Pre-print Version/ Submit to Journal of
Library & Information Science (2016/11/18)
URI: http://data.odw.tw/article/20161122

結構資料的再次使用：語意、連結與實作

黃韋菁、李承鑫、莊庭瑞
中央研究院 資訊科學研究所
E-mail: {andreahg, cjlee, trc}@iis.sinica.edu.tw

關鍵詞：開放資料連結(LOD)、知識庫、資料品質、資料溯源、語意再現、知識本體、CKAN

【摘要】

持續創造資料的語意與連結，藉由全球資訊網散布同可由常人和機器處理並理解的結構性資料，以增進資料集的「再次使用價值」 (Reuse Value)，是目前廣受重視的課題，也是本研究由理論探討邁向系統實作的動力與目的。本文簡述與「開放資料連結」 (Linked Open Data, LOD) 相關的國際計畫與技術發展，介紹以「開放資料連結」方式建置的五項跨領域知識庫和七項專業知識庫。我們並解析資料品質、後設資料 (metadata) 及資料溯源 (provenance) 的關聯脈落。我們的實作網站 data.odw.tw 收納並轉換典藏品的目錄資料為富語意結構的連結式資料，其中我們使用並擴充 CKAN (The Comprehensive Knowledge Archive Network) 此資料集管理系統，作為連結式資料的儲存與展示平台。CKAN 此軟體的程式碼係以「開放原始碼」 (Open Source) 方式釋出，而我們的資料來源也採「創用 CC」(Creative Commons) 公眾授權方式釋出，這讓我們可以在開放的基礎上發展可被自由使用、自由擴散的方法與內容。我們強調從原始目錄資料到語意連結資料的分段轉換步驟，並將各步驟的轉換程式以開放原始碼方式釋出。

前言

「資料連結」(linked data)關聯了資料(Data)、常人(Human)以及機器(Machine)三者在知識呈現與語意處理的共有課題。假使我們想要詢問台灣國寶級植物——台灣一葉蘭的地理分布為何？在目前的典藏台灣聯合目錄中[2]，確有該植物標本的紀錄，其後設資料記載了採集地點以及相關描述資訊。該項目錄資料經知識呈現及語意處理後[3]，可以連接外部地理資源以及相關藏品資訊，用來回答典藏目錄裡不同的台灣一葉蘭標本藏品之地理分布情形(圖1)[4]，同時經由外部地名資料庫，也可以顯示採集地如宜蘭大同等地的其他資訊，進一步提供典藏目錄裡這些標本的地理分佈資訊以及採集脈絡：例如，該植物的採集時間前後跨越 1983-2010 近三個世代、以及標本資料建置的相關人員資訊等。

我們使用 CKAN (The Comprehensive Knowledge Archive Network) 這項「資料藏庫」(Data Repository) 系統軟體工具，在上面架構了「開放資料連結」(Linked Open Data, LOD)的新功能，整合提供可被常人以及機器皆可使用的連結資料服務(Lee, Huang & Chuang, 2016)。如今，典藏目錄裡台灣一葉蘭的資訊，在此資料連結架構下，提供了同時適合機器與常人皆能理解的資料連結介

面，進而能自資料連結的語意關係進行相對應的知識查詢。換言之，「開放資料連結」的最終目的是支持語意網的運作，本文所介紹的方法與實作，有助於常人對資料進行理解 (interpretation of data)、資料能被機器操作 (machine actionable)、資料間可建立語意連結 (semantics linked)、且資料能被語意檢索 (semantic query)等工作，走向人機語意連結與互通的語意網願景。

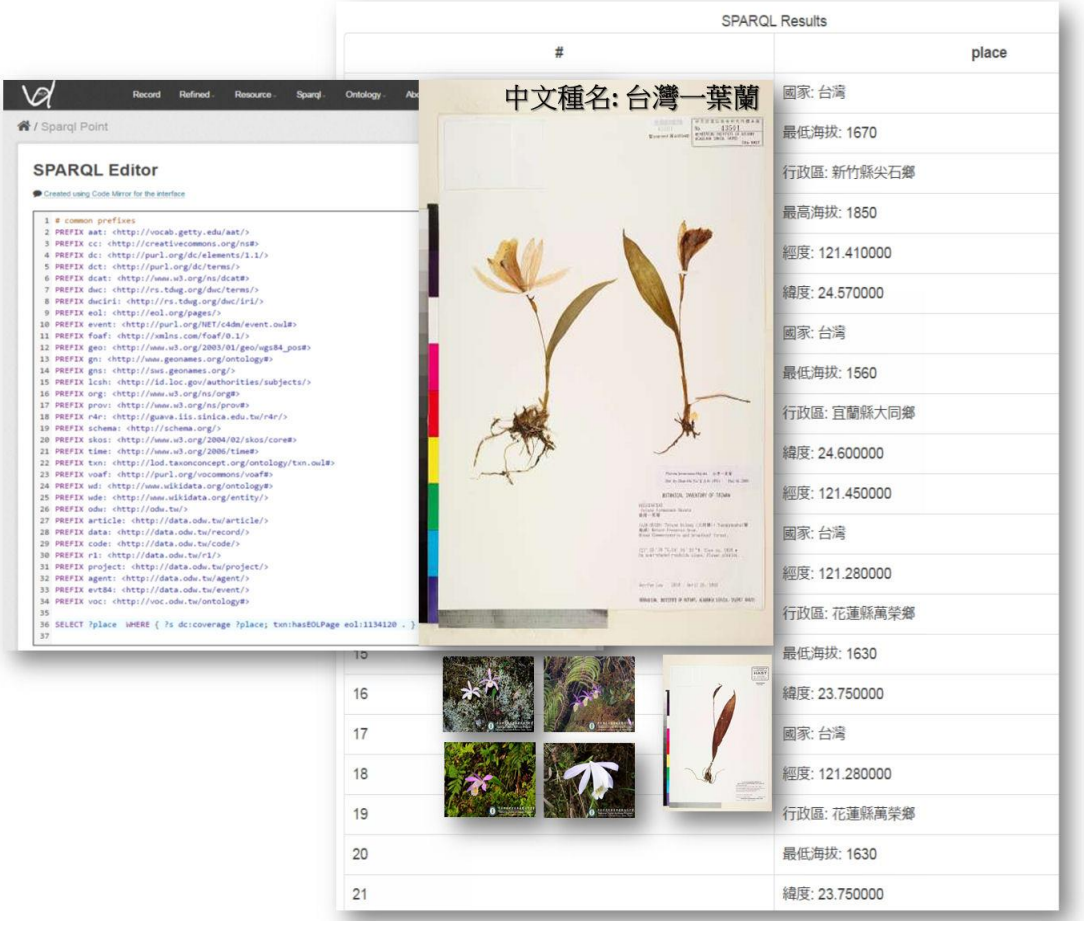

圖1: 以 SPARQL 查詢台灣一葉蘭標本藏品之地理分布

在此語意網與資料連結的願景中，我們看見圖書、典藏、博物館界(Libraries, Archives and Museums, LAM) 在後設資料語意標示的傳統，珍視服務使用者應用的實務操作經驗，以及積極建構以知識為主的索引典 (Thesaurus) 與特定學科主題的控制語彙 (Control vocabularies) 等特點。這些優勢皆彰顯圖書、典藏、博物館界進行「開放資料連結」工作的優勢與價值。本研究使用典藏台灣聯合目錄後設資料為案例研究，得利於後設資料原始檔案為半結構化 XML 形式，以及其使用「都柏林核心集」(Dublin Core, DC) 15 欄位，已整合了異質資料來源。儘管如此，我們的實作也證實開放式的資料連結過程中仍面臨許多挑戰，誠如 Hallo, Luján-Mora, Maté & Trujillo (2016) 研究指出，圖書典藏博物館界在「開放資料連結」發展所面臨的困難包括：(1)技術工具的支持、(2)資料品質控制機制、(3)資料模型與語彙的實作、(4)人性化的瀏覽與查詢界面、(5)定義資料開放授權的困難，以及(6)缺少新技術知識的技術人員等。藉由實作方面的經驗分享，我們希望可以引發更多討論，並或可推進這方面的工作。

「開放資料連結」以及「開放資料連結知識庫」的發展

始創資料連結的 Tim Berners-Lee 於 2006年指出：「驚人數量的資料呈現未連結的狀態」[5]。十年後2016年的今天，我們或可去掉 「未」字，改為：「驚人數量的資料呈現連結的狀」[6]。儘管如此，圖書典藏博物館領域在面對「開放資料連結」卻仍面臨許多挑戰。例如，2013年 Marden, Li-Madeo, Whysel & Edelstein (2013) 分析當時15個文化資產 的「開放資料連結」計畫後指出，文化資產資料無法廣泛被使用的最大障礙是：大多數機構尚未以開放資料連結方式發佈與使用資料。

三年後此令人擔憂的狀況獲得改善，主要誘因包括：為增加資料曝光度吸引更多使用者、示範資料集能完成資料連結程度、普遍聽聞「資料連結」趨勢而嘗試、測試資料連結是否能優化搜索引擎效能等 (Mitchell, 2016; Godby, 2016) 。如「線上電腦圖書館中心」（Online Computer Library Center, OCLC）針對 20 個國家 90 個圖書博物館機構的最新報告指出[7]，相對於 2014年的調查，資料連結計畫已快速成長兩倍；歐盟計畫 Europeana[8]自2008年11月至2016年4月，整合歐盟約3500個機構，五千二百萬筆藏品物件的後設資料，於2010年更發展 Europeana Data Model (EDM) 開始邁向資料連結 (Haslhofer & Isaac, 2011)，目前則積極進行 Europeana Semantic Enrichment Framework 的工作[9]針對語意加強、資料品質以及評估三大方向進行。

學術界如自2014年起美國的 LD4L Labs (Linked Data for Libraries Labs) 計畫[10]，由哈佛大學圖書館創新研究室、史丹佛大學圖書館、康乃爾大學圖書館三機構共同合作發展超越傳統後設資料的全新蒐尋方法，於今年起擴大該計畫[11]並規劃將技術成果提供給另外三機構[12]共同進行 LD4P (Linked Data for Production) 計畫[13]。該計畫主要目的為發展超越傳統後設資料的全新蒐尋方法，針對學術資源如傳統專題著作、期刊發表、檔案資料、研究資料集、圖檔影音媒體、文化器物、新聞雜誌、甚至網路典藏等，進行資料脈絡與關係之語意網路平台系統整合與建置，學術資源語意資訊倉儲(Scholarly Resource Semantic Information Store, SRSIS) 以及知識本體的建置與維護。預期將不同單位間的資料連結的工作流程標準化，並藉由 LD4P 六機構所產生的連結資料，同步發展技術服務。

自上述美國圖書館界在資料連結計畫的發展趨勢，學術界也不忽略這一波整合學術資源資料連結目的。「開放資料連結」在學術應用層面上，若以歷史研究領域為例，可知資料連結與語意網技術所提供的控制語彙與知識本體，可解決史料欠缺正規化與隱含知識推理的探究問題、資料整合則提供散落各處的獨立史料的連結機會、資料互通則提供史學家新的資料搜尋與資料擷取的機會、RDF(Resource Description Framework)資料模型提供了史料不論採取「來源導向的再現模式」(source-oriented representation) 或「模式 導向或是目標導向的再現方式」(model-oriented or goal-oriented representation) 一個更彈性且易於因應不同情境脈絡變化的設計選擇。

例如，在資料轉換與更新過程中，史學家主要面臨的挑戰是保持原始資料完整性，以及能追溯資料轉換的過程，而 SPARQL 這種 RDF 查詢語言，其所提供的 CONSTRUCT (根據查詢結果自動建構 RDF 圖）與 SELECT (選擇顯示查詢結果欄位值) 等資料網絡建構與選取方式，可提供選擇呈現不同問題觀點的需求與結果。SPARQL 查詢方式不需要改變知識庫原先的狀態即可根據不同觀點需求進行，也因此提供了傳統知識庫中使用較無彈性所建構的資料模型的替代方案。而對於資料轉換的追溯，連結資料所重視的「資料溯源」(provenance) 則提供了解決方案 (Meroño-Peñuela et.al., 2014) 。

近幾年資料連結與語意網技術與巨量資料的結合不僅在定義上密切相關，同時「開放資料連結」也提供了整合異質性資料成為可理解的巨量資料(understandable big data) 的角度，用以偵測資料不一致性，並可透過推理引擎或連結外部資料產生新知識，皆賦予巨量資料更多資料處理與利用價值 (Emani, Cullot & Nicolle, 2015)。以南加州大學處理文化機構巨量資料為例，當面臨資料差異與資料異質整合的問題時，透過其所發展的開放原始碼工具 Karma，可針對不同資料來源與格式的資料進行整合，並利用知識本體語彙進行語意對應 (Knoblock and Szekely, 2015)。另一方面，研究亦發現自2014年起，行動裝置結合「開放資料連結」與語意網技術的 APP 大量快速發展，技術方面也從早期行動裝置僅扮演用戶端，語意資料處理有賴遠端伺服器，提昇至近日發展為行動裝置用戶端也具備語意推理功能 (Yus & Pappachan, 2015)，「開放資料連結」貼近常人的每日生活為期不遠。

