大數據時代下的思維變遷，並以圖書館定位思考為例

一、 在閱讀完本書的當下.....

日常生活中的我們，對於資訊唾手可得的情景早已司空見慣，從起床開啟電腦到入睡前的手機滑動瀏覽，龐大的資料活動從不停歇。或許是體認也體驗了這樣的事實，在閱讀完本書之後，原本受益也受制於這片廣大資訊海的自己，在一直摸不清自我在資訊時代中定位的當下，有了一個明朗的思考方向。「大數據」，這個也稱作「巨量資料」或「海量資料」的名詞，已成為現在資訊時代下重要的新興概念之一，大數據的思考模式改變了我們的固有思維也改變著我們的生活，我想本書可以做為入門者認識大數據樣貌的第一步，它並不如同時下一般的教科書，正襟危坐地講述晦澀主題、傳授專門學問，而是循序漸進地將大數據的概念釐清，用生動而清晰的實例繪出架構、說明現況以及預測未來，聲聲喚起了資訊時代份子都該有的自覺。

二、 何為「大數據」?

翻閱幾本探討大數據的相關書籍，不同作者給予幾種定義，在城田真琴所著的「大數據的獲利模式」中指出:巨量資料狹義是指「用現有的一般技術難以管理的大量資料群」。難以進行管理的主因，以

3V—量（Volume）、多樣性（Variety）、出現頻率或更新頻率（Velocity）表示。廣義來說，指的是「從 3V 的角度難以進行管理的資料，以及為了儲存、處理與分析這些資料的技術，此外，包括分析這些資料並從中萃取有用資訊或富有洞見的人才與組織之全盤概念」。在胡世忠所著的「雲端時代的殺手級應用：Big Data 海量資料分析」中，作者提出了 4V 理論：1. Volume： 巨量性。2. Velocity： 即時性。3. Variety：多樣性。4. Veracity：不確定性。不管是 3V 又或 4V，可以發現巨量資料的特性不外乎是由數量巨大、結構複雜、類型多樣的資料所構成的資料集合。在本書中，巨量資料依循上述定義，歸納出了三大思維模式的改變:

樣本＝母體

大部分的人都有讀過詩人徐志摩的名句：「數大就是美」，他所談的是文學，而大數據談的則是宏觀世界裡的真相。過去人們為了獲取一些調查或分析研究上的資料，但礙於經費、人力、時間、設備等技術上的限制，因而在資料的蒐集和分析上發展出了沿用多時的抽樣統計學，透過代表性或隨機性的「抽樣」或「統計學」的方式來完成。此舉可以降低蒐集資料的成本，利用少量的資料樣本推斷出母體的樣貌，但相對的卻會造成了調查上的不精確，雖其準確度的誤差在一定的範圍內影響並不明顯，但卻可能造成整體系統性的偏差、或難以做

更多研究延伸。現今處理巨量資料的能力已逐漸成熟，做到樣本等於母體的理想，而巨量指的是相對的概念，是要有完整或盡量完整的資料集合，加以整理、分析後，便能成為有用的資訊，我們便可以查看細節或進行全新的分析，了解資訊片段間的相互關係，過去抽樣統計中模糊不清的地帶將會因資料的完整性而明朗。

雜亂

過去小量資料的世界裡，所追求的是精準的測量，以求能夠準確地蒐集、紀錄和管理資料。然而在巨量資料的時代裡，必須顛覆精準的概念，不精確而種類繁雜的資料形態是主體，也就是巨量資料的雜亂性。資料的雜亂性包含資料中參雜的錯誤、不同源頭和類型的資料、以及資料本身格式上的不一致。具體上來說，資料由各種文字、影音、網頁等類型所組成，而每種類型的資料又包含許多不同的格式，當資料的來源變得更多元時，資料本身的可靠度便降低，錯誤率提高。我們應當接受資料增大時所同時產生的雜亂事實，讓數據的重點由精確走向可能性。本書中提到一個具體的例子就是 Google 的語言翻譯，透過大量而雜亂的資料，擴大語言翻譯的規模和精確度，除此之外，社會網絡媒體的標籤功能、資料庫架構的設計，也因而將著重的點由質轉為量，創造更大的價值。因此容忍巨量資料的雜亂的性質，以及取捨、放寬允許的誤差值，就能讓更多資料獲得在各方面的使用，創

