具語意鏈結能力之文化領域政府開放資料服務設計*

陳純美[1]　廖峻鋒[2]
[1]文化部資訊處　[2]逢甲大學資訊工程系
[1]grace@moc.gov.tw, [2]cfliao@fcu.edu.tw

摘要

政府開放資料服務主要目的為將公部門資訊提供公眾免費或最小限制的再利用，以促進公民參與並創造新商業機會。我國具備政府資訊電子化程度高、網際網路基礎建設普及的優勢，近年來亦積極推行相關服務。然而，目前各級單位相關服務均個別獨立且大部份無法符合 W3C 所建議以語意與鏈結為基礎的五星資料服務品質指標，不但難以與世界接軌，且開發人員使用開發資料時亦事倍功半。本論文匯整針對此一問題的初期探索成果，延續已上線之文化部開放資料服務，重新設計領域模型並提出支援具語意鏈結能力之開放資料服務平台架構，最後並實作初步原型。期能藉此拋磚引玉，引發更多討論，並做為下一階段各級政府機關開放資料服務參考。

關鍵詞　鏈結開放資料、開放政府資料、語意網

1. 前言

有鑒於資訊及行動通訊技術快速普及於日常生活的趨勢，自 1990 年代起，世界各國相繼推動電子化政府(eGovernment)，將政府資訊服務結合網際網路，藉以提高行政效率並增強與民眾之間的互動。 而整體發展趨勢則是從早期「公共事務管理」推移到當前的「公共服務創新」，並逐漸走向「公共價值創造」的發展目標。開放政府資料(Open Government Data, OGD)的目標正是基於「公共價值創造」的理念，將政府蒐集及產生的資料，主動發布在網路，提供公眾重用(Reuse)，藉以創造新的商業機會，提高政府透明度和公民參與。政府以開放政府資料促進及吸引企業和公民參與，共同協作及重複利用開放的政府資料，企業及公民可以更容易地存取與重新使用這些資料，進而創造新的加值產品和服務，政府透過網路公開資料，可從資訊創新的角度切入與整合相關資源，藉以推動政府資訊有效益的散播與使用。

近年來，在世界各先進國家政府推動下，OGD 已成為世界各國電子化政府的重要政策之一。美國政府在 2009 年首先推出 OGD 服務網 Data.gov [1]，英國政府隨後於 2010 年 1 月推出 Data.gov.uk[2]，歐盟也鼓勵會員國開放政府資料，並建置共同入口網站[3]。截至 2013 年 7 月為止，全球 OGD 資料集累計已超過 100 萬份[4]，但因發布格式彼此不相容，有些屬非結構化的專屬格式如 word、pdf，有些已結構化為欄位，但無法標定欄位的意義，使得資料難以重複使用、資料品質參差不齊，且無法貼近公眾需求[5]。針對開放資料的品質，Lee 提出了開放資料五星的發展評估模式(5 Stars Development Scheme)，做為 OGD 服務改善資料品質的指標[6](如表 1)。資料發展評估模式目前已成為各國政府經常引用的重要指標。Machado 等人指出，經由五星鏈結開放資料，可有效提升資料品質，進而誘發公眾產生創新及混搭應用[7];經由鏈結技術則可以組合許多不同領域詞彙，降低資訊取得的困難度及費用[8]。

表 1: 開放資料五星的發展評估模式[6]

分級	說明
一星	將資料以任何格式放上網路，使用者依規範使用開放的資料，如 word、pdf、jpg 等。
二星	資料格式為程式可讀取的結構化資料(如以 Excel 取代表格的影像掃瞄檔)。
三星	資料格式以開放(而非專屬特定公司)的資料格式呈現(如以 CSV 取代 Excel)。
四星	以 W3C 定義的開放標準規範格式 RDF 來描述資料並以 SPARQL 語言選定所要查詢的資料。
五星	在符合四星外，目前資料項目也能夠具備指向其他相關資料項目的鏈結。此處的鏈結指的是可定址的 URI。資料的使用者可透過 URI 參考到其它資料項目，其功能有如網頁的超鏈結功能。