在「知識庫」 (Knowledge Base) 或稱「知識圖庫」 (Knowledge Graph)方面，近期也逐步採「開放資料連結」方式建置，提供如前述 Europeana 等計畫所著重的可豐富資料之連結對象。因此我們針對以「開放資料連結」方式提供，全球知名的五項跨領域知識庫專案，以及七項專業領域知識庫（主要為地理資訊為主，但亦簡略介紹我們因文化典藏需求，所使用的文化藝術類索引典，以及生物主題知識庫) ，進行考察比較。主要目的是探索這些知識庫因開放連結而展現一體兩面的效果：一方面積極對外連結以豐富自身知識庫語意，另一方面因豐富的知識資源亦成為其它資料庫及知識庫的連結對象。表1 列出這些「開放資料連結」知識庫的基本資訊。

表1: 五項全球「開放資料連結」跨領域知識庫與七項專業「開放資料連結」知識庫基本資訊 (2016/11/06) [3]

LOD 知識庫		起始	組織	資料性質	主要來源	個體量	三元組量	更新頻率
專家建構	OpenCyc	2008	商業	跨領域	自建	41,029	2,412,520	超過一年未更新
	Getty AAT	2014	商業	文化藝術	自建	45,327	13,259,890	LOD 後一年3-5次
	Getty TGN	2014	商業	地名	自建	2,495,100	204,614,290	
	Ordnance Survey	2010	政府	地理資訊	自建（二者視為同一專業知識庫)	2,938,707	58,377,209	視需求
	Open Names	2015	政府	地名		925,157	21,360,688	一年兩次
混合	EOL (TraitBank)	2014	學會	生物	整合專業資料庫/資料協作為輔	10,753,384	359,292,712	統計更新約一週
協同合作	Freebase	2008	商業	跨領域	Wikipedia	49,947,799	3,124,791,156	2015關閉
	YAGO	2007	大學	跨領域	Wikipedia	5,130,031	1,001,461,786	超過一年
	DBpedia	2007	大學	跨領域	Wikipedia	5,109,890	402,086,316	約一年/部分即時
	DBpediaPlace	2007	大學	地名	Wikipedia	816,252	53,895,946	
	Wikidata	2012	NGO	跨領域	Wikipedia	19,367,201	1,371,170,022	即時
	LinkedGeoData	2010	大學	地理資訊	OpenStreetMap	> 3 billion	1,384,887,500	約一年
	GeoNames	2010	NGO	地名	資料協作為主/整合地名資料庫	>6.2 million	93,896,732	即時

以開放資料連結的跨領域知識庫：五項知名專案的簡介

早期即由學術資源資料的開放而成為「開放資料連結」知識庫之先驅代表者為 Open-Cyc[14]：自2002年起以開發人工智慧的企業公司 Cyc，使用開放原始碼軟體建置知識本體與常識知識庫 OpenCyc，而後於2008年邁向連結資料[15]另以「開放資料連結」版本釋出[16]，其資料來源即為該公司提供給學術社群免費使用而所創建的 ResearchCyc[17]。另外 2007年美國軟體公司 Metaweb ，開發 Freebase [18] 以 HTTP/JSON 為基礎的 API 和以 RDF 為端點[19]，提供機器可抽取的資料庫，開放給一般使用者自由編修資料，並提供結構化的使用者參與介面 (Bollacker, Evans,

Paritosh, Sturge & Taylor, 2008)。2010年 Google 購買 Freebase 以此基礎建立 Google 知識圖庫，2014年底宣布將 Freebase 資料匯入 Wikidata [20]，在 Google 支持合作下藉由 WikiProject Freebase 陸續將資料與 Wikidata 整合[21]。

相對於商業公司在「開放資料連結」知識庫的進展，學術界方面則以德國學術圈發展最為活躍。「開放資料連結」知識庫的知名代表為德國馬克斯普朗克研究所的 YAGO(Yet Another Great Ontology)[22]，以及德國萊比錫大學、柏林大學與開放連結軟體公司合作的 DBpedia[23]。YAGO 源於網路搜尋引擎尚未對知識單位進行搜尋年代，即以自動抽取 維基百科 及 WordNet 知識單位方式，建立大規模實體與關係連結的知識本體 (Suchanek, Kasneci & Weikum, 2007, 2008)。其中最引起我們關注的是 YAGO 以時空為其語意資料模型的首要元素 (first-class citizen)，以事件和情境脈絡進行物件的語意再現 (Hoffart, Suchanek, Berberich, Lewis-Kelham, De Melo & Weikum, 2011; Hoffart, Suchanek, Berberich & Weikum, 2013)，是本研究在實作時設計物件語意強化版 (Refined Versions, R 版) 的啟蒙雛型（於實作步驟五詳述）。近年來 YAGO 亦和其他知識庫同步發展連結資料的多語系統 (Mahdisoltani, Biega, & Suchanek, 2015)，主要採取的方法是自然語言處理，因不在現階段本研究範圍內，故不再詳述。

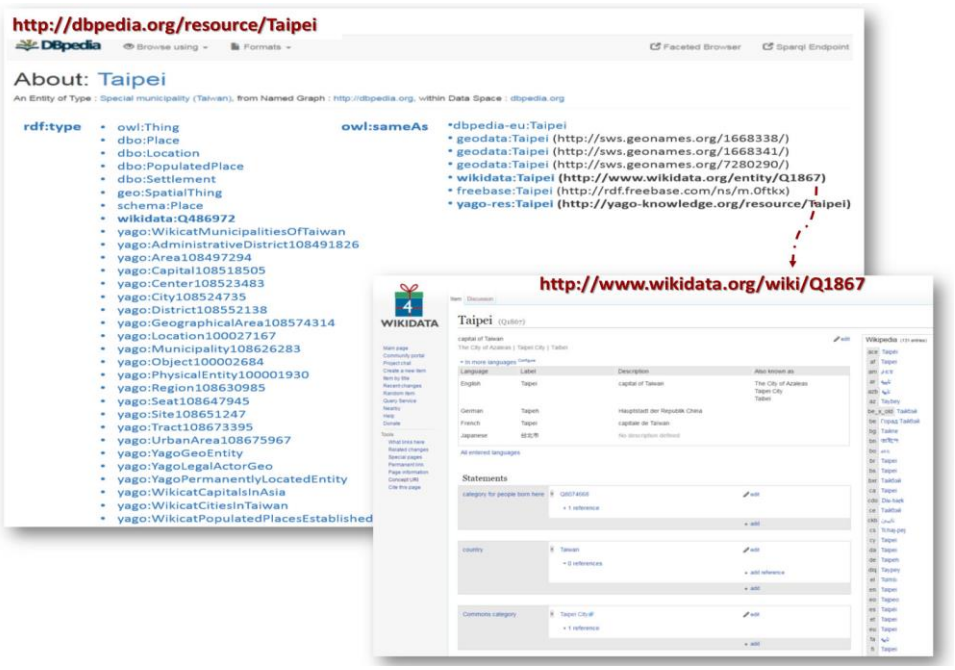

圖2: 以台北為例，DBpedia 中 Taipei 對外連結以及 Wikidata 連結外部 LOD 知識庫的情形

事實上最能反映「開放資料連結」與語意網的知識庫，非 DBpedia 莫屬。DBpedia 以 「維基百科語意網的反射鏡」[24]、「開放資料網核心」、「資料網結晶點」著稱，不僅與上述 OpenCyc, Freebase 與 YAGO 連結，自2007年1月初版之後約每年釋放一個版本，而後2011年也經 DBPedia Live[25]即時反應部分維基百科的資訊更新，相關研究更明確指出 DBpedia 在資料互通性與外部資料連結與相互連結[26]、以及多語機制設計等方面，提供了「開放資料連結」在解決問題與技術框架的領先示範 (Auer, Bizer, Kobilarov, Lehmann, Cyganiak & Ives, 2007; Bizer et.al., 2009; Lehmann et.al., 2015)。表1中，我們亦列出 DBpedia 子資料集 DBpeia Place 作為本研究實作連結地名時的參考，並將在探討地名「開放資料連結」知識庫時進行更詳細的分析。

以「開放資料連結」方式建置跨領域知識庫的許多專案中，目前最引起我們關注的是 Wikidata。在維基百科(Wikipedia) 建立十年後，由維基媒體基金會於 2012 年推動以資料知識庫成為維基百科的知識架構基礎。Wikidata 清理維基百科裡的事實性資訊，整合集中使其成為可重新利用的知識庫，提供多種資料格式如 JSON, XML, RDF 等。一方面 Wikidata 類同 DBpedia 或 Freebase 一樣，抽取維基百科的結構化資訊 (如 Infobox 區塊內的資訊)，另方面也抽取資訊來源以及資料情境例如時間有效性，使其資料溯源的機制更加完備。Wikidata 並設計每個實體 (entity) 具其概念網址(Concept URI)、以及其屬性名稱與屬性值所構成的陳述(statement)。這些陳述的設計相對其他知識庫更為彈性，例如可描述其屬性值是未知，或是「無/沒有」，例如澳洲「沒有鄰國」等 (Erxleben, Günther, Krötzsch, Mendez & Vrandečić, 2014; Vrandečić & Krötzsch, 2014)。其未來潛力可自上述 Freebase 的加入，以及 Europeana 的語意策略運用 (Charles, Manguinhas, Alexiev, Charles & Dammers, 2015))、DBpedia 增加比對連結(Ismayilov, Kontokostas, Auer, Lehmann, & Hellmann, 2016)、或是與專業知識庫如 VIAF 與 GeoNames 的高度連結 (Voß, 2016) 等方面，已得多方期待。以上均是本研究實作目標連結知識庫時考量的因素。

不僅如此，若自資料品質角度觀察比較這五項知識庫，Färber, Bartscherer, Menne & Achim Rettinger (2016) 以正確性、可信度、一致性、相關性、完整性、適時性、易了解性、互通性、可取得性 、授權 、相互連結等十一項指標研究後發現[27]：在正確性方面，在 RDF 文件驗證、文字語法驗證及「三元組」(Triple) 語意，五大知識庫大致表現優良。YAGO 在易了解性，如資源描述、多語標籤、提供多樣可理解 RDF 格式等表現最佳。Freebase 則是一致性及相互連結兩項指標的冠軍；而 DBpedia 在可取得性以及互通性上，不僅避免使用「空白節點」(blank node)所造成無法根據 URI 參照取得資源 (dereference) 的問題、同時提供多種資料格式、大量使用外部語彙、且幾乎所有第二層類別 (Class) 資料均連結外部資源的類別，因此在此兩項指標中領先，並與 Wikidata 同列授權指標表現優異者。

然而，最引起我們關注的 Wikidata 在可信度、相關性、完整性、適時性、授權等五項指標上比其他知識庫表現更佳。換言之，綜觀十一項指標，其中除正確性為五大知識庫持平外，Wikidata 在十項指標中具有五項指標最佳的優勢，若待後續 Freebase 資料陸續匯入後，亦可能延續 Freebase 在二項指標優勢而大幅超越其他知識庫。此研究結果亦呼應本研究第一階段評估綜合性知識庫作為連結目標時，選擇 Wikidata 而非 DBpedia 的方向。以下簡要討論為何在可信度方面，由專家建置的 OpenCyc 並未勝出，且 DBPedia 在資料的一致性上也未能超越 Freebase 與 Wikidata 的可能原因。

持平而論，以何種指標為檢驗項目，是所有品質研究可討論的議題，但是我們也可從 Färber 等人的指標設計上發現，以協同參與的機制建置資料、以及資料溯源資訊的完備，將會深遠影響資料品質。例如，該研究評估在一致性指標中，由於 Freebase 和 Wikidata 可由參與成員編輯，因此在用戶端界面中增加新陳述 (statement) 時，即可針對一致性進行簡易檢驗。另外，在可信度面向中，針對知識庫層級資料的匯入與策展，OpenCyc 與 Wikidata 可信度得到最佳評價，其原因為 OpenCyc 得利於專家建置而獲高可信度品質檢驗，而 Wikidata 有雙重的大眾參與機制為品質把關（資料匯入前通過維基百科社群檢驗，匯入後通過 Wikidata 社群檢視）。在陳述層級的可信度上，具「資料溯源」 陳述機制為基準的專案，如 Freebase, Wikidata 與 YAGO 都能有較突出的表現，其中又以 YAGO 能儲存每一陳述的資料來源與資訊擷取技術[28]，是五項跨領域知識庫中最為獨特的代表。

以開放資料連結的專業領域知識庫：地名資訊專案簡介

以上我們探究的是跨領域知識整合的「開放資料連結」知識庫，然而根據資料特性，專業領域知識庫常是語意資料連結的主要目標。鑒於本研究尚處於建置初期，基本的後設資料如時空資訊必須先組織整理，才能初步連結各項典藏品的語意。也因此地理資訊相關的知識庫如 GeoNames[29], LinkedGeoData[30](OpenStreetMap 資料的 RDF 呈現[31]), DBpediaPlace (DBpedia 子資料集) 與 Getty Thesaurus of Geographic Names (TGN)[32]，以及英國 Ordnance Survey 的 Open Names Linked Data [33]成為我們首要的觀察對象。目前研究方法主要以文獻探討，並實際查詢各知識庫所（經由其 SPARQL 查詢端點），查證我們想要比對的資料。Getty TGN 為文化資產專業領域知識庫的連帶產品，本節僅引用文獻比較其與其他地名知識庫的差異，具體介紹則於下一節介紹 Getty 語彙時詳述。