造出更大的效益。

相關性

這項思維的改變可說是一大劇烈的轉變。人們在過去的世界裡，在意的是事物的因果關係，常會預先建立一套假說，再找出原因、驗證和歸納預測的結果，但在巨量資料的時代裡，原因已變得相對不重要，焦點在於觀察事物彼此有何相關的影響現象。但所謂的相關性並非絕對的預支未來，而是有一定的可能性，透過現代科技的電腦分析，在沒有歸因作為思考前提來理解的預設之下，將更有助於我們對真實世界的理解，未來，透過豐富的資料將會代替了過去的假設，成為瞭解的起點。書中所舉的例子是網路書店龍頭亞馬遜，利用電腦找出巨量資料有何相關性的性質來取代耗費人工所寫書評的創舉，藉著交易記錄大量資料的蒐集、使用及分析，用來為讀者推薦新書，而締造了銷售量的業績。如同書中所述，在大數據的時代裡，我們必須拋下過去對因果關係的執著，轉而擁抱簡單的相關性。我們不必知道「為何如此」，只要知道「正是如此」就足夠了。

三、　大數據的價值

在瞭解大數據的意涵之後，便要瞭解大數據具有哪些過去資料所非具備的價值。首先，為數眾多的各樣資訊充斥在生活裡，必須先都

成為資料才能呈現出其價值，當一切成為資料，用途無窮盡。

資料化

在文字資料部分，Google 將圖書的內容文字掃描成電子檔，讓電腦可以處理、演算和分析，使得資料化後的圖書文字不再只是頁面影像，具有資料化後的價值；在位置資料的部分，當空間也資料化之後，除了可以提高行動服務的效能，也能創造出更多其他的相關價值；除此之外，在互動的資料部分更是驚人，人類的情緒、態度和行為將一覽無遺。資料化成為了現代基礎建設，種種應用的潛力將無遠弗屆。

價值與蘊涵

大數據時代所指的價值是「所有資料」本身就有其價值。本書中提到資料的價值取決於我們用盡所有的方式來使用，是所做選項產生的價值總和，也就是「選項價值」，一反過去資料完成原始用途便放手刪除的概念，而是透過三種方式:重複使用資料、合併資料集、找到「買一送一」的情況，讓資料的價值真正釋放。了解到資料價值在於其選項意義之後，接著在下一章節說明資料的價值是在於使用以及使用的方式，提出資料的價值鏈，從資料持有人、資料專家到有巨量資料思維者，這三類型人才各有優勢，但卻不固守於其中某個環節，而是將他們的資料技術交叉運用在各式各樣的領域裡，產生新型態的

中介機構，或是讓非營利組織的重要性增加，也改變了傳統專家的地位。最後總結出商業經營的新方向，企業若能擁有資料的一定規模或是善用低成本而有創意的經營模式，才能創造出最大的資料價值。

四、　心得-思考大數據與圖書館

閱讀完本書之後，發覺大數據概念對於服務於圖書館這個資訊機構的自己的來說，別有一種深刻的體認。面對現在資訊量的擴張和其唾手可得的性質，圖書館面臨的挑戰是許多圖書館員已能遇見的。巨量資料時代的來臨是一項即將動搖到圖書館地位的現象，當 Google 和許多網路資源正利用龐大的資料發展旺盛，許多現代人也同時開始質疑圖書館存在的意義和價值。然而在圖書資訊學領域裡卻努力將危機化為轉機，讓巨量資料成為強化圖書館服務的力量。國內有許多相關的研究，在這裡提出幾點個人的淺見，用以說明圖書館服務和大數據概念結合的可能和思考。