我國自 1998 年起就開始推行電子化政府。基於世界各國推動開放政府資料的趨勢，行政院已正式將開放政府資料服務列為推動第四階段電子化政府發展的重要主軸。其中，文化部被列為先行試辦單位，已於 2012 年 11 月釋出第一版文化領域開放資料服務，提供 EXCEL、XML 及 JSON 三種格式[9]。然而，根據我們對國內各級政府機關所做之初步調查(如表 2)顯示，截至 2013 年 6 月底，我國

*本論文提出之方法純為學術研究探索，不代表任職機關之立場與發展方向; 本論文接受教育部 102 年度資訊軟體人才培育推廣計畫與國科會研究計畫編號 102-2221-E-035-039 經費補助。

政府機關開放的資料集均為三星以下之非結構化或結構化格式，且大部份並未依 W3C(World Wide Web Consortium)格式發布，使得資料難以重複使用，也難以與國際接軌。此外，目前資料集為個別獨立，無法呈現相互關聯性，使用者也無法確認不同詞彙是否為相同涵義，增加維護個別資料集的困難度。最後，在資料提供時，也缺乏共通的 API 設計，必須針對特定的需求，做專屬設計，使用者無法在不同資料集間任意查詢、組合所需資料。

本研究主要目的即為針對上述問題，基於目前文化開放資料服務建置成果，以五星資料發展評估模式[6]做為指引，設計領域模型並提出支援建構具語意鏈結能力之開放資料服務平台，最後實際實作出初步原型。本論文包含了此初期探索的進行程序及相關架構議題討論與研析，期能做為下一階段政府各機關開放資料服務架構的參考。

2. 背景與相關研究

自 1998 年 W3C 啟動 e-Government Activity 以來[8]，電子化政府(e-Government)已成為世界各國政府提升行政效率的重要措施，且近年有從 e-Government 走向 e-Governance 的趨勢。e-Governance 指的是善用資訊通訊技術提供政府行政、資料交換及整合的服務，包含政府對政府(G2G)、政府對公民(G2C)及政府對廠商(G2B)等面向[10]。因此開放資料相關技術可說是實現 e-Governance 的重要前提。開放資料一般以個別資料集(Datasets)的型態發布在網路。其中，資料集指的是相同類別資料的集合。政府開放資料並非直接將大量未整理資料放上網路，而是由資料提供單位經過特定流程並確保資料品質後發布。W3C 也提供一般性的設計指引來導引整個資料設計的過程[6]。以下我們分別針對國內外 OGD 服務的現況進行簡要說明與討論。

2.1 國外開放政府資料現況

根據聯合國經濟及社會事務部指出，在 193 個會員國裡，於政府網站上推動開放資料的國家已超過四分之一[11]。其中美國、英國兩國可說是開放政府資料的先驅。美國的開放資料發展是由小規模但具影響力的公民駭客(Civic Hackers)社群開始，將美國國會議員的訪問、活動等公開在網站上的資料，轉換成可供搜尋、再利用的資料格式，受此影響，美國一些州開始釋出州級的官方資料。歐巴馬總統在就任當天，更立即簽署了「透明與開放政府備忘錄」，作為美國聯邦政府在未來四年的開放政府資料政策準則並於同年 5 月建立了統一入口網站 Data.gov[1]。 英國內閣辦公室則是在 2012 年 6 月公布的政策白皮書指出，開放資料將是未來政策發展的重點。目前英國政府的開放資料規劃團隊以 Web 之父 T. B. Lee 為首，大力推行五星級資料指標[2]。

歐盟開放資料入口網站則提供了鏈結資料目錄，使開發人員更容易連接不同來源的資訊。鏈結資料目錄以 SPARQL(SPARQL Protocol and RDF Query Language)語法提供服務。此外，歐盟最大的文化遺產入口網站 Europeana 則於 2012 年 2 月開始進行 Linked Data Pilot 計畫[12]，結合來自 15 個國家的資料，將其轉換成五星資料，並置於 data.europeana.eu 開放使用。

2.2 國內開放政府資料現況

國內於 2012 年起設定以開放政府資料為電子化政府第四階段主軸，於資訊服務產業層面，期許以建立基礎資料開放民間加值運用機制，帶動我國資訊服務產業往高附加價值發展，提升產業競爭力。開放政府資料服務由台北市政府首先於 2011 年 8 月推出國內第一個開放資料平台 Data.Taipei，接著台中市政府、文化部及新北市的開放資料服務相繼於 2012 底陸續上線。行政院研考會於 2013 年 4 月推出了政府資料開放平臺公開測試版 Data.gov.tw，收錄國內政府機關開放資料集目錄。