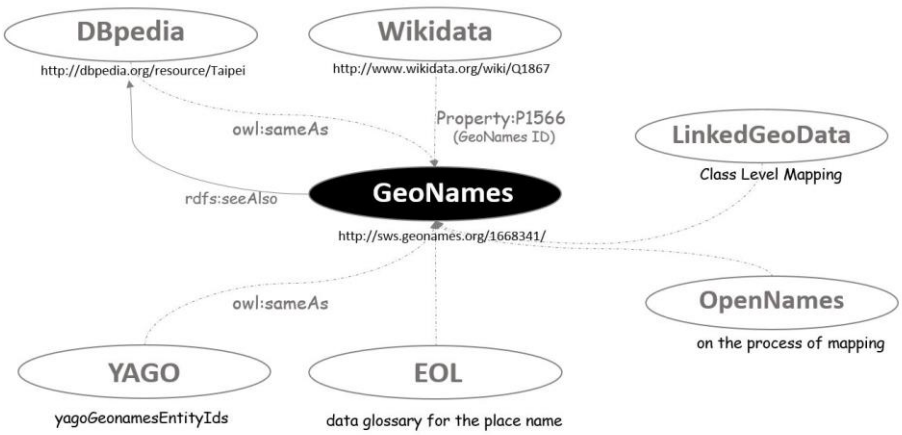

圖3: GeoNames 與知識庫互連情形

從「開放資料連結」一項統計資料研究[34]，不同領域間最佳「開放資料連結」實作分析結果中得知，GeoNames 是第二大連結對象 (Schmachtenberg, Bizer & Paulheim, 2014)。換言之，我們可以合理假設大多數「開放資料連結」發佈時，均需地名資料庫作為空間參照對象，而多數的連結均選擇 GeoNames。反之，地名資料庫也需「開放資料連結」與語意網技術正規化地理相關資訊的語意關係、以及連結外部知識庫資料（地理相關或非地理相關），進而互補單一知識庫的不足(圖3)。例如 2014年統計顯示，約 95% 的 GeoNames 資料與 DBpedia Place 的資料並未重複、33%的 DBpedia Place 資料與 GeoNames 資料無雷同之處 (Moura & Davis Jr, 2014); 分析資料特性差異的研究指出 GeoNames 以地理特徵為主軸、DBpediaPlace 主要描述都市區域資訊、Getty TGN 則偏重歷史文化相關的地名資訊 (Zhu, Hu, Janowicz & McKenzie, 2016)。

地名知識庫的連結選擇，也必須考量資料集的區域特性。若資料集所描述的資源多屬當地資源，該地區所屬的官方權威檔知識庫佔優勢的可能性較大。以英國為例，Ordnance Survey 是英國政府官方製圖單位，2010年4月開始開放地理資料[35]同年10月以「開放資料連結」釋出資料集[36]，藉自建地理行政區域本體論 (Goodwin, Dolbear & Hart, 2008)，陸續發展郵政編碼知識本體、幾何關係知識本體等[37]。2015年釋出基於空間關係本體所建置的「開放地名連結資料」 (Open Names Linked Data)，2016年8月約釋出約92萬筆開放地名所產生的二千多萬筆地名三元組[3]。地理學家的近期研究分析指出 (De Sabbata & Acheson, 2016)：比較 GeoNames, Getty TGN 以及 Ordnance Survey 的 Open Names 的結果顯示，由於英國地名在 GeoNames 與 GTN 中相對於其他國家地名資料量而言，皆屬於高密度資料，因此以英國地區性地理資料的對比代表性，應可適用其

他地區。然而比較三個知識庫後，不論是地名數量、空間分佈、或是建置地名資訊創造者的不同釋義，Open Names 均呈現出較佳的表現。換言之，有地緣關係的地理知識庫理論上是有地區特性資料集考慮進行連結的較佳選項。然而以本研究的實作為例，初期雖考量選擇台灣地名資料庫，但本地地名知識庫品質、已否發佈為「開放資料連結」等限制，均是作為連結首選的阻礙因素。

OpenStreetMap[38]的特點是使用者數量與資料精細度等方面均較 GeoNames 優勢。且 LinkedGeoData 對空間資訊特徵如道路、結構關係、地貌等已進行連結資料的建構，對外連結 GeoNames, DBpediaPlace 以及聯合國農糧署[39]等知識庫，並已發佈「開放資料連結」的空間服務 (Stadler, Lehmann, Höffner & Auer, 2012)。本研究的實作中使用的資料集，其空間部分處理的資訊較為單純 (地名萃取自藏品項目後設資料中都柏林核心集的 coverage 欄位)，因此雖在地圖視覺界面運用中我們選擇使用 OpenStreetMap，但現階段僅實作連結 GeoNames。完整嚴謹的學術比較各地名「開放資料連結」 知識庫的研究，目前仍是開放的研究議題。

以開放資料連結的專業領域知識庫：文化資產與生物資訊專案簡介

基於本研究實作所用資料集多為文化典藏品資料，且台灣推動中文藝術與建築索引典 (AAT-Taiwan)[40]多年有成，本節亦嘗試解讀美國文化藝術專業機構 Getty 所發佈的索引典的「開放資料連結」現況。藝術與建築索引典 (Art & Architecture Thesaurus, AAT) 主要針對文化資產物件提供詞彙、常用概念以及相關資訊；地名索引典 (Getty Thesaurus of Geographic Names, TGN) 則提供居住地、地理特徵、考古區域的地名與相關資訊；藝術家名稱聯合列表 (Union List of Artist Names, ULAN) 則提供藝術家和其他文化相關代理者的結構化人名與傳記類型資訊。

Getty 語彙自1980年代開始至2014-2015年陸續釋出開放連結資料，一方面該項工作保持其索引典在語彙結構上具有每一筆紀錄均有唯一概念的特性，以及概念間完整的相似、上下位、附屬等關係的語意架構，同時每個詞彙來源皆依照文獻保證原則以確保品質，而目前該三語彙之間的整合，也透過「開放資料連結」的方法進行比對連結，如 TGN 中的地方類型與 ULAN 中藝術家的國籍資訊的整合，近期也與 Europeana 在「開放資料連結」上合作 (Baca & Gill, 2015)。另一方面 Getty 索引典的「開放資料連結」化，也化解控制語彙被批評為一過時的知識與經驗產品、或被質疑無法順應網路時代資訊可取得性等問題。更進一步分析，語彙索引典的語意結構以開放連結方式之後，或可提供網際網路時代處理巨量資料中許多統計方法無法解決的問題，也因此「開放資料連結」 方法成為索引典資源永續再生的契機 (Clarke, 2016)。

另外，基於本次研究實作採用的資料有超過三分之一為生物主題，我們選擇了「生物大百科」(Encyclopedia of Life, EOL)[41]作為生物主題的連結目標。這裡也討論 EOL 近期所發展的開放連結物種特徵知識庫 TraitBank[42]，以及其強化 EOL 語意連結的能力。

不一致性(inconsistency)一直是資料品質與整合的巨大難題，然而資料語意的不一致性卻也可能是追求語意豐富的新契機。以 EOL 為例，全球生物物種的分類學系統因年代、地域、學派不同等因素，各有其獨立架構。基於不同觀點的解釋，EOL 允許多重分類，單一物種可被不同命名與分類系統所定義 (Parr, et.al. 2014])。例如，台灣一葉蘭在 EOL 網頁收錄了17種不同的分類架構[43]，而從其獨特的分佈地域角度觀察[44]，即使是台灣本地分類也因資料策展原始單位的不同而有差異[45]。而此台灣特有物種的當地分類架構目前尚未包括在 EOL，對於國際生物物種的分類解釋層面而言，透過「開放資料連結」方法，是否也能成為填補全球生物知識的缺口，亦是值得我們繼續觀察的重點。

實際探討 EOL 在發展「開放資料連結」層面上可知，允許多重分類架構同時存在，成為發展開放式連結資料的重要環節。2014年 EOL 開始建立物種特徵資料庫(TraitBank)，在物種語意分類架構上，不是設計完整的語意架構整合資料，而是在物種語意分類架構延續多重分類方法。EOL 一方面採用許多不同生物知識本體以解釋單一複雜物種特徵，一方面也因現有國際語彙的不足而適時新增自訂語彙。EOL 也因此預期朝此趨勢發展，不僅增加生物領域的知識本體更廣泛與深入的研究應用，同時也能互補特定物種特徵資料庫分類與屬性資料的不足，並進一步增加新資料類型 [46] 與促進跨領域知識的整合 (Parr, et.al. 2016) 。換言之，與其長期陷於眾多語彙無法取得共識的困境，對於資料的語意與連結，設計允許百家爭鳴的機制反而是最符合現實與應用的需求。而此機制也將反映到本研究允許多重語意精煉版本同時存在的設計理念。

後設資料與資料溯源的資料品質議題

品質雖是所有人對資料的基本要求，事實卻是我們很難駁斥 Van Hooland & Verborgh (2014)「沒有完全乾淨的後設資料」[47] 這項論點。我們的實作案例的來源資料集，理論上雖已採用都柏林核心集十五項欄位，以統一整合異質資料來源[48]，並以 XML 結構化格式為內部資料儲存，但亦是無法完全避免標題亂碼、欄位空值、屬性值矛盾等錯誤[49]。後設資料品質議題不僅牽涉到資料建置時期的時空背景，如早期資訊技術處理資料與語意的限制、不同時期對資料要求與需求不同等因素，同時也會根據不同使用者情境，對資料品質有不同解讀。Yasser (2011) 指出，最常見的後設資料品質問題，包括不正確的資料屬性、屬性值以及系統功能性所造成的資訊遺漏，或因資料比對所造成語意資訊的喪失，以及資料再現格式不一致等。

若更進一步分析，不同領域研究者對資料品質研究的定義，以及其指標檢驗項目亦是各自不同。例如表2 所整理的資訊、資料、後設資料、以及連結資料品質的不同觀察面向[50]。資訊管理學者的資訊物件包含資料內容與後設資料，並視後設資料為提供資訊物件的程序工具(Stvilia, Gasser, Twidale & Smith, 2007)；在廣義資料品質方面則注重方法論、資料種類、以及系統面等因素，也因此其所需考量的角度 (28 種面相) 也更為廣泛 (Batini, Cappiello, Francalanci & Maurino, 2009)。

表2： 資料品質檢驗的不同面向

Information Quality	Data Quality	Metadata Quality	Linked Data Quality
Stvilia et al.(2007): 22 dimensions	Batini et al. (2009): 28 dimensions	Tani et al. (2013): 10 parameters	Zaveri et al. (2016): 18 dimensions
Naturalness (I)			Interoperability (RP)
Accessibility (R)	Accessibility	Accessibility	Availability (A)
Accuracy (R)	Accuracy	Accuracy (S)	Semantic Accuracy (I)
Accuracy/Validity (I)	Applicability	Pertinence	Syntactic Validity (I)
	Appropriate amount of		
Complexity (R)	Clarity		
Precision/Completeness(R)	Completeness	Completeness(S)	Completeness (I)
Informativeness/Redundancy(R)	Comprehensiveness		Understandability (C)
Informativeness/Redundancy(I)	Conciseness		Conciseness (I)
Structural Consistency (I)	Consistency	Similarity	Consistency (I)
	Convenience		
Structural Consistency(R)	Correctness		
Verifiability (R)	Credibility		Trustworthiness (C)
Currency (I)	Currency		
Semantic Consistency(I)	Derivation Integrity		
	Ease of operation		
Naturalness (R)	Interactivity	Conformance(S)	Interlinking (A)

Semantic Consistency(R)	Interpretability		Interpretability (RP)
Precision/Completeness(I)	Maintainability	Preservability	
Complexity(I)	Objectivity		
Relevance/ Aboutness(R)	Relevancy	Relevance	Relevancy (C)
Authority (Reputational)	Reputation		
Security(R)	Security		Security (A)
	Speed		Performance (A)
	Timeliness	Timeliness	Timeliness (C)
	Traceability		RP Conciseness (RP)
Cohesiveness (I)	Uniqueness	Significance	
	Usability		Licensing (A)
Volatility(R)	Volatility		
			Versatility (RP)
(I): Intrinsic; (R): Relational; (S): Metadata Spec.; (RP): Representational; (A):Accessibility; (C): Contextual			

另外，圖書館界對於後設資料品質的討論包括書目資料簡易儲存、目錄資料的元素設計、以至遠端整合異質資料庫的資料脈絡重要性等，關於後設資料品質的判定，亦是眾說紛紜，有學者嘗試根據數位圖書館品質架構 (Digital Library Quality Frameworks) 歸納出適合後設資料語意以及數位物件的十個資料品質參數 (Tani, Candela & Castelli, 2013)。 相對之下，「開放資料連結」的資料品質則是新興議題，Zaveri 等人(2016) 提出語意再現與連結是連結資料品質的基本要素，其中包括互通、語意正確、互連、可被解釋、再現的簡明、以及資料的多功能性等，均是連結資料品質所著重的面向。同時為達「開放資料連結」 再次使用目的，授權資訊明確與否亦成為資料品質指標判別要素，前人這項研究所歸納18個連結資料品質參考面向，亦成為目前 W3C 資料品質語彙 (Data Quality Vocabulary) 比對 ISO/IEC 25012 資料品質模式的主要對象[51]。

即使是國際通用後設資料語彙如都柏林核心集，在語意層面也受到語意概念模糊，如 source 與 coverage(包含時間資訊)定義易混淆、欄位語意重疊(semantic overlaps)如 creator, contributor, publisher 間定義可相互套用、或不同單位對 relation 欄位解釋不同，因而使用方式呈現高度差異等 (Park & Childress, 2009)。在此前提下，若以柏林核心集 15 項欄位為資料模型所產生的後設資料，在 Chuttur(2014) 的實證研究下指出，資料品質零錯誤的可能性為微乎其微。換言之，以上所探討品質定義多樣化、資料生成時空脈絡迴異、國際通用語彙具先天缺陷等問題，都密切攸關本研究實作上資料語意再現的一大目標：結構性資料的再次使用以及永續價值。而對於資料品質與價值之間的衝突是否能藉由「開放資料連結」方法，轉化為和諧並存，也因實作將典藏品目錄以「開放資料連結」方式再現的過程，促使我們探究後設資料的資料溯源議題。