首先，必須先釐清圖書館的定位。在傳統的定義上，圖書館是收藏圖書和各種有形資訊媒體的場域。然而隨著電子資料型式的出現，圖書館也透過電子書、資料庫或其他網路媒體來提供資訊服務。因此，在現在數位時代裡，圖書館可以重新定義為能夠獲取多種來源、多種格式資訊的中心。除了提供資訊，圖書館透過學科館員等專家提供讀

者服務，研究讀者的資訊需求、資訊搜尋行為和最佳的資訊組織方式，讓圖書館的資源更貼近使用者。綜合上述來說，圖書館便是資訊使用者與資訊之間的中介者。接下來將做大數據應用於圖書館的討論，希望除了能從圖書館內各種資訊資源的角度來探討資訊組織或檢索系統的改良，也能從讀者資訊需求和搜尋行為的方向來分析圖書館整體服務的提升。

圖書館可以說是典型的巨量資料的儲存場域，因為圖書館本身便是一個資料蒐集、儲存、組織和流通的中心，其內部保存的紙本圖書館藏、出版刊物或蒐集來的各式資料，類型多樣、格式不一、資料複雜，都可算是屬於巨量資料範疇。除了內部提供讀者使用的各式資料之外，在抽象的服務上，圖書館整體運作方式、書目資料內容、圖書館的流通服務、參考服務或專業訓練課程等等，這些運作活動在資料化之後均可做為巨量資料的分析題材，具有相當的價值。因此，大數據在圖書館的分析面相相當多元，在此提出幾點分析來做為簡單的討論。

當圖書館所能提供的資訊眾多且繁雜時，透過巨量資料性質的應用來改善圖書資訊檢索系統，便能提高資訊對於讀者的易近性(Approachability)，根據書中對於大數據特性的描述，包含樣本等於母體、雜亂和相關性，我們可以將圖書館的館藏資源套用同樣的模式，做出適當的整理、保存、分析或對資料加以利用各種工具以開發出資料庫檢索系統，協助使用者在眾多的資訊海確認資訊需求，並搜尋和使用需要的資源。相關實例如資訊選粹服務和社會性標籤，均是透過蒐集龐大資料被利用的資訊，做出相關性的分析，提升資訊系統服務的主動性。

1. 資訊選粹服務（Selective Dissemination of Information，簡稱 SDI）

是圖書館一種針對讀者個別興趣，選擇最新資訊的新知報導服務。如右圖所示，為臺北市立圖書館的新書通知服務，會以主動積極的方式，定期提供資訊服務，節省使用者檢索資料的時間，加強館藏資源的利用，過去 SDI 定義為採取人工或自動的方式，針對個別使用者提供現況通知服務，選擇相關研究興趣的新資訊，定期傳遞給讀者，篩選以及發送最新的資訊，幫助讀者利用或取得所需資訊。

但在系統技術成熟的背景之下，圖書館的資訊選粹服務相較於過去的限制，現今已有能力蒐集更多的資料，利用為數眾多的書籍流通資料和讀者資料的交叉分析和比對，找出閱讀的相關性，將預測讀者興趣和專題選粹的服務做得更為精準，類似書中提到的網路書店相關書籍推薦的完善功能，提供更完整的閱讀建議。目前圖書館界最常利用的概念便是「資料探勘（data mining）」，探討如何發掘此一讀者個人化的書籍推薦。陳垂呈(2005)提到:在探勘過程中，會比對此一讀者借閱資料與其他借閱資料的相似度，依據其是否符合所設定的條件，來分別設定其與此一讀者借閱資料的關聯性為高或低，並視其他非此一讀者曾借閱過的書籍項目為影響屬性，然後對讀者的借閱資料進行分類分析。此探勘結果，對圖書館在擬訂最適性之讀者個人化書籍推薦時，可以提供非常有用的參考資訊。

2. 社會性標記(Social Tagging)