為了解國內中央及縣市政府開放資料的一般狀況，我們針對各級政府機關資料開放服務做了一個全盤性普查。截至 2013 年 6 月為止，國內主要已提供開放資料之機關資料概況如表 2 所示。由表中可看出目前政府開放資料服務主要分佈於資料發展評估模式的一星至三星，根據研考會 101 年度電子治理委外服務計畫案，建議於遠程目標再決定是否與何時需要達到四星和五星的開放資料格式標準[13]，而當前仍以確保資料平台的互通性、資料存取的易用性與穩定性等，為開放資料的重點項目[14]，因此相對於世界各國顯得較為保守。

國內開放社群近年來亦積極參與政府開放資料推動。例如財團法人青平台基金會於 2012 舉辦了 Open Campus，邀請在開放資料領域的專家們討論如何掌握資料及善用資料的力量。Code for Tomorrow 一個非營利組織，推廣利用開放資料進行程式開發，進而達成「寫程式改造社會」的理念。零時政府(g0v.tw)則定位為一個全民參與的公民運動，透過對政府開放資料的分析，以公民的角色對政府的施政進行監督。

3. 開放資料領域模型

本研究聚焦於文化領域之政府開放資料，首先必須針對文化領域專有的重要概念名詞、屬性及其關係建立領域模型。建立某領域專屬的領域模型必須投入一定量的人力與時間成本。Hyland 與 Wood 指出，由於領域模型建立後，對於組織資料電子化有極重大的助益，相對於大部份的公司必須持續投

表 2: 截至 2013 年 6 月底之國內政府機關開放資料狀況(1:一星, 2:二星, 3:三星)

名稱	WORD[1]	EXCEL[2]	KML[1]	KMZ[1]	WMS[1]	PNG[1]	PDF[1]	TXT[2]	CSV[3]	XML[3]	JSON[3]
內政部								V	V	V	
外交部									V	V	
國防部								V	V		
財政部									V	V	
教育部									V		
法務部									V		
經濟部				V	V				V		
交通部								V	V	V	
文化部		V								V	V
故宮博物院										V	
國科會			V			V			V	V	
研考會								V	V	V	
台北市	V		V							V	V
新北市		V							V	V	V
台中市	V	V	V						V	V	
台南市									V	V	V
宜蘭縣							V		V		