在資料價值、常人信賴、以及機器自動處理的多方期待，以及重視資料品質[52]的時代中，「資料溯源」顯然是資料品質和資料再次使用的通關護照。歐洲最古老圖書館之一的牛津博德利圖書館的積極採用可為例證 (Burgess, 2016)。資料溯源也已成為數位策展、資料引用的必要環節(Poole, A. H. (2016)。這同時，地理學家也擔憂若缺乏資料溯源，機器所提供的知識將會降低常人解讀地方資訊的能力 (Ford & Graham, 2016) 。若再就本研究個案為例（圖4），若想達成資料再次使用以及開放連結的目標 (Context III)，首先我們必須確認使用的資料來源 (Context II)，因此需了解此資料的原始資料 (context I) 。再次使用而產生新資料時 (Context III) ，為確保此新資料能再次被他人使用，我們也必須提供他人確認我們新資料來源的資訊、以及轉換資料的過程 (Context I + II) 。追溯資料的歷史與脈絡的資訊即是所謂的資料溯源 (Carata, 2014)。

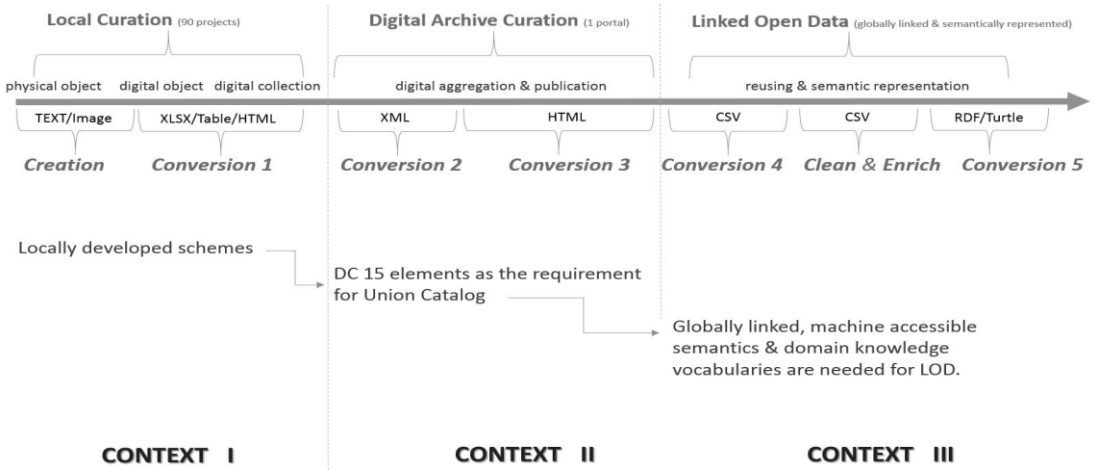

圖4: 結構資料在不同情境下的再次使用與資料轉換過程

實際進入語意、連結、資料溯源的作法，將需要考量包括資料與工作流程二種形式的資料溯源 (data and workflow provenance)、資料溯源的資料模型設計、以及資料溯源的儲存與再現 (Storage and Representation) (Omitola, Gibbins, & Shadbolt, 2010)。在資料溯源的資料模型設計方面，我們體認本研究案例的資料來源脈絡，將與其數位資源再次使用的效果相關。這裡我們使用之前建立的「再次使用關聯性知識本體」(Relations for Reusing Ontology, R4R) 為基礎 (Huang and Chuang, 2014)：資料溯源資訊和要被再次使用的物件以同時打包的方式，一起提供常人與機器，以面對該數位物件被再次使用的不同情境。例如台灣一葉蘭 (代碼 data:d2148340)[53] 藉由 *r4r:hasProvenance* 將資料溯源以 W3C 資料溯源本體論 (Prov-O) [54]描述，指出該數位資源分別在 1993, 2011, 2016 三個時間點，被不同地方不同單位進行了再次使用、格式轉換、以及語意連結的工作，因此也提供使用者資訊來源佐證 (圖5、圖6)[55]。

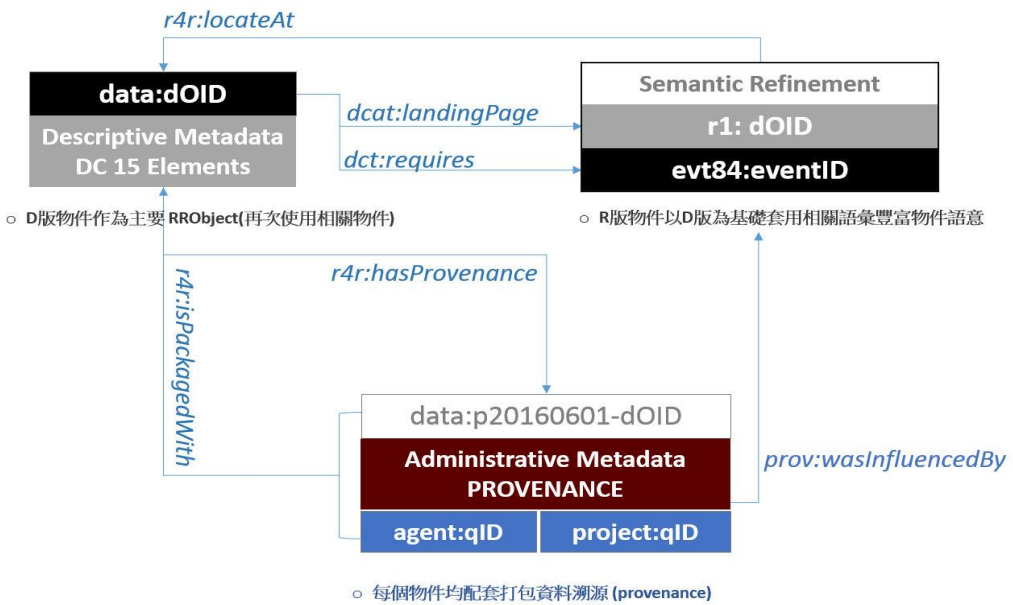

圖5: 資料模型包含 D 版 R 版資料與相對應的資料溯源

圖6:台灣一葉蘭 (data:d2148340)之資料溯源

實作案例: data.odw.tw

我們之所以回顧近期「開放資料連結」相關計畫與技術發展、探究全球開放式連結資料知識庫的差異與品質、也分析資料品質與資料溯源的關係，主要目的是一方面整理提供「開放資料連結」國際發展趨勢，便利進一步的觀察與思考，另一方面也釐清我們目前實作案例 data.odw.tw 中許多設計原則的選擇因素與脈絡。以下將透過實作的五大步驟：(1) 探究共享脈絡下的資料再次使用的關聯性 (2) 設計不同情境下的模式系統架構 (3) 資料剖析、清理與比對 (4) 以使用者為核心的資料庫知識平台技術架構 (5) 透過知識本體以協助理解資料語意的再現與再次使用。我們期望這樣的實作案例分享，對目前正投入或預備投入「開放資料連結」 的研究者與實作者，將有所助益，對於尚未接觸「開放資料連結」者，也能因此思考將已有的資料集附予新價值，讓單一或片段知識透過語意連結，成為全球知識網的連結點。

步驟一: 探究共享脈絡下的資料再次使用的關聯性

以台灣一葉蘭的標本典藏品為例，該品項歷經了不同機構的資料策展、資料發佈、資料再次使用(curation, publication, reusing)三個動態脈絡，以及相對應的三個描述資料語意的階段：資料再現、資料保存、資料釋義(Representation, Preservation¸ Interpretation)，在此不同的共享脈絡之下，「再次使用的關聯性」，亦即再次使用之關聯性本體論 R4R (Relations for Reusing)[56]是我們採用的理論基礎(圖7)。

R4R 是一個簡易知識本體，以描述資源發佈 (Publication) 和再次使用 (Reusing) 的一般性關係。R4R 由兩個分立概念 RRObject 和 RRPolicy 組成。其中 RRObject 包含可分別獨立或可關聯的三組件：文件(Article)、資料(Data)、軟體碼(Code)；RRPolicy 包含可分別獨立或可關聯的二組件：資料溯源資訊 (Provenance) 和授權資訊 (License)。關係陳述主要由 *r4r:isPackagedWith* 和 *r4r:isCitedBy* 兩個基本關係以定義。前者借由資料套裝資料溯源與授權資訊，進行宣告資源是處於可再次使用的狀態。後者則對資源間的引用關係做描述。

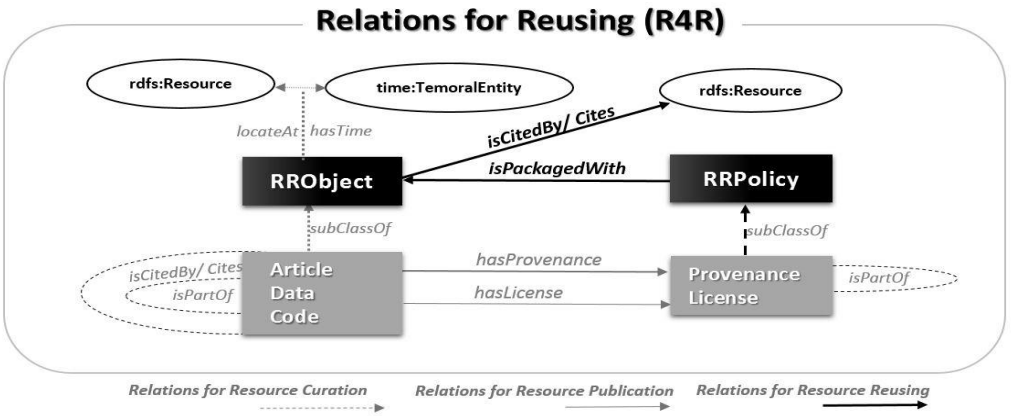

圖7: 再次使用之關聯性本體(R4R Ontology)

步驟二：設計不同情境下的模式架構

我們以資料策展者角度出發，學習探索以關聯式資料庫、及開放檔案格式與開放程式碼二種不同情境的思考架構，探究發佈資料連結不同模式的運作，茲分別敘述如下:

模式一：使用關聯式資料庫進行「開放資料連結」的發佈

研究初期我們嘗試使用關聯式資料庫工具 D2RQ 發以佈連結資料[57]，並試作數位典藏索引典與 AAT 語意描述典藏品的連結資料，結合 W3C 資料溯源本體論，描述資料重整活動並設計 dat ontology[58]，提供機器可操作資料格式，並測試 SPARQL 語意查詢能力，如回答以下這類問題：銅琺瑯方瓶有哪些語意概念？概念侈口 (器口向外張) 描述了哪些器物？器物 A 和器物 B 有哪些相似的特質？

模式二：開放式連結資料模式的系統架構

對於本研究而言「開放資料連結」最具魅力的核心觀念是資料藉由開放與連結的方法，即使是其最小單位，以三元組方式呈現的單一敘述，只要能自由被常人和機器理解與操作，都是資料永續使用。也因此我們採用以開放檔案格式以及開放程式碼為基礎的資料釋出策略，將可開放的資料，同時整理為批次大量下載的結構資料檔案，如 CSV 格式檔案。並以此為基礎，使用開放原始碼程式工具，進行資料整理、清理、發佈等各階段工作。資料連結的發佈，亦採用開放檔案格式如 JSON, Turtle, XML 等供常人與機器下載使用，並同時提供常人使用的網站瀏覽介面，以及機器介接資料使用的 SPARQL 端點(圖 8)。

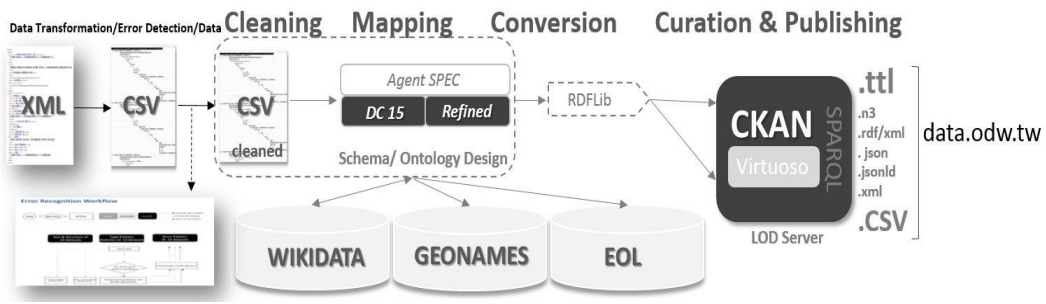

圖8: 模式二開放式連結資料的模式系統架構

步驟三: 資料剖析、清理與比對

資料格式轉檔、清整與除錯

依模式二架構，我們對84萬筆以創用 CC 授權的藏品目錄資料，進行由資料庫匯出為 XML 文件，再轉換為 CSV 格式資料表單。採用 CSV 為中介表單格式的優點包括：可參照其他資源表單、表單可人工勘誤、表單增修歷程可管理、軟體工具多、資料連結的產出方式有彈性等。過程中我們遇到由 XML 至 CSV 資料轉換可能遇到的問題，如 XML 樹狀結不易轉換 CSV 扁平結構、資訊遺失、或過多欄位等，因此測試多種版本後，目前採用的 CSV 版本為類似 XML 結構的格式如圖9 所示。