社會性標記也算是大數據概念的一大應用，在 Web 2.0 風潮中，社會性標籤是關注的重點之一，意指讓使用者自由給定詞彙。何謂社會性標記？從詞面意義分析來說，社會性標記包含了社會性（social）及標記（tagging）的意涵，「標記」與圖書資訊組織中的分類索引概念相近，即針對資源的內容，提取重要概念給予詞彙，以利資源之組

織與檢索；而「社會性」則是強調標記的分享及互動特性，亦即使用者的標記過程與產出皆會受到眾人並對眾人產生影響。簡言之，社會性標記是指在公開的環境之中，每個使用者個人對於各類資源都可進行標記或分類的動作，進而集結成眾人共同標記或分類的機制。

社會性標記通常運用各式各樣的「標籤雲」模式呈現於圖書館的檢索系統。標籤雲是一種將單一詞彙，以不同顏色大小的字型顯示，用簡潔單一的視覺畫圖形成線主題所引的網路應用方式，標籤為指向相同的主群體連結，為一導覽工具。也就是使用者可以透過此一機制為其所感興趣的數位內容賦予一個或多個的分類標籤，同時可以透過標籤的連結瀏覽所有被賦予相同分類標籤的資料或被關注的議題。如果點選「標籤雲」的顯示模式，除了可以顯示全體使用者對該筆資料所建立的分類標籤或關鍵字詞之外，亦可透過標籤字體大小與顏色的變化，分享其他使用者對該筆資料的觀點，以及這些觀點被關注的程度。

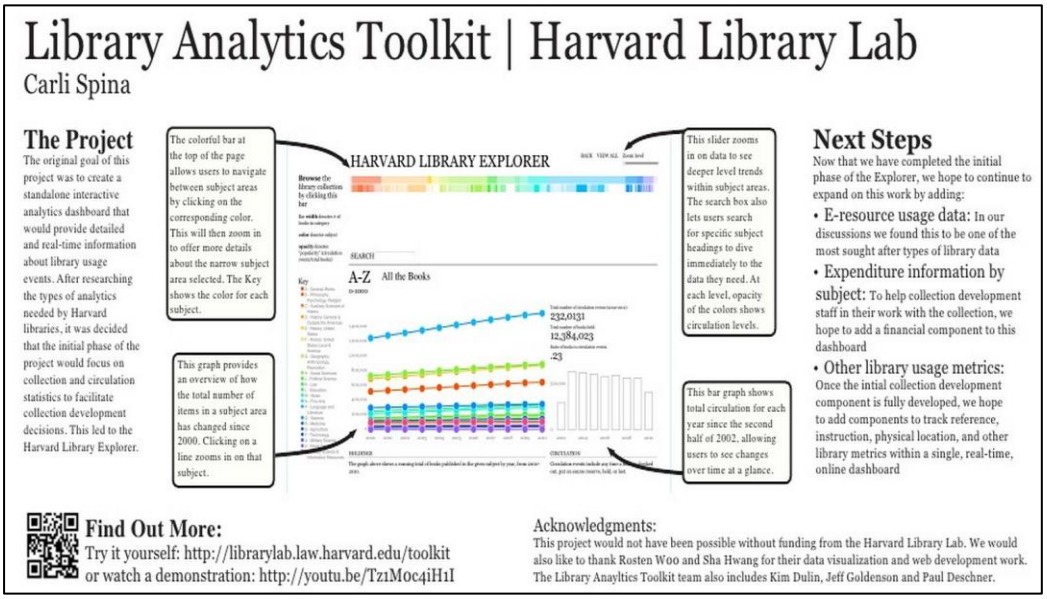
前述中提到，社會性標記與圖書資訊組織的分類與索引概念有關，相較於過去以專家為基礎的資訊組織方式，此種方式是以使用者為中心的資訊組織模式，集結所有資訊使用者共同依照自主判斷來做的主題分析，雖然使用者標記品質可能參差不齊亦或是資訊正確性不足，但大數據所具有的特性便是巨量而雜亂，能從宏觀的角度看出使用者

與資訊之間的相關性，貼近使用者的認知。(下圖為國立臺北藝術大學及國立臺灣師範大學館藏目錄上之「熱門關鍵字」呈現)