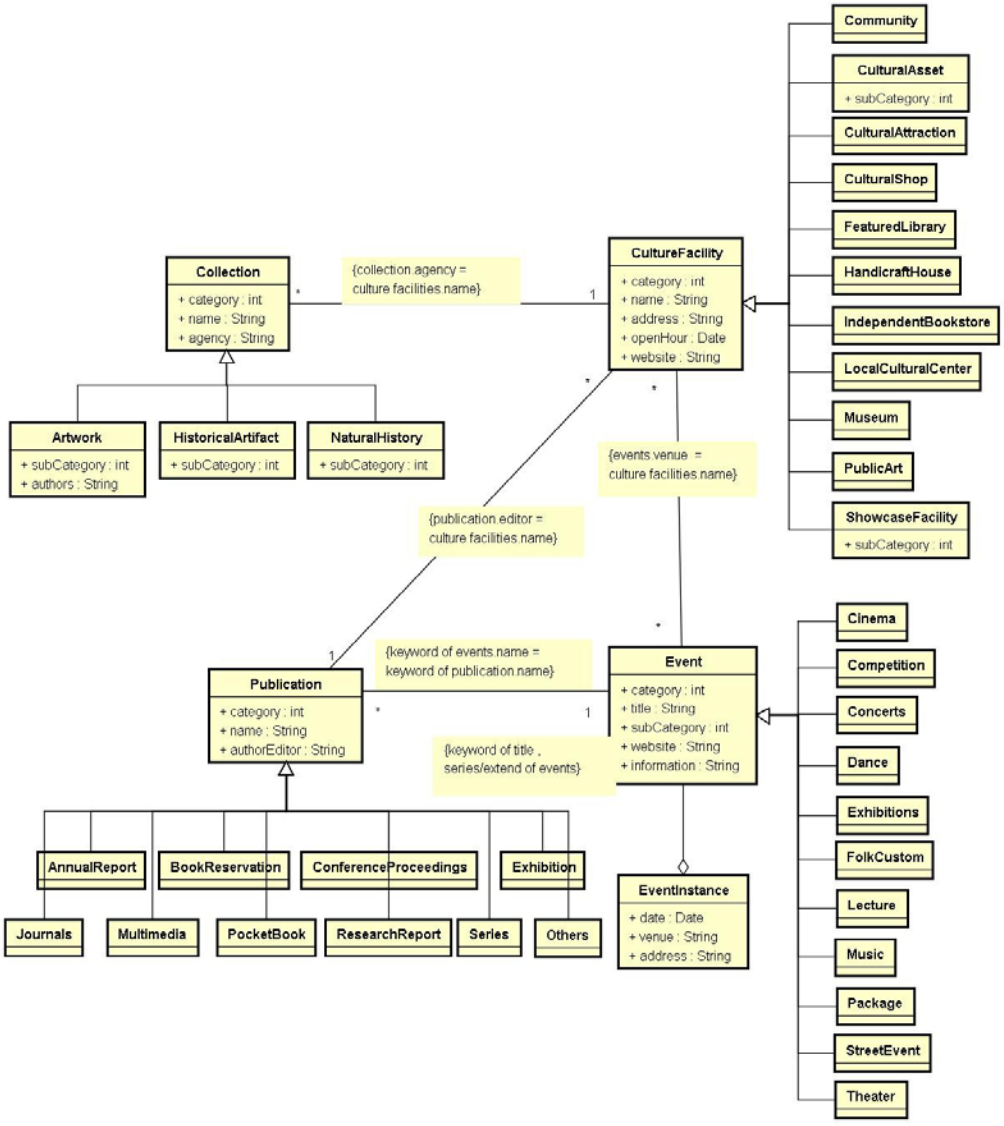

圖 1: 文化領域開放資料模型

入大量資源進行資料整理與維護，領域模型所投入的一次性成本顯得十分微小[16]。建構領域模型的程序可分成四大階段，分述如下:

步驟 1: 蒐集及分析領域資料案例 首先訂定欲查詢之領域資料來源的計畫文件及網站，作為資料蒐集的案例。我們以「文化雲先期計畫」之專案 (如文化地圖建置案、藝文活動報名系統、典藏共構系統、文化資料開放及文化資源庫)文件資料共 5 件、國內文化機關官網及主題網站，如文化部及所屬機關、國立故宮博物院、文化臺灣、藝文資源服務網、文化資料開放服務網等共 10 個網站、文化專案相關會議文件約 12 件，作為領域資料檢視及網站訪視的資料來源。資料蒐集完成後，我們採人工檢視計畫文件及網站訪視方式，將重要領域名詞抽取出來，依資料特性進行大略分組，刪除、合併概念相近或重覆的名詞，之後將剩下的重要領域概念名詞表列，並進行下一步驟。

步驟 2: 分析領域元素屬性及關係 在此一步驟，我們綜整步驟 1 得出之重要領域概念名詞，進一步分析其屬性及關係。關聯式資料庫及物件導向分析領域已存在許多關於建立領域模型的方法與程序。由於以 RDF(Resource Description Framework)為基礎的資料儲存結構較接近物件/類別型態，因此本研究採行物件導向領域驅動設計(Domain –Driven Deisgn)[15]，輔以 CRC (Class Responsibility Collaborator) 卡方法進行分析。首先，我們藉由 CRC 卡方法發想個別概念名詞可能的責任、屬性，並研析各概念的相關性，並繪製為初步 UML 類別圖。

步驟 3: 領域專家訪談與驗證 為了驗證步驟 2 初步分析的可行性，我們針對不同領域專長之文化人士共計 11 人訪談。訪談對象包含文化部所屬機關業務推展人員及文化相關專案承商。在統整專家意見後，對初步領域模型進行改善。例如，「徵件」與「競賽」分別為不同類別，但深入比對其包含的內容，具有很高的相似度，經專家訪談後，大部份認為應合併為一項。此外，原本分析了文化人才資料部份，但因專家均認為資料涉及個人資料，不適合列入開放資料，因此予以刪除。值得注意的是，步驟 2 與步驟 3 為持續改善領域模型品質至所需水準的遞迴程序。