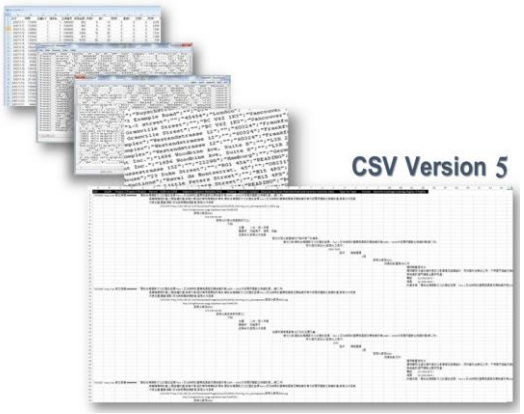

圖9: CSV 轉置第五版

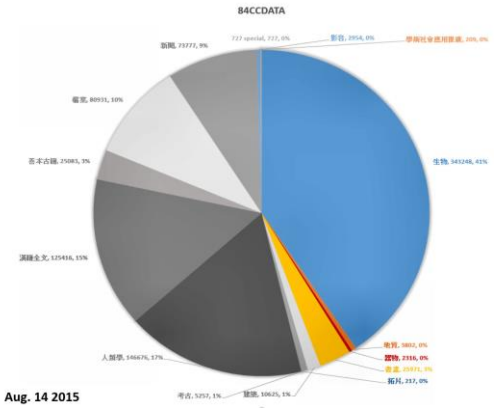

圖10: 84萬筆資料剖析

另外我們也分析這批資料的特性如圖10所示。其包括14的主題，內容中生物佔全部資料41%，其次為人類學及漢籍全文。而這也是最初選擇外部專業知識庫 EOL 及 AAT 的背景因素。誠如前面所討論的資料品質問題，當資料已轉換為 CSV 後再進行設計程式檢驗歸納資料錯誤模式[59]，並產出錯誤藏品清單則相對容易。在此過程中我們發現資料具有都柏林核心集定義混淆、名稱模糊、編碼不一致 (如時間欄位中時間表式法的不一致)、語意超載 (如 Subject 中包括 creator, contributor 欄為值[60]) 、資料重複、或來自資料輸入程式錯誤等問題[61]。然而，資料品質牽涉廣泛，況且本研究成員並非原始資料創造者，許多資料脈絡無法取得與判別，或若因專業知識不足亦可能導至除錯反錯結果（如生物命名規則中，斜體表示、加底線、問號等是允許的），若不慎亦可能在資料清理過程中，將正確資料視為錯誤或亂碼而過度清整。當然，無可避免的問題是，現有資源如時間人力經費等因素是否能支持資料清整的考量？亦可能成為促進永續資料再次使用的障礙。然而，若要達成解決前述資料品質與價值的衝突，除運用後設資料溯源外，如何才能同時保持資料品質、且減少資料清整的替代方法成為我們的新挑戰。

首先，一般對無效連結 (broken/dead links) 的看法是錯誤連結或連結資源消失，因此若非更正連結資訊，就是清除連結資料。在「開放資料連結」脈絡下，無效連結似乎更是必要的清整對象。然而美國國會圖書館的 Susan Manus 卻認為保存無效連結有其正當性 [62]：一方面無效連結可協助搜尋推定原始典藏品不同版本的位置，另一方面無效連結的網址(URL)本身即是一種描述資源的後設資料。網址傳達的訊息包括網站結構，特定資源發佈日期、文檔標題、作者、描述性關鍵字等資訊。即使主機僅為 IP 地址 URL，亦可能表示託管該域的地理區域設置。無效連結的網址包含如此豐富的語意資訊，因此成為我們在「資料溯源」 設計原則中保留所有無效連結的主要

理由。但如何以豐富語意的陳述來描述無效連結，使機器互通資料時不會回傳錯誤(404 訊息)，或是在「開放資料連結」群體中在品質檢視時被視為錯誤，仍是我們需要研究的課題。

其次，開放資料具有協助改善資料品質如資料永續保存、增進外部驗證資料機會等益處 (Janssen, Charalabidis & Zuiderwijk, 2012)，因此我們採取保留原資料 CSV，以此基礎在 data.odw.tw 已連結方式發佈原始（亦稱 D 版）資料集，以不更動原始資料為原則，僅增加資料溯源資訊，讓任何使用者欲再次使用該資源時，可根據使用者認定的資料品質定義與應用的需求，自行進行資料清理。或如同我們前面所提，經資料清理之後的 R 版資料集，允許多重分類架構原則，在此資料修正的脈絡下，R 版的另一功能為提供多重修正版本語意陳述，設計使資料清理版本也可因不同清理時間、方法、或對品質解釋不同而提供不同資料清理版本。

資料語意比對：

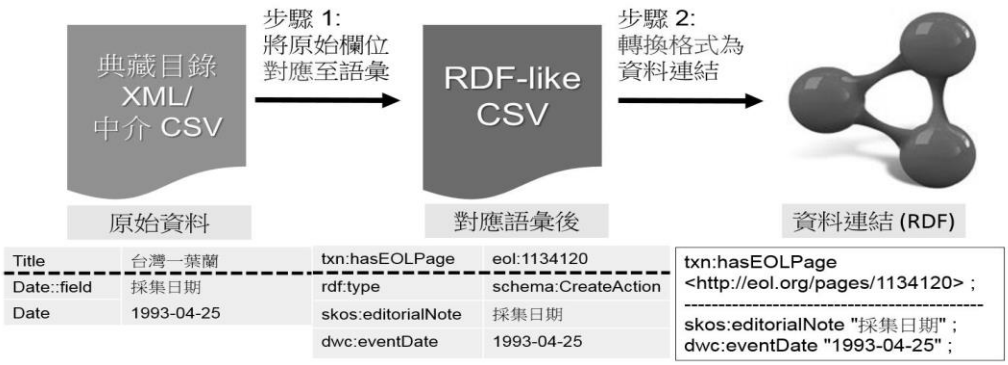

圖11:: 語彙對應與格式轉換

資料品質的改善亦可借由「開放資料連結」方法達成，例如資料的語意定義使用語意網資料作為可信賴的參照對象、連結外部知識庫與協同合作再現語意、根據要求可自動驗證資料衝突、或以知識本體整合資料內容等 (Fürber & Hepp, 2013)。關鍵在於語彙的運用以及資料語意的比對。實作中我們主要進行的工作包括時間資料正規化 (如 date 欄位值若為西元時間以 ISO8601為標準、民國年或朝代則對應到 Wikidata)、生物主題資料連結到 EOL 等。語彙對應與格式轉換包含兩步驟流程如圖11所示，由於以設定檔 (profile) 定義對應關係，因此更換語彙時僅需調整對應的設定檔。從空間資料 (coverage) 欄位值抽取到的地名資訊則對應到 GeoNames， 使其原始資料的空間資訊理解度與查詢度大幅提升(如可回答：採集於大同鄉的台灣一葉蘭標本物件有那些?)[3]，同時也完成典藏機構與計劃名稱在 Wikidata 的系統代號建置與比對。完成 CSV 比對後，利用 Python 程式語言搭配函式庫 RDFLib 進行 Turtle 資料格式的檔案轉換。

步驟四: 以使用者為核心的資料庫知識平台技術架構

CKAN 資料平台軟體：

本研究使用開放原始原始碼套件 CKAN 建立 data.odw.tw 網站，以資料連結形式儲存與呈現典藏品資訊。CKAN 是目前開發最活躍、使用組織最多的開放原始碼資料平台軟體，包括英美澳洲及我國多個地方政府均以其作為開放資料平台之基礎。據官方網站於2016年9月統計[63]，全球已有超過140個政府機構、社群或學術單位使用 CKAN 建置資料平台。CKAN 由開放知識基金會（Open Knowledge Foundation, OKF）於2005年發展，目前由 OKF 成立之 CKAN Association 維護，透過 GNU AGPL 3.0授權條款釋出程式原始碼，本研究使用其最新版本2.5.2。

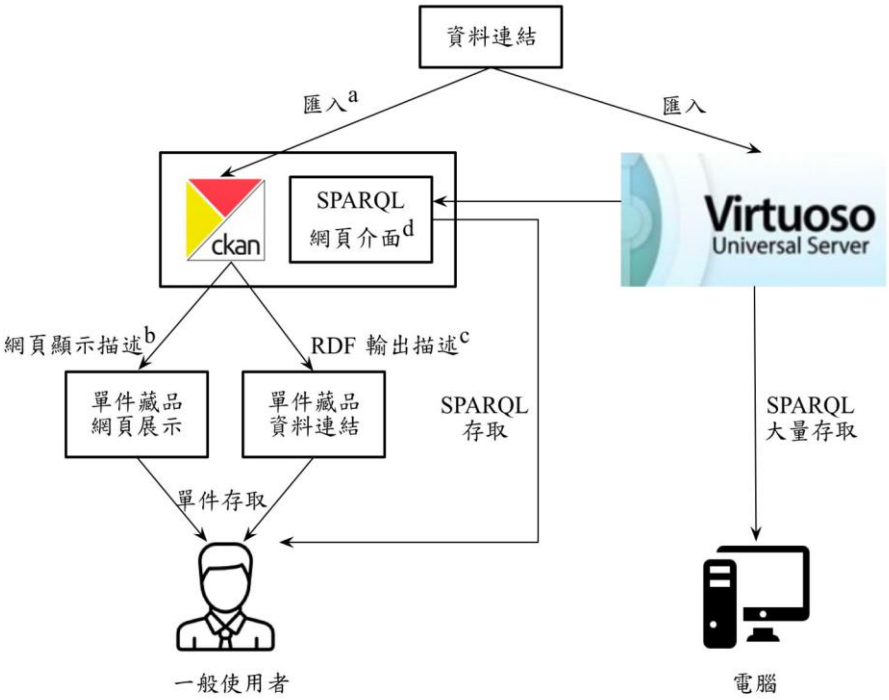

圖12 CKAN「開放資料連結」系統架構[64]

因其開放原始碼之特性，機構可自行建置（self-hosted）系統提供服務，同時可避免被特定專有軟體（proprietary software）套牢（lock-in）。在功能方面，除基本資料發佈與存取外，CKAN 亦支援資料應用程式介面（Application Programming Interface, API）、搜尋與條件篩選器、標籤、版本控制、分享與權限控制等功能，而可直接將資料以圖表形式呈現之「資料視覺化」功能更是其一大特色。以 Pylons 網頁開發框架寫成的 CKAN 具有現代網頁應用程式架構與極佳的自訂彈性，而 CKAN 更有為數眾多的擴充套件（extension），**提供包含自訂後設資料**、自動生成數位物件識別碼（Digital Object Identifier, DOI）、通用大量資料採集（harvesting）介面及資料連結輸出等研究資料管理所需之各項功能。

CKAN 資料連結支援

CKAN 在2010年發行的初期版本 [65] 即具有將資料集（及所包含之資料）之後設資料發佈為資料連結之功能（支援 RDF/XML 與 Notation 3格式）。而為進一步完善資料連結功能，OKF 於2013年啟動 ckanext-dcat 擴充套件[66]的開發，不僅提供更多資料連結格式（RDF/XML、Notation 3、Turtle 與 JSON-LD）輸出，更對應 CKAN 的資料採集介面以支援大量資料輸入。使用者除可自瀏覽器瀏覽以網頁形式呈現之藏品資料連結外，亦可於網址後方加上對應格式之副檔名[67]，即可取得該筆藏品之資料連結，相當方便。

另值得一提的是，雖由 ckanext-dcat 名稱可知該套件對應語彙以 W3C 制定之 DCAT（Data Catalog Vocabulary）為主，但由於採用標準資料採集介面，故保留了擴展的彈性。本研究即在此基礎之上，透過實作位於 ckanext.dcat.harvesters.rdf 之 DCATRDFHarvester 類別，及位於 ckanext.dcat.profiles 之 RDFProfile 類別，開發自訂資料採集介面，加上些許調整原 ckanext-dcat 套件之採集邏輯後，使其具備匯入以多種語彙描述之典藏品資料連結的能力（支援匯入格式與輸出時相同）。

操作流程與系統架構

本研究建置之開放式資料連結平台架構如圖12所示。轉換為資料連結之藏品描述，會先經由 ckanext-dcat 套件提供之採集介面，以 ckanext-harvest 規範之採集機制匯入至 CKAN 平台後，使用者便可自網頁瀏覽單筆藏品之資料連結，該展示介面係由 ckanext-scheming[68]與 ckanext-repeating[69] 兩套件加以定義。本系統亦結合 ckanext-spatial[70] 與自行開發之 ckanext-tempsearch[71]套件以分別支援空間、空間搜尋，並可結合 CKAN 既有的過濾條件與關鍵字搜尋功能進行整合查詢。而單筆藏品之資料連結則是透過 ckanext-dcat 的輸出描述，再由 Python 函式庫 rdflib[72]輸出為資料連結。

另一方面，為顧及供機器操作之 SPARQL 語意查詢功能，本系統同時將資料連結檔案匯入 OpenLink Virtuoso Open-Source Edition[73]（版本07.20.3217），並整合網頁查詢介面[74]於 CKAN 平台供使用者進行 SPARQL 查詢測試。相關程式碼均以 MIT 或 GNU AGPL 3.0授權條款釋出，可於 https://gitlab.com/iislod 取得。