3. 電子資源系統的多向使用者行為分析

圖書館中各類型的巨量資料是研究使用者資訊行為以及評估和分析系統的珍貴題材，如同書中提到的使用者留下的「資訊廢氣」具有分析價值一樣，從圖書館資訊系統中，也可以透過讀者在操作使用的過程中留下的各種資料來做使用者行為的分析。例如資訊計量研究，可以研究某位讀者重複造訪圖書館或網站的次數，找出讀者相關的各

類型的統計；也可以分析統計讀者使用系統的時間，看出系統對於讀者使用習慣的影響；分析檢索詞彙的檢索次數或資源被點閱的次數，瞭解熱門資源的種類等等。舉一實例如下圖所示，此為哈佛大學 Library Analytics Tool-Kit Project，2010 年先進行圖書館自動化紙本館藏使用率分析，2013 年則計畫將 Aleph 與 COUNTER 統計數據匯入於 Library Analytics Tool-Kit 中，並於 Tool-Kit 完整呈現美國國會主題詞表，利用國會圖書館類表及主題表架構的分析數據，觀察讀者尋找資訊的行為軌跡。



由此可知，圖書館員若能有效利用系統提供的功能與工具進行巨量資料的探勘，並將分析後的結果提供給管理者作為決策分析的參考，便能使其成為強化圖書館管理及服務效益的最佳利器。透過各種資料分析的「量」化研究，可以獲得改善圖書館服務的建議，確認讀者需

要的資訊資源以及改良資訊檢索或瀏覽系統的介面等等，得以讓圖書館這個介於讀者和資訊之間的中介者，達到提升使用者滿意度的最終目標。接著在最後，我想提出圖書館在面對巨量資料時代時，對於內部或外部的龐大資訊海所做的相關討論和研究議題，從資訊組織的角度出發，觀看圖書館可能提供給資訊使用者的指引和服務。

4. e-Research

大數據時代的來臨，科技與技術的革新使得資料氾濫的情況同時發，促使各個學門在學術研究上的模式和環境改變。圖書館位處於資訊中介者的角色，數位化型態的研究資料成為研究者的核心題材，在此圖書館便須因應「e-Research」時代的浪潮並確認「資料度用」的挑戰，其中尤以學術(大學)圖書館尤重。「e-Research」的這個名詞的誕生便是用以呼應巨量時代的內涵，其特徵是大量研究資料的運算、保存和再使用，並同時可以促進學術研究的進展。而「資料度用」則是圖書館可以運用其特長之數位典藏、機構典藏、開放取用的知識，協助研究社群設計與實踐包含資料描述、存儲管理、重複使用在內的規劃。

因此，面對 e-Research 的環境，學術圖書館發展資料度用是重要的課題，在臺灣大學圖書館做的研究提出三種面向的建議，可作為同

型大學圖書館館員的參考的指引，分別是：1)精進圖書館員知能以勝任資料度用任務；2)實施需求調查以奠定資料度用基礎；以及 3)規劃合作策略以推動資料度用服務。雖然學術圖書館在 e-Research 與資料度用的服務或應用有其積極的角色，但是卻仍須回到服務的最根本，也就是關注使用者的研究需求，不同的學科領域研究者的需求會有差異，因此進行研究者的需求調查也是重要的過程。資料度用為支援 e-Research 程序的重要環節之一，在了解使用者的資訊需求之後，便須建構良善的資料典藏系統，接著，再擴展系統的多樣化及多功能整合性的 e 化研究環境。

e-Research 對於研究的意義，不僅僅是資訊科技對研究過程的涉入和輔助，它同時也帶來研究行為和研究文化生態的轉變。面對資料繁複的資訊世界，資訊服務系統並非萬能，系統開發者的工作目標並不是建立一個無人介入的全自動化環境，而是希望能藉由資訊科技的深入輔助，減少研究者對資訊取用的障礙及複雜度，讓研究者與資源間能夠更有效地互動，提昇研究工作的效能，並更進一步地加速研究週期的循環。隨著 e-Research 系統各方面發展，系統平臺扮演的角色，應該超越以往的後援功能，而是與研究活動的各個程序環環相扣，真正在大數據時代的環境下讓使用者與資訊能夠有正確的連結，也讓學術圖書館的價值因資料度用的實踐得以在學術研究領域內展現。

