017). Practices of research data curation in institutional repositories: A qualitative view from repository staff. *PLOS One*, *12*(3), e0173987.

Li, J., Lu, M., Dou, G. & Wang, S. (2017). Big data application framework and its feasibility analysis in library. *Information Discovery and Delivery*, *45*(4), 161-168.

Beyer, M. A., Laney, D. (2012). The importance of “Big Data”: A definition. Retrieved from

https://www.gartner.com/doc/2057415/importance-big-data-definition

Marzullo, K. (2016). Administrative issues strategic plan for big data research and development. White House blog. Retrieved from https://www.whitehouse.gov.

Narendra, A. P. (2016). Big data, data analyst, and improving the competence of librarian. *Record and Library Journal*, *1*(2), 83-93.

Noh, Y. (2015). Imagining library 4.0: Creating a model for future libraries. The Journal of *Academic Librarianship*, *41*(6), 786–797.

Palmer, M. (2006). Data is the New Oil, ANA marketing maestros. Retrieved from https://ana.blogs.com/maestros/2006/11/data_is_the_new.html

Provost, F., & Fawcett, T. (2013). Data science and its relationship to big data and data-driven decision making. *Big Data*, 1(1), 51-59.

Showers, B. (2015). *Library analytics and metrics: using data to drive decisions and services*. UK: Facet Publishing.

Weiss, A. (2018). *Big data shocks: An introduction to big data for librarians and information professions.*
Maryland, Lanham: Rowan & Littlefield.

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., & Bouwman, J. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3.

Yoon, A., & Donaldson, D. R. (2019). Library capacity for data curation services: A US national survey. *Library Hi Tech*, *37*(4), 811-828. https://doi.org/10.1108/LHT-12-2018-0189

Zhan, M., & Widén, G. (2017). Understanding big data in librarianship. *Journal of Librarianship and Information Science*, *51*(2), 1-16. https://doi.org/10.1177/0961000617742451