系統限制與發展方向

如此搭建之平台功能雖尚屬完整，修改既有程式範圍亦不致過大，但使用較多擴充套件[75]，且新增語彙等操作均須直接修改程式，提升維護難度；未來規劃將部分較常變動之設定獨立為描述檔案，以降低程式複雜度。而一藏品對應產生一 CKAN 資料集的設計，所產生的大量資料集亦對匯入工作造成負擔（於 Intel E5-2620 2.1GHz、16GB 主記憶體伺服器實測，匯入84萬件藏品約需2個月時間），未來將朝改以多筆藏品彙整於一 CKAN 資料集方式匯入。

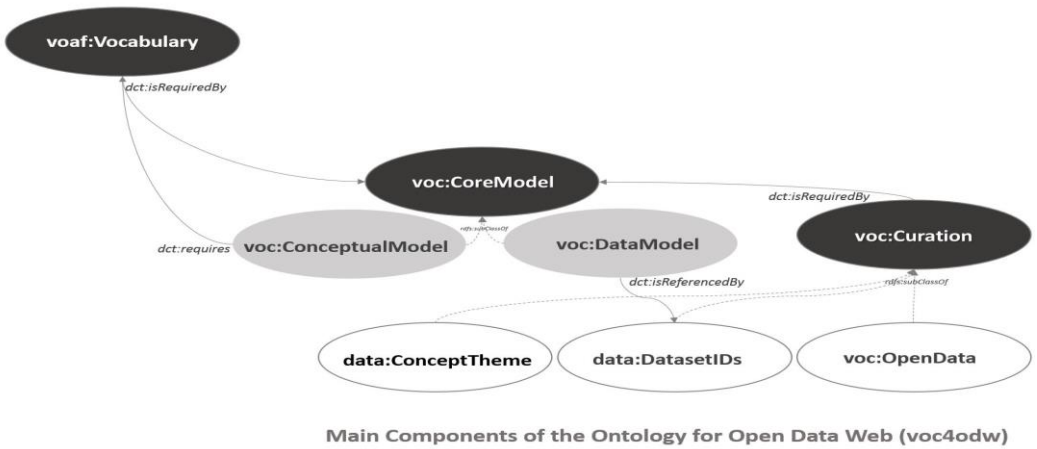

圖13: voc4odw 知識本體論主要架構

步驟五: 透過知識本體以協助理解資料語意的再現與再次使用

我們實作一項「開放資料網知識本體」 (Ontology for Open Data Web, voc4odw) [76]由核心主模型 (Core)、策展 (Curation) 與與國際語彙 (voaf:Vocabulary) 三大模型組成 (圖13)。主模型為該知識本體主要架構，並作為策展與國際語彙間的橋梁。策展模型是目前該知識本體中，連結資料、常人和機器三者溝通的管道。國際語彙則是關聯主模型參照外部常見國際語彙，並提供概念模型引用外部語彙的知識參照。

策展模型的課題主要包括資料識別、分類及資料發佈。如台灣一葉蘭資料識別為 data:d2148340、事件 ID 為 evt84:phyCre-d2148340、策展分類(*dcat:themeTaxonomy*)為 data:Biology。

而為回應全球開放資料與高規格要求資料與過程的可複製性，因此 R4R 設計再次使用機制包括 articles, data 與 code 的打包組合、以及台灣一葉蘭多樣資料發佈格式，如 XML. JSON-LD, Turtle 或 SPARQL 端點等，均由策展模型結合 R4R 描述。

其次，關於主模型中的二大元素：概念模型與資料模型，且讓我們再細看台灣一葉蘭的情境：1993年4月25日(*dwc:eventDate*)，台灣一葉蘭(data:d2148340)此實體物件(dct:PhysicalResource)在一次採集活動(evt84:phyCre-d2148340)中，於地點(*gn:parentCountry*) 台灣(gns:1668284)的(*gn:parentFeature*) 宜蘭 (gns:1674197) 大同 (gns:1667637)，被製成 (*event:product*) 標本(dwc:PreservedSpecimen)。此標本採集活動，若用常用語彙描述，是一個物件被創造的事件(schema:CreateAction); 採集過程若用生物領域語彙 Darwin Core 來描述此脈絡 (voc:Context) 就是常人觀察的活動(dwc:HumanObservation)。

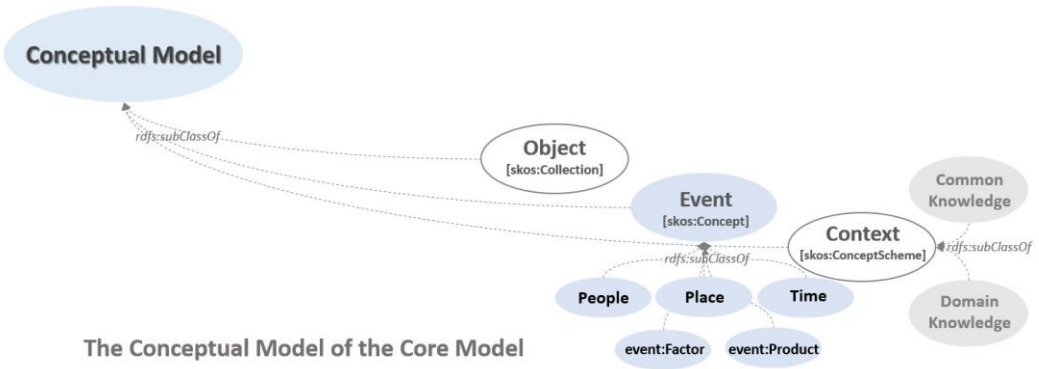

圖14: 概念模型

依附在開放資料網知識本體下，台灣一葉蘭透過概念模型中事件、物件、脈絡三大元素描述資源語意。事件，由人時地三概念組成，而脈絡則由常用語彙如 schema.org 以描述，Darwin Core 則用來描述專業知識。如圖14所示，概念模型包括 SKOS 簡易知識組織系統的概念化模型，並闡釋由專業知識或一般性知識所關連的事件，藉此提供概念成形的架構。

觀察台灣一葉蘭(data:d2148340)的數位化進展，會幫助我們了解資料模型(圖15)。一葉蘭從實體標本物件於(*prov:generatedAtTime*) 2011年5月13日由 (prov:wasGeneratedBy) 中研院台灣本土植物數位化典藏計畫(project:q21095859)被數位化，並在(prov:hadPrimarySource)該原始計畫網站上呈現其後設資料的資訊。這些描述是資料模型主要敘述一葉蘭數位演化的過程，藉由後設資料溯源資訊的追溯，如前圖4所示，台灣一葉蘭在不同階段的脈絡所代表的不同角色如：

- Context I: 原始資料 (prov:PrimarySource);
- Context II: 目錄呈現 (dcat:Catalog, prov:Revision);
- Context III: 開放連結 (D 版為 data:Reused 與 r4r:RRObject ; R 版為 data:Refined 與 r4r:Data)

在資料模型中(圖15)，除資料溯源外另一重要模型的機制設計是衍生資料的兩個子類別設計：都柏林核心集描述版本 D 版的 data:Reused、以及強化語意 R 版的 data:Refined。

首先，D 版的 data:Reused 運用 R4R 語意描述模組化機制，提供基礎的都柏林後設資料15欄位描述。在此所描述的資料為自原始資料中(voc:PrimaryData)所抽取的衍生資料

(voc:DerivationData)。抽取資料的目的是再次使用該資源，因此宣告為 data:Reused，同時為此物件在 data,odw.tw 中賦予唯一 URI 而定義為 r4r:RRObject (再次使用相關物件)。

例如台灣一葉蘭在 D 版中 rdf:type 為 data:Reused 與 r4r:RRObject，使用都柏林後設資料的11個欄位，在 RDF 描述架構中 Subject 為此資源的 URI，Property 為都柏林後設資料所對應的 URI，三元組的前二者均為連結，最後 Object 在 D 版則預設為文字。雖然國際「開放資料連結」與 Semantic Web 社群並不鼓勵發佈「開放資料連結」為文字值，但考量保存與策展原始資料最初原型的必要，我們將典藏品按原始資料不添加 Object 語意前提下，只增加該藏品資料溯源資訊，達成提供第三方後續可不被我們語意框架限制而自由再次使用該資源的目的。

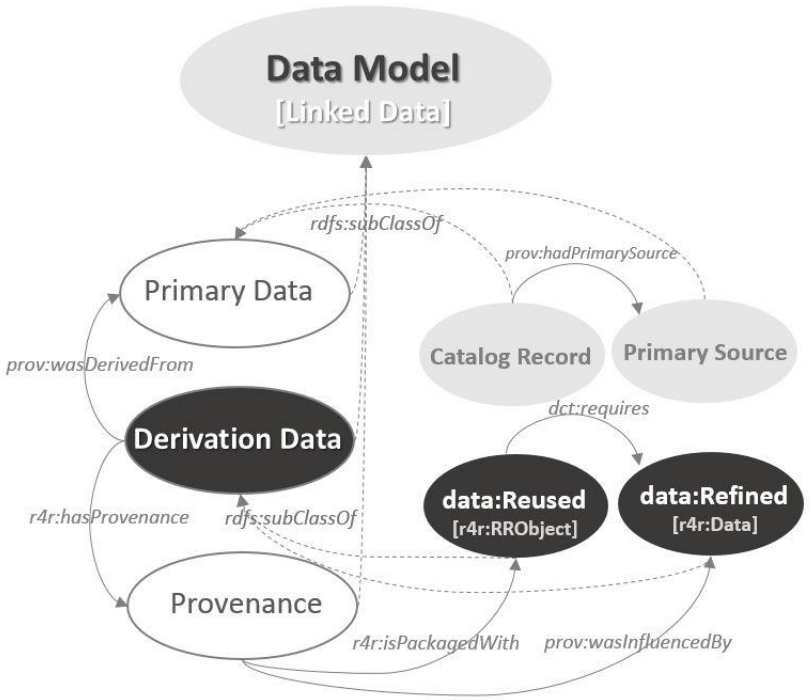

圖15: 資料模型

其次，R 版語意資料首先為自 D 版中所抽取的再次使用資料 (r4r:Data)，因此和 D 版共享同一資源 URI。D 版(data:Reused) 需要(*dct:requires*）R 版（data:Refined）的語意強化與連結，因此 R 版三元組特色為三者主要是 URI 連結、或正規化後的資源如時間。進一步說明 R 版和 D 版在 R4R 的關係可知，二者為 r4r:RRObject 資源唯一 URI 下之 r4r:Data，因此 R 版和 D 版二者並無上下位關係，而這也是現階段設計每一物件藏品的資源唯一 URI 在 CKAN 中宣告為 dcat:Dataset 描述此物件藏品的資料集是集合不同版本、該資源不同檔案格式的各式資料。

例如台灣一葉蘭(data:d2148340)在 D 板中 *coverage* 的欄位值為文字 "行政區: 宜蘭縣大同鄉"在目前 R1版中則借由概念模型的事件描述，以 GeoNames ontology 與地名資源 URI 描述[77]，這二者所包含的三元組各均屬於 data:d2148340資料集，而目前提供的資料集格式則包含 XML, JSON-LD 與 Turtle。換言之，台灣一葉蘭雖在 R1中連結 EOL 豐富語意，但目前因時間人力資源等限制只試作生物類三計畫，且回顧比對工程之時間點 EOL TraitBank 尚未釋出「開放資料連結」 資料，也因此我們僅以連結網頁而非資源 URI 方式處理[78]，後續則可在不同 R 版完成所有 EOL 比對連結，或連結 TraitBank URI，或連結其他知識庫。簡言之，不同比對連結時機、不同語

義解釋架構、或不同推導過程，則可以不同語意化資料版本策展（例如，R1, R2, R3...）。以下歸納 R 版的多重機制主要有三:

多重清理版本機制：清理本身即是一種語意解釋過程，例如前述資料品質問題中提到的標題亂碼案例，在 D 版有一案例為 data:d4653940。我們並未對該標題進行清理或刪除原因有三：(1) 標題應由資料創造者定義，(2) 該資源原始資料[79]顯示此為一植物學遺傳研究頁面，若技術資源足夠時，因我們已提供完整的後設資料溯源，大量自動化訂正錯誤是可能的，(3) 若未來採用開放大眾參與編修，此資源亦可能被修正，因此未來亦可能在不同時間由不同修正者提供不同修正版本。

多重知識連結機制：連結知識庫的目的不在於為特定資源提出完全正確的語意，各專業知識內自有理論解釋，而是要提供檢視資源的多種面向，了解其中有哪些概念能有助於處理當下使用者問題。再以台北的描述為例，TGN 知識樹分類以台北上層為國家／島嶼／特別都市、Wikidata 連結14種不同知識庫的 Taipei 編號、而 DBpedia 則描述台北有34種 rdf:type 知識分類 (圖2)。這種多重知識的觀點，為連結至不同 知識庫的典藏品資源賦予了強化語意知識的基礎。另外也因知識本身會隨著時間演化，今日的事實可能是明日的謬論，對典藏品而言，連結的是固定 URI 而不是知識內容，一當知識庫知識內容更新，典藏品不需更新知識本身，透過資料連結，知識重新發現重新表達。

多重語意架構機制：對語意架構、語彙選擇、漸進式增加資料的語意深度等不同的需求劃分開來，並且各自獨立，即是 R 版多重語意架構機制主要目的。舉例而言，R1 目前只針對生物類資源使用 dwc 語彙，然而描述物種語彙至少有30種[80]，R1只處理都柏林核心 15 欄位中時空資訊以及部分生物類標題，因此類似生物類中 subject 所描述「界門綱目科屬種」等知識架構語意，如前章所述，多重 R 版機制的設計使其更具彈性。另外，目前84萬筆資料由14個內容主題構成，後續若時機許可亦可套用不同專業領域的語彙發佈不同R版，同時若有其他非典藏目錄新資料集加入，亦可據資料特性調整，套用現有 D 版架構、根據使用者需求調整 R 版，同時發佈該資料集之 R 版。此多 R 機制允許互相獨立，甚至是互不相容的語意需求，顯示了這種資料連結新式架構，最適策展者與使用者雙方需求最終可調和的本質。