五、　小結

大數據的時代創造了如同拼圖般的可能，其能應用在許多不同的場域，讓各種瞭解大數據價值的產業有了新的發展契機，雖然價值無盡，但不外乎的，這樣的情景也會產生許多負面議題需要我們做深入的思考。書中提到了幾點大數據可能產生的危機，首當其衝的當是隱私權的侵害，像是巨量資料的分析導致監控無時無刻充斥在生活中，也讓過去頗有成效的匿名技術這個保護使用者隱私權的方式不再適用；除此之外，由於巨量資料具有預測的功能，因此可能導致誤判犯罪事實的情況發生；再者，對於巨量資料的準確性有過度的信心常導致許多謬誤的發生。由此可知，不管是隱私權議題亦或是準確度的價值，巨量資料裡的數字並非無所不能，最終我們仍必須回歸到「人性」的出發點。

本書的最後提出了三大管控的機制來面對大數據時代的負面效應，可以給各個領域的我們有一些方向來做應變。首先是在個資保護上，必須由資訊使用者肩負使用責任；再者是在利用巨量資料進行預測時，依舊以尊重每個人的能動性為原則；最後，在巨量資料時代下，必須培養「演算學家」這樣的審計師來做控管的工作，透過這些因應的策略，人們才能避免被巨量資料牽制，因而有了駕馭的可能。

閱讀完這本書之後，除了讓我了解到自己身處於這個巨量資料體系下的處境之外，也對巨量資料有了更全面的認識。巨量化的資料就像數以萬計的拼圖，拼湊出原先我們看不到的圖像，大數據所帶來的利益絕對不容小覷，甚至能扭轉原先深根柢固在社會裡的各種認知和事實，讓人類的視野無限擴張。在了解到其樣貌和價值之後，我也回顧了自己熟悉的領域，思考圖書館在這樣的時代下，所能保有存在價值的可能，以及如何提升和強化服務。我想，在未來，能有更多創新的突破去運用各種形態的巨量資料，便是一項重要的成敗關鍵，也期許自己能在這樣的時代底下運用創新的概念加強工作領域內的效能。當然，一切事物都有正反面，不管大數據所帶來的效益是多麼龐大，最終還是必須回歸到人性的原點，讓巨量資料僅只成為利器，並充分建立在謙卑的基礎之上。

六、 參考資料

1. Franks, B. (2013). 駕馭大數據：從海量資料中挖掘無限商機: 碁峰.
2. 胡世忠. (2013). 雲端時代的殺手級應用：海量資料分析: 天下雜誌.
3. 城田真琴. (2013). Big Data 大數據的獲利模式 : 圖解.案例.策略.實戰 (鐘慧真、梁世英, Trans.): 經濟新潮社.

4. 陳雪華、陳光華. (2012). e-Research: 學術圖書館創新服務. 台北市: 國立臺灣大學圖書館.

5. 柯皓仁、楊雅雯、吳安琪、戴玉旻、楊維邦. (2002). 個人化及群體化圖書館資訊服務初探. 國家圖書館館刊, 91(1), 161-195.

6. 林倩妏、卜小蝶. (2010). 標籤雲在圖書資訊應用服務之初探. 國立台灣師範大學研究所.

7. 陳垂呈. (2005). 利用資料探勘技術發掘圖書館個人化之書籍推薦. 教育資料與圖書館學, 43(1), 87-107.

8. 楊雅雯、柯皓仁、楊維邦. (2000). 個人化數位圖書資訊環境 – 以 PIE@NCTU 為例. Paper presented at the 2000 年台灣區網際網路研討會.

9. Library Watch (http://ifii-enews.blogspot.tw/2013/08/mining-data-for-library-decision-support.html)