步驟 4: 產出領域模型 經由步驟 2、3 的數次諮詢、修改與驗證後，本階段將包含所有必要屬性及各式關係(如聚合、繼承、依賴等等)，繪製為精確版 UML 類別圖。例如圖 1 為本研究經由此步驟所產出的文化領域模型類別圖。

文化領域開放資料可分為四大類別：典藏品(Collection)、 文化設施(Culture Facility)、藝文活動(Event)及出版品(Publication)。其中，Collection 代表博物館典藏藝術作品或文物，包含藝術品(Artwork)、歷史文物(Historical Artifacts)及自然史(Natural History)等 3 個子類; Culture Facility 為具有實際形體的文化硬體設施，單一設施可能具有跨類別的功能，社區(Community)、有形文化資產(Cultural Asset)、文化景觀(Cultural Attraction)、文創商店(Cultural Shop)、特色圖書館(Featured Library)、工藝之家(Handcraft House)、 獨立書店(Independent Bookstore)、地方文化館(Local Cultural Center)、博物館(Museum)、公共藝術(Public Art)及展演設施(Showcase Facilities)等 11 個子類;Events 代表文化與藝術相關的活動分為電影(Cinema)、徵件競賽(Competition)、演唱會(Concert)、舞蹈(Dance)、展覽(Exhibition)、民俗活動(Folk Custom)、講座(Lecture)、音樂(Music)、綜藝(Package) 、街頭活動(Street Event)及戲劇(Theater)等 11 個子類。「綜藝」為單一藝文活動資訊內容即包含各種不同類別的活動訊息，如藝術節。「街頭活動」為週期性的街頭文化藝術活動資訊，如創意市集;Publications 泛指與文化機關的各式出版品。分為年報 (Annual Report) 、典藏/圖錄/年鑑目錄(Book Reservation)、 論文集(Conference Proceeding)、展覽專輯(Exhibition)、期刊(Journal)、多媒體(Multimedia)、口袋書 (Pocket Book)、研究報告(Research Report)、叢書(Series)及其他(Others)等 10 個子類。

4. 開放資料服務平台

在美國人工智慧學會(AAAI) 2011 年開放政府知識秋季論壇中，專家學者為開放政府資料的運作建立了一個三階段模式:開放(Open)、鏈結(Link) 與重用(Reuse)[5]。基於此一模式，在建立領域模型後，接下來必須考慮的問題是如何基於領域模型新增資料至開放平台或轉換既有資料為適當的開放格式(開放階段);為了避免重覆使用不同字彙來指稱相同概念，並增進資料間的鏈結，必須藉由領域專家的協助來進一步重整資料(鏈結階段);最後這些資料會經由標準通訊協定及查詢方式(如 SPARQL)將資料開放給第三方加值，開發可創造更高的附加價值的應用程式(重用階段)。

本研究基於上述模式提出了一個可兼容於既有二、三星資料及植基於 RDF 的四星的鏈結資料開放平台(如圖 2 所示)。首先，在開放階段最主要的目的即為將資料以標準格式放上開放資料平台。根據 W3C 的建議，為了建立品質較佳的開放資料，在建立新資料集時應遵守下列四項原則[6]: 1) 為開放資料中的重要領域概念建立 URI(Uniform Resource Identifier)，並且 2)該概念應