國際語彙的運用 (voaf:Vocabulary)

鑑於 Schaible 等人 (2014) 實證研究指出，目前語彙再次使用策略有六： (1) 再次使用常用語彙，(2) 自訂語彙並使其與外部常用語彙連結，(3) 語彙再次使用最大化，(4) 語彙再次使用最小化，(5) 語彙個別概念再次使用最小化，(6) 僅再次使用專業領域語彙。同時並建議與其限制語彙數目，不如採取再次使用共用語彙的策略。這與 Srinivasan 等人 (2010) 提出的「多重知識本體」(multiple ontologies) 觀點相互輝映。該觀點認為對於一物件的不同了解和詮釋間的張力是應該被接受的[81]，因此我們使用 voaf:Vocabulary (在資料連結雲中所使用的語彙) 作為主要類別，借此關聯主模型 (Core Model) 至外部常見國際語彙(圖13)。

可再次使用的語彙資源可透過 Linked Open Vocabulary (LOV)[82]了解與選擇目前國際語彙的使用。此資源網站目前並未囊括已「開放資料連結」 發佈所有資料集中所有使用語彙以及索引典資源。然而對使用者而言，LOV 可依據語彙作者或單位、語彙名稱、語彙單詞如 class 或 property 名稱、專業知識分類查詢、SPARQL 查詢、以及蒐錄語彙時必須通過機器與 LOV 專業人員的審查等優點 (Vandenbussche, Atemezing, Poveda-Villalón & Vatant, 2015)。本研究設計 voc4odw 時，得利於使用此國際開放語彙服務，進而採用25個國際語彙如表3所示: 包含 W3C 標準語彙如:

csvw, dcat, org, prov, skos, time; 一般常用語彙如: cc, dc, dct, event, foaf, r4r, schema.org; 以及專業知識語彙如: aat, dwc, geo, gn, txn。

表3: voc4odw 知識本體命名空間

Prefix	Namespace	Description
Common Knowledge		
cc	http://creativecommons.org/ns#	1. Creative Commons Rights Expression Language
csvw	http://www.w3.org/ns/csvw#	2. W3C CSVW Namespace Vocabulary Terms
dc	http://purl.org/dc/elements/1.1/	3. DC 15 (Dublin Core Metadata Element Set)
dcat	http://www.w3.org/ns/dcat#	4. W3C Data Catalog Vocabulary
dct	http://purl.org/dc/terms/	5. DCMI Metadata Terms
dctype	http://purl.org/dc/dcmitype/	6. DCMI Type Vocabulary
event	http://purl.org/NET/c4dm/event.owl#	7. Event Ontology
foaf	http://xmlns.com/foaf/0.1/	8. FOAF Vocabulary Specification
geo	http://www.w3.org/2003/01/geo/wgs84_pos#	9. W3C WGS84 Geo Positioning: an RDF vocabulary
gn	http://www.geonames.org/ontology#	10. GeoNames Ontology
gns	http://sws.geonames.org/	11. GeoNames Entity
lcsh	http://id.loc.gov/authorities/subjects	12. Library of Congress Subject Headings
org	http://www.w3.org/ns/org#	13. W3C Organization Ontology
prov	http://www.w3.org/ns/prov#	14. W3C Provenance Ontology (PROV)
r4r	http://guava.iis.sinica.edu.tw/r4r/	15. Relations for Reusing Ontology (r4r)
schema	http://schema.org/	16. Schema.org
skos	http://www.w3.org/2004/02/skos/core#	17. W3C Simple Knowledge Organization System (SKOS)
time	http://www.w3.org/2006/time#	18. W3C Time Ontology
voaf	http://purl.org/vocommons/voaf#	19. Vocabulary of a Friend (VOAF)
wde	http://www.wikidata.org/entity/	20. Wikidata Entity
Domain Knowledge		
aat	http://vocab.getty.edu/aat/	1. Art and Architecture Thesaurus
dwc	http://rs.tdwg.org/dwc/terms/	2. Darwin Core Terms
dwciri	http://rs.tdwg.org/dwc/iri/	3. Darwin Core terms
eol	http://eol.org/pages/	4. The Encyclopaedia of Life (EOL)
txn	http://lod.taxonconcept.org/ontology/txn.owl#	5. Taxon Concept OWL Ontology
Local Namespace		
voc	http://voc.odw.tw/ontology#	1. Ontology for ODWeb (voc4odw)
agent	http://data.odw.tw/agent/	2. Organization/Agent Entity in ODW
article	http://data.odw.tw/article/	3. Textual Description with *rdf:type* r4r:Article in ODW
code	http://data.odw.tw/code/	4. Code Description with *rdf:type* r4r:Code in ODW
data	http://data.odw.tw/record/	5. Linked Data for ODWeb
evt84	http://data.odw.tw/event/	6. Event Entity in ODW
project	http://data.odw.tw/project/	7. Project Entity in ODW
r1 (n)	http://data.odw.tw/r1/ (r2, r3…)	8. Refined Version(s) of ODW Entity
refined	http://data.odw.tw/refined/	9. Directory of the Refined Versions
catdat	http://catalog.digitalarchives.tw/	10. Union Catalog of Digital Archives Taiwan

建議與結論

總結現階段實作案例的成果面向有三：一、資料面向：提供使用者可選擇的多種連結資料格式、以及相關的轉換程式碼，協助使用者進行符合自身需求的資料語意處理與再次使用，本實驗產生的結構資料三元組約四千五百萬筆；二、技術面向：使用並擴充 CKAN 軟體套件，發展為「開放資料連結」的儲存與展示平台，收納整合前述之連結式資料，建置常人與機器皆可瀏覽與操作的「開放資料連結」系統 data.odw.tw；三、語意面向：已建置強化語意版的資料連結集（R 版），並連結知名的知識庫 (如 Wikidata, GeoNames, EOL 等) ，此資料集約有二千五百萬筆。並透過知識本體 voc4odw 的設計，運用語意描述的模組化機制，提供基礎都柏林核心集的描述版本 (D 版) 與資料溯源，更進一步針對時空語意描述加強、設計 R 版彈性多重機制（資料多重清理、知識多重連結、語意多重架構），讓資料策展者或使用者能重新建構資料語意連結，走向資料盡用、語意整合、知識連結的語意網世界。

目前的個人行動裝置多已可處理資料連結與結構語意，預期不遠的未來將是數十億人在機器協助下進行即時、富語意、多樣態的資源連結與快速互動。我們應盡速培養熟悉並可連結資料、常人和機器三者的技術人才，才能面對快速進展的數位環境與全新型態的挑戰。

致謝

特別感謝中央研究院資訊科學研究所陳克健研究員所帶領的典藏台灣聯合目錄團隊過去幾年的支持與協助。我們尤其感謝洪崇熙先生在資料收集與處理的幫助、以及曹晉豪先生與陳心萍小姐在資料清理與比對的協力。

附註

[1] 官方網站位於 http://ckan.org。

[2] http://catalog.digitalarchives.tw/

[3] 本文所有 SPARQL 查詢請參考 http://data.odw.tw/examples/JLIS-2016-query.html

[4] 包括: 宜蘭三星大同鄉棲蘭山林道; 花蓮秀林萬榮; 苗栗泰安南庄，新竹尖石鴛鴦湖、及嘉義縣等地。

[5] “a surprising amount of data isn't linked in 2006” at https://www.w3.org/DesignIssues/LinkedData.html , 2007 年約 12 個資料集

[6] 超過 130,502,164,357 的「開放資料連結」 Triples 來自 2740 資料集，而此數字不包括總「開放資料連結」 資料集超過 9960 與各機構組織未開放的 Linked Data 的統計內。數字來自: http://stats.lod2.eu/ (http://lodstats.aksw.org/) 2016/11/03

[7] OCLC 視 LOD 與知識庫計畫為該機構 Data Science 要項 (http://www.oclc.org/research/themes/data-science.html)，此研究報告為 Karen Smith-Yoshimura 於在 2016 年 4 月 CNI Spring Membership Meeting 簡報: Linked Data Implementations—Who, What and Why? http://www.oclc.org/content/dam/research/presentations/smith-yoshimura/oclcresearch-linked-data-implementations-cni-2016.pptx

[8] http://www.europeana.eu/

[9] https://docs.google.com/document/d/1JvjrWMTpMIH7WnuieNqcT0zpJAXUPo6x4uMBj1pEx0Y/

[10] http://www.ld4l.org/

[11] https://www.ld4l.org/ld4l-labs/

[12] Columbia University, Library of Congress, Princeton University

[13] http://www.ld4l.org/ld4p/

[14] www.opencyc.org ; http://sw.opencyc.org/

[15] www.dataversity.net/opencyc-hooks-into-linked-data-web/

[16] 透過 Open Source Texai Project 發佈 RDF 相容的格式。https://sourceforge.net/projects/opencyc/files/

[17] www.cyc.com/platform/researchcyc/

[18] www.freebase.com

[19] rdf.freebase.com/

[20] www.wikidata.org

[21] www.wikidata.org/wiki/Wikidata:WikiProject_Freebase

[22] www.mpi-inf.mpg.de/departments/databases-and-information-systems/research/yago-naga/yago/

[23] dbpedia.org

[24] wiki.dbpedia.org/online-access/DBpediaLive

[25] live.dbpedia.org

[26] 超過 30 個以上外部連結資料庫如 Amsterdam Museum, BBC, Eurostat Linked Statistics, CIA World Factbook,GeoNames, GeoSpecies, LinkedGeoData, New York Times, OpenCyc, WordNet, YAGO ... 等。http://wiki.dbpedia.org/Downloads2015-10

[27] 正確性 (Accuracy), 可信度 (Trustworthiness), 一致性 (Consistency), 相關性 (Relevancy), 完整性 (Completeness), 適時性(Timeliness), 易了解性(Ease of understanding),互通性(Interoperability), 可取得性(Accessibility), 授權(Licensing), 相互連結, (Interlinking)

[28] See yago:extractionTechnique 與 yago:extractionSource at wiki.cfcl.com/Projects/YAGO/Predicates

[29] sws.geonames.org

[30] linkedgeodata.org

[31] www.openstreetmap.org

[32] vocab.getty.edu/tgn

[33] data.ordnancesurvey.co.uk/datasets/opennames

[34] http://lod-cloud.net/

[35] https://www.ordnancesurvey.co.uk/blog/2010/04/os-opendata-goes-live/

[36] http://data.ordnancesurvey.co.uk/datasets/os-linked-data/about; 但連結資料則自 2009 年 10 月對外公布 http://lists.w3.org/Archives/Public/public-lod/2009Oct/0136.html

[37] http://data.ordnancesurvey.co.uk/ontology

[38] openstreetmap.org

[39] aims.fao.org/standards/agrovoc/linked-open-data

[40] http://aat.teldap.tw/

[41] http://eol.org/

[42] http://eol.org/traitbank

[43] http://eol.org/pages/1134120/

[44] http://eol.org/pages/1134120/maps

[45] 參見台灣一葉蘭 data:d2148340 與 data:d4542169 中 dc:subject 描述差異。

[46] EOL 新增資料類型方法尚包括使用文字探勘、公民科學參與、標本資料數位化等項目。

[47] “All metadata is dirty.”

[48] 全部典藏品整合超過 90 計畫單位，就 84 萬筆 CC 授權資料而言，則來自 74 個跨領域單位。

[49] http://catalog.digitalarchives.tw/item/00/47/03/74.html

[50] 此表源自本研究於 2015 年 12 月調查報告 (http://goo.gl/pPUXcd) ，當時並比較 W3C 的 Data Quality Vocabulary 所提十面向 (https://www.w3.org/TR/2015/WD-vocab-dqv-20150625/)， 目前最新 dqv 尚未進入正式推薦標準語彙，(https://www.w3.org/TR/2016/NOTE-vocab-dqv-20160830/)，且觀察不同版本間的變動差異甚多，因此先不列入此表。

[51] https://www.w3.org/TR/vocab-dqv/#mapping-ISOZaveri

[52] https://www.w3.org/DesignIssues/UI.html

[53] http://data.odw.tw/record/d2148340 本文所使用的命名空間請參照表三。

[54] http://www.w3.org/TR/prov-o/

[55] 目前台灣一葉蘭 (data:d2148340) 後設資料溯源為 data: p20160530-d2148340 與 data:p20160912-d2148340，若有新的版本，後設資料溯源會根據資源產生日期生成。

[56] R4R 本體論中英文以及 RDF/Turtle 檔案下載可參看: http://guava.iis.sinica.edu.tw/r4r/

[57] http://dat.digitalarchives.tw/

[58] http://dat.digitalarchives.tw/ontology.html

[59] https://gitlab.com/iislod/

[60] http://catalog.digitalarchives.tw/item/00/00/46/14.html

[61] http://catalog.digitalarchives.tw/item/00/3a/3d/14.html

[62] http://blogs.loc.gov/thesignal/2012/03/the-value-of-a-broken-link/

[63] 「CKAN instances around the world」頁面：http://ckan.org/instances/。

[64] 註 a：以經修改之 ckanext-dcat 採集描述（profile），透過 ckanext-harvset 採集機制達成。註 b：以 ckanext-scheming 與 ckanext-repeating 套件達成。註 c：以 ckanext-dcat 輸出描述與 rdflib 函式庫達成。註 d：以 ckanext-sparql 套件達成。Icon made by SimpleIcon (http://www.flaticon.com/authors/simpleicon) and Freepik (http://www.flaticon.com/authors/freepik).

[65] 版本 0.11 於 2010 年 1 月釋出。見 http://docs.ckan.org/en/latest/changelog.html#v0-11-2010-01-25。

[66] https://github.com/ckan/ckanext-dcat。最早的提交（commit）時間為 2013 年 7 月 3 日。

[67] 以「鋼鐵沈思少女」(http://data.odw.tw/record/d4502674）此一藏品為例，若欲取得 Turtle 格式之資料連結，只需在其後加上.ttl（http://data.odw.tw/dataset/d4502674.ttl）即可。

[68] https://github.com/ckan/ckanext-scheming。

[69] https://github.com/open-data/ckanext-repeating。

[70] https://github.com/ckan/ckanext-spatial。

[71] https://gitlab.com/iislod/ckanext-tempsearch。

[72] https://rdflib3.readthedocs.io。

[73] http://virtuoso.openlinksw.com/。

[74] 該查詢介面位於 http://data.odw.tw/sparql。

[75] ckanext-dcat、ckanext-harvest、ckanext-scheming、ckanext-repeating、ckanext-spatial、ckanext-tempsearch，計有六個擴充套件。

[76] Ontology for ODWeb: http://voc.odw.tw/ontology/; 本文所使用的 namespace 請參照表 3。

[77] {evt84:phyCre-d2148340 gn:parentFeature gns:1667637}

[78] {data:d2148340 txn:hasEOLPage <http://eol.org/pages/1134120> }

[79] http://literature.tfri.gov.tw/atlas/content1.jsp?item=45600

[80] http://lov.okfn.org/dataset/lov/terms?q=species&vocab_limit=0

[81] “which accepts the tensions that lie between different interpretations and understandings of an object.”

[82] http://lov.okfn.org/

Reference

Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., & Ives, Z. (2007). Dbpedia: A nucleus for a web of open data. The semantic web, 722-735.

Baca, M., & Gill, M. (2015). Encoding Multilingual Knowledge Systems in the Digital Age: the Getty Vocabularies. Knowledge Organization, 42(4).

Batini, C., Cappiello, C., Francalanci, C., & Maurino, A. (2009). Methodologies for data quality assessment and improvement. ACM computing surveys (CSUR), 41(3), 16.

Bizer, C., Lehmann, J., Kobilarov, G., Auer, S., Becker, C., Cyganiak, R., & Hellmann, S. (2009). DBpedia-A crystallization point for the Web of Data. Web Semantics: Science, Services and Agents on the World Wide Web, 7(3), 154-165.

Bollacker, K., Evans, C., Paritosh, P., Sturge, T., & Taylor, J. (2008). Freebase: a collaboratively created graph database for structuring human knowledge. In Proceedings of the 2008 ACM SIGMOD international conference on Management of data (pp. 1247-1250). ACM.

Burgess, L. C. (2016). Provenance in Digital Libraries: Source, Context, Value and Trust. In Building Trust in Information (pp. 81-91). Springer International Publishing.

Carata, L., Akoush, S., Balakrishnan, N., Bytheway, T., Sohan, R., Seltzer, M., & Hopper, A. (2014). A primer on provenance. Communications of the ACM, 57(5), 52-60.

Charles, V. (2016) Linked Data for Europeana Cultural Heritage: the Europeana approach, Presentation given on April 28th in Paris at International Conference organised by ISSN IC: "Bibliographic metadata getting linked…", http://www.slideshare.net/ValentineCharles/linked-data-for-europeanacultural-heritage-the-europeana-approach.

Charles, V., Manguinhas, H., Alexiev, V., Charles, V., & Dammers, M. (2015). Wikidata, a Target for Europeana's Semantic Strategy. Glam-Wiki 2015.

Chuttur, M. Y. (2014). Investigating the effect of definitions and best practice guidelines on errors in Dublin Core metadata records. Journal of Information Science, 40(1), 28-37.

De Sabbata, S., & Acheson, E. (2016). Geographies of gazetteers in Great Britain. In 24th GIS Research UK (GISRUK 2016) conference, University of Greenwich, April 2016, http://hdl.handle.net/2381/38182.

Dextre Clarke, S. G. (2016). Origins and Trajectory of the Long Thesaurus Debate. Knowledge Organization, 43(3).

Emani, C. K., Cullot, N., & Nicolle, C. (2015). Understandable big data: A survey. Computer Science Review, 17, 70-81.

Ermilov, I., & Pellegrini, T. (2015). Data licensing on the cloud: empirical insights and implications for linked data. In Proceedings of the 11th International Conference on Semantic Systems (pp. 153-156). ACM.

Erxleben, F., Günther, M., Krötzsch, M., Mendez, J., & Vrandečić, D. (2014). Introducing Wikidata to the linked data web. In International Semantic Web Conference (ISWC) (pp. 50-65). Springer International Publishing.

Färber, M., Bartscherer, F., Menne, C., & Rettinger, A. (2016). Linked Data Quality of DBpedia, Freebase, OpenCyc, Wikidata, and YAGO. Semantic Web (Under Review), Cited version status (Minor Revision): http://www.semantic-web-journal.net/system/files/swj1465.pdf.

Ford, H., & Graham, M. (2016). Provenance, power and place: Linked data and opaque digital geographies. Environment and Planning D: Society and Space, 0263775816668857.

Fürber, C., & Hepp, M. (2013). Using semantic web technologies for data quality management. In Handbook of data quality (pp. 141-161). Springer Berlin Heidelberg.

Godby, C. J. (2016). Seeding the Linked Data Cloud: The present and future of library. Days of Knowledge Organization, Oslo and Akershus University. Retrieved from http://edu.hioa.no/korg2016/korg2016_godby.pdf .

Goodwin, J., Dolbear, C., & Hart, G. (2008). Geographical linked data: The administrative geography of Great Britain on the semantic web. Transactions in GIS, 12(s1), 19-30.

Hallo, M., Luján-Mora, S., Maté, A., & Trujillo, J. (2016). Current state of Linked Data in digital libraries. Journal of Information Science, (42), 117-127.

Haslhofer, B., & Isaac, A. (2011). data. europeana. eu: The europeana Linked Open Data pilot. In International Conference on Dublin Core and Metadata Applications (pp. 94-104).

Hoffart, J., Suchanek, F. M., Berberich, K., & Weikum, G. (2013). YAGO2: A spatially and temporally enhanced knowledge base from Wikipedia. Artificial Intelligence, 194, 28-61.

Hoffart, J., Suchanek, F. M., Berberich, K., Lewis-Kelham, E., De Melo, G., & Weikum, G. (2011, March). YAGO2: exploring and querying world knowledge in time, space, context, and many languages. In Proceedings of the 20th international conference companion on World Wide Web (pp. 229-232). ACM.

Huang, A. W. C., & Chuang, T. R. (2014). Relations for Reusing (R4R) in a Shared Context: An Exploration on Research Publications and Cultural Objects. Semantic Digital Archives, SDA@ JCDL/TPDL (pp. 49-60).

Ismayilov, A., Kontokostas, D., Auer, S., Lehmann, J., & Hellmann, S. (2016). Wikidata through the Eyes of DBpedia. Semantic Web (Under Review/ Decision/Status: Minor Revision) : http://www.semantic-web-journal.net/system/files/swj1462.pdf.

Janssen, M., Charalabidis, Y., & Zuiderwijk, A. (2012). Benefits, adoption barriers and myths of open data and open government. Information Systems Management, 29(4), 258-268.

Knoblock, C. A., & Szekely, P. A. (2015). Exploiting Semantics for Big Data Integration. AI Magazine, 36(1), 25-38.

Lee, C.J., Huang, A.W.C., & Chuang,T.R., (2016) A Linked Open Data Repository Built with CKAN, CKANCon 2016, October 4th, Madrid, Spain (http://ckan.org/ckancon-2016/)

Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P. N., ... & Bizer, C. (2015). DBpedia–a large-scale, multilingual knowledge base extracted from Wikipedia. Semantic Web, 6(2), 167-195.

Mahdisoltani, F., Biega, J., & Suchanek, F. (2015). Yago3: A knowledge base from multilingual wikipedias. In 7th Biennial Conference on Innovative Data Systems Research. CIDR Conference.

Marden, J., Li-Madeo, C., Whysel, N., & Edelstein, J. (2013). Linked open data for cultural heritage: evolution of an information technology. In Proceedings of the 31st ACM international conference on Design of communication (pp. 107-112). ACM.

Meroño-Peñuela, A., Ashkpour, A., Van Erp, M., Mandemakers, K., Breure, L., Scharnhorst, A., ... & Van Harmelen, F. (2014). Semantic technologies for historical research: A survey. Semantic Web, 6(6), 539-564.

Mitchell, E. T. (2016). The Current State of Linked Data in Libraries, Archives, and Museums. Library Technology Reports, 52(1), 5-13.

Moura, T. H., & Davis Jr, C. A. (2014). Integration of linked data sources for gazetteer expansion. In Proceedings of the 8th Workshop on Geographic Information Retrieval. ACM.

Omitola, T., Gibbins, N., & Shadbolt, N. (2010). Provenance in linked data integration. In Proceedings of the Workshop on Linked Data in the Future Internet at the Future Internet Assembly Ghent, Belgium, December 16-17, 2010 (LDFI-2010)..

Park, J. R., & Childress, E. (2009). Dublin Core metadata semantics: An analysis of the perspectives of information professionals. Journal of Information Science, 35(6), 727-739.

Parr, C. S., Schulz, K. S., Hammock, J., Wilson, N., Leary, P., Rice, J., & Corrigan Jr, R. J. (2016). TraitBank: Practical semantics for organism attribute data. Semantic Web, 7(6), 577-588.

Parr, C. S., Wilson, N., Leary, P., Schulz, K., Lans, K., Walley, L., ... & Holmes, J. (2014). The Encyclopedia of Life v2: providing global access to knowledge about life on earth. Biodiversity Data Journal, 2, e1079.

Poole, A. H. (2016). The conceptual landscape of digital curation. Journal of Documentation, 72(5), 961-986.

Schaible, J., Gottron, T., & Scherp, A. (2014). Survey on common strategies of vocabulary reuse in linked open data modeling. In European Semantic Web Conference (pp. 457-472). Springer International Publishing..

Schmachtenberg, M., Bizer, C., & Paulheim, H. (2014). Adoption of the linked data best practices in different topical domains. In International Semantic Web Conference (pp. 245-260). Springer International Publishing.

Srinivasan, R., Becvar, K., Boast, R., & Enote, J. (2010). Diverse knowledges and contact zones within the digital museum. Science, Technology & Human Values.

Stadler, C., Lehmann, J., Höffner, K., & Auer, S. (2012). Linkedgeodata: A core for a web of spatial open data. Semantic Web, 3(4), 333-354.

Stvilia, B., Gasser, L., Twidale, M. B., & Smith, L. C. (2007). A framework for information quality assessment. Journal of the American society for information science and technology, 58(12), 1720-1733.

Suchanek, F. M., Kasneci, G., & Weikum, G. (2007, May). Yago: a core of semantic knowledge. In Proceedings of the 16th international conference on World Wide Web (pp. 697-706). ACM.

Suchanek, F. M., Kasneci, G., & Weikum, G. (2008). Yago: A large ontology from Wikipedia and Wordnet. Web Semantics: Science, Services and Agents on the World Wide Web, 6(3), 203-217.

Tani, A., Candela, L., & Castelli, D. (2013). Dealing with metadata quality: The legacy of digital library efforts. Inf. Process. Manage., 49(6), 1194-1205.

Van Hooland, S., & Verborgh, R. (2014). Linked Data for Libraries, Archives and Museums: How to clean, link and publish your metadata. Facet.

Vandenbussche, P. Y., Atemezing, G. A., Poveda-Villalón, M., & Vatant, B. (2015). Linked Open Vocabularies (LOV): a gateway to reusable semantic vocabularies on the Web. Semantic Web, (Preprint), 1-16.

Voß, J. (2016). Classification of Knowledge Organization Systems with Wikidata. In Proc. of the 15th European Networked Knowledge Organization Systems Workshop (NKOS 2016). CEUR-WS. org, Hannover, Germany.

Vrandečić, D., & Krötzsch, M. (2014). Wikidata: a free collaborative knowledgebase. Communications of the ACM, 57(10), 78-85.

Yasser, C. M. (2011). An analysis of problems in metadata records. Journal of Library Metadata, 11(2), 51-62.

Yus, R., & Pappachan, P. (2015). Are Apps Going Semantic? A Systematic Review of Semantic Mobile Applications. In 1st International Workshop on Mobile Deployment of Semantic Technologies (MoDeST 2015), co-located with the 14th International Semantic Web Conference (ISWC 2015), Bethlehem, PA (USA) (Vol. 1506, pp. 2-13).

Zaveri, A., Rula, A., Maurino, A., Pietrobon, R., Lehmann, J., & Auer, S. (2016). Quality assessment for linked data: A survey. Semantic Web, 7(1), 63-93.

Zhu, R., Hu, Y., Janowicz, K., & McKenzie, G. (2016). Spatial signatures for geographic feature types: Examining gazetteer ontologies using spatial statistics. Transactions in GIS, 20(3), 333-355.