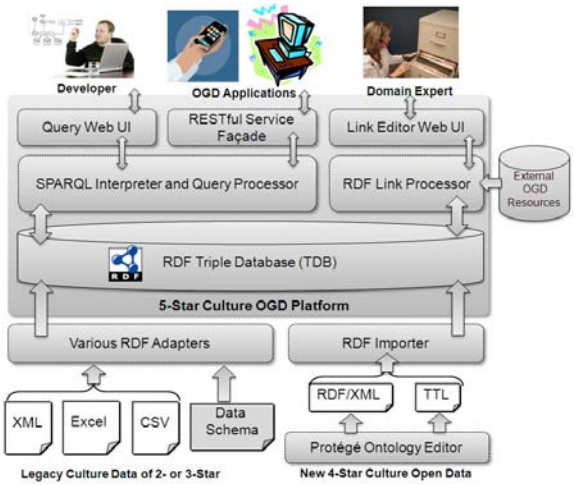

圖 2: 文化領域五星開放資料服務平台架構

可透過 HTTP 定址;3) 基於 RDF/SPARQL 標準來提供相關重要概念的資訊; 4)為這些概念建立鏈結。在新建資料時，上述原則可以大幅提升資料品質，可藉由領域專家的專業知識，使用視覺化編輯工具新增資料，並直接匯出產生為品質較佳的 RDF 標準格式，如 :N3、RDF/XML 或 TURTLE，這些格式均可完整表達 RDF 資料，其中 TURTLE 是較為易懂且格式最精簡，因此目前選用其為共通 RDF 格式。在大部份情況下，機構都會有大量的既存資料，其品質介於二至三星間(遵循私有或開放格式的結構化資料)。此時，仍然可以先請領域專家將上一節所取得的領域模型以 Protégé 建立 Ontology，並根據此一 Ontology，為每個資料集寫作一個 Adapter，將既有資料自動轉換為 RDF。

在轉換為 RDF 後，資料會全部集中在 TDB (RDF Triple Database)中，此時，外部應用程式可透過 SPARQL 查詢語言對 TDB 進行十分彈性的查詢並回傳資料。通常開發人員會透過 Web 查詢介面先下達、編寫合適的 SPARQL 指令後，再將其編碼為 GET 要求一部份，以 HTTP GET 形式取得資料(如圖 2 上半部)。然而，這種全自動轉換的開放資料，很容易產生過多類似的詞彙來描述相同的概念，並且也導致缺少有意義的鏈結[16]。因此本平台也提供 Link Editor 的 Web 介面，讓領域專家定期手動檢視整理資料，並和外部開放政府資料建立鏈結，以確保資料品質。領域專家對現有 RDF 的改善工作重點包含:1)重用現有的通用字彙: 為了支援此一工作，Link Editor 列出包括 CKAN、FOAF、VoID、Dublin Core(要再加入 Link/REF)等重要的全球共通概念字彙網站，便於領域專家加以應用。2)為相關概念建立鏈結: 目前本平台提供的 Link Editor 支援透過 rdfs:seeAlso、rdfs:subPropertyOf 與 owl:sameAs 等屬性，讓領域專家藉以建立和網內外相關概念鏈結。

5. 系統實作與討論

為驗證上述領域模型及服務平台架構設計，我們實作了一個具語意鏈結能力的文化領域開放資料服務平台的初步雛型。依圖 2 的架構設計，此雛型分為三層實作:資料處理層、資料儲存層及服務介面層。資料處理層包含了處理各式介接格式的 RDF Adapter，主要功能為讀入各式既存資料，連同領域模型輸出的 Data Schema，匯入資料儲存層。我們使用 Apache Jena Semantic Web 框架為基礎開發來開發各式 Adapters 及 RDF Importer，將所有資訊轉換為 TURTLE 格式。針對 XML、微軟 Office 文件格式及 CSV 我們分別使用 JDOM、Apache POI 及 Apache Commons CSV 等開源專案來實作。

目前雛型一共匯入了 2013 年藝文活動資料共 13 類別，2 萬多筆資料項目。在資料儲存層使用 Apache Jena 專案中用來快速處理儲存 RDF 資料的 TDB 做為資料儲存庫。在服務介面層，此平台除了提供機器可存取的 RESTful 的 SPARQL 網路服務介面外，也提供簡易的 Web SPARQL 查詢介面，提供開發人員測試使用(圖 3)。客戶端程式可透過標準 HTTP GET 命令，並將 SPARQL 編碼含入 GET 的 sparql 參數傳送，既可取得所需要的資料，目前此平台提供 JSON、XML、Text 及 CSV 等四種格式。舉例來說，客戶端可透過指令(如圖 3)尋找所有音樂類的藝文活動的標題與編號，其輸出的結果如圖 4。由於已經與場次資訊建立鏈結，因此可以透過活動編號連結，並取得場次資訊，透過鏈結再次查詢的結果如圖 5 所示。

6. 結論

本論文主要目的在於以文化部現有之開放服務為基礎，設計一個具備語意鏈結能力之文化領域開放資料服務平台。首先，我們基於相關資料，經過分析與領域專家的建議，將文化領域專有的重要概念名詞、屬性及其關係建立了一套領域模型。其次，基於五星發展評估模式設計了具語意鏈結能力之開放資料服務平台架構，並其於此一架構實作一個具語意鏈結能力的文化領域開放資料服務平台的初步雛型。本文所匯整的相關成果，可對於跨機關資料整合，協助政府機關建立符合民眾所需之開放資料集，誘發企業及公民參與共同協作及重利用開放的政府資料有所幫助。

開放政府資料平台目前在實務上仍存在許多問題與挑戰極待解決。首先，領域模型建立、管理及維護均需耗費大相當人力，未來需藉由資訊技術協助，建構半自動化的語彙分析比對工具，自動至各重要共同語彙網站(如 DBpedia 或 Dublin

圖 3: 文化領域五星開放資料服務查詢畫面

圖 4: SPARQL 查詢結果(JSON 格式)

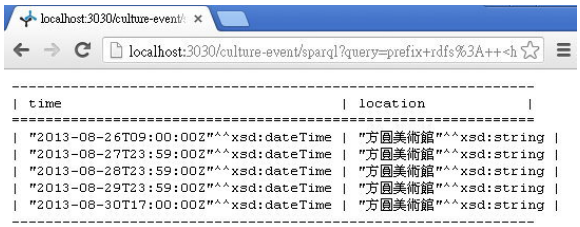

圖 5: 追蹤鏈結查詢結果(Text 格式)

Core) 搜集分析，提供專家建立領域模型建議與驗證[17]。此外，目前我們並未考量在服務上線後的延展性(Scalability)議題。針對此問題我們認為可善用開放資料服務主要存取行為「讀取」的前提，將架構做最佳化，例如利用 HTTP 協定的客端/伺服端快取機制等。最後，政府提供資料的正確性與不可篡改特性十分重要，本研究目前並未針對此議題提出解決機制，但我們認為此一機制亦為未來政府開放資料服務大量上線前的必要基礎設施。

參考文獻

[1] J. Hendler, J. Holm, C. Musialek, and G. Thomas, "US Government Linked Open Data: Semantic.data.gov ," in *IEEE Intelligent Systems*, Vol. 27 , No. 3, 2012.

[2] N. Shadbolt et al., "Linked Open Government Data: Lessons from Data.gov.uk," in *IEEE Intelligent Systems*, Vol. 27 , No. 3, 2012.

[3] European Union Open Data Portal, http://open-data.europa.eu/en .

[4] Linking Open Government Data, IOGDS Statistics Page, http://logd.tw.rpi.edu/iogds_data_analytics .

[5] L. Ding, V. Peristeras and M. Hausenblas, "Linked Open Government Data," in *IEEE Intelligent Systems*, Vol. 27 , No. 3, 2012.

[6] T. B. Lee, "Linked Data - Design Issues, " http://www.w3.org/DesignIssues/LinkedData.html , 2009.

[7] A. L. Machado and J.M.P. de Oliveira,"DIGO: An Open Data Architecture for e-Government," in *Proc. IEEE International Enterprise Distributed Object Computing Conference Workshops*, 2011.

[8] T. B. Lee, "Putting Government Data online, " http://www.w3.org/DesignIssues/GovData.html , 2009.

[9] 文化資料開放服務網， http://cloud.culture.tw/opendata/ .

[10] B. V. Terrazas, L. M. V. Blazquez, O. Corcho, and A. G. Perez, "Methodological Guidelines for Publishing Government Linked Data," in *Linking Government Data*, Springer, 2011.

[11] United Nations E-Government Survey, http://unpan3.un.org/egovkb/global_reports/12report.htm , 2012.

[12] Europeana, http://www.europeana.eu/portal/ .

[13] 項靖,楊東謀,王慧茹,張榮容,許文馨,黃靜吟，*資訊分享與共榮：政府機關資料公開與加值應用*，行政院研究發展考核委員會 101 年度電子治理委外服務計畫案，2012。

[14] 行政院研究發展考核委員會，*政府資料開放加值應用研究分析書面報告*，http://www.rdec.gov.tw/ct.asp?xItem=4540507&ctNode=14813&mp=100 ，2013.

[15] E. Evan, *Domain-Driven Design*, 2005.

[16] B. Hyland and D. Wood, "The Joy of Data : A cookbook for publishing Linked Government Data on the Web," in *Linking Government Data*, Springer, 2011.

[17] R. Cyganiak, F. Maali, and V. Peristeras, "Self-service linked government data with dcat and gridworks," in *Proc. ACM International Conference on Semantic Systems*, 2010.