

以語意網技術整合網站異質資料 供鏈結資料發布

Using Semantic Web Technology to Integrate Heterogeneous Data on Websites for Publication of Linked Data

葉慶隆 Yeh, Ching-Long

大同大學資訊工程學系副教授

Associate Professor, Department of Computer Science and Engineering, Tatung University

摘要

全球網從文件網提升至資料網的過程背後，語意網標準扮演主要的概念架構角色。語意網在現有異質資料層上，以對映、轉換、包裝等程序，透過共通的詞彙和資料模型建設語意資料層，在上層提供高階服務。本文採用過去成功的方法步驟，以國家發展委員會檔案管理局網站異質資料為研究目標，將上述概念具體實現以建立系統架構。網站資料分成結構性和文字資料，依循資料的準備、儲存及發布等步驟，按部就班在各階段探討如何引進標準技術及利用開放源工具，開發系統元件。對於結構性資料，研究如何在不影響現有資料庫系統運作的情況下，即時擷取關聯資料庫資料轉成 RDF (Resource Description Framework) 格式，提供鏈結資料發布之用。非結構性資料有賴於自然語言處理中的資訊擷取應用，將非結構性資料轉成鏈結資料。知識圖譜源自語意網／鏈結資料技術，提供鏈結資料發布以外的應用，本文也探討結合知識工程方法，設計相關知識系統，開發智慧型服務。

Abstract

Semantic Web standards and technologies play the essential role of constructing conceptual architecture when enhancing the Web from document-based link to web of data. On the existing heterogeneous layer, we use procedures such as mapping, converting, and wrapping to build up the middle semantic data layer with common vocabulary and data models as the basis to provide higher-level services on the top layer. In this paper, we employ the successful method and steps, take the National Archives Administration website as the research goal and implement the above concepts to establish a system architecture. The resources on the web site are divided into structured data and textual data. Following the steps of data preparation, storing, and publishing, we investigate how to bring the standard technologies and leverage open-source tools to develop system components. For the part of structured data, we explore how to generate RDF data from relational databases for the publication of Linked Data without affecting the operation of the existing database system. For the part of textual data, we rely on the application of information extraction in natural language processing to convert textual data to Linked Data. Knowledge Graph originated from Semantic Web/Linked Data technologies provides applications other than Linked Data publication. In this paper, we also discuss the use of knowledge engineering methodology to design knowledge systems and develop intelligent services.

關 鍵 字 | 語意網、鏈結資料、異質性資料、詞彙集、RDF 資料儲存

Keywords | Semantic Web, Linked Data, Heterogeneous Data, Vocabulary, RDF Data Store

壹、前言

全球網（World Wide Web, WWW）係由超文字傳輸通訊協定（HyperText Transfer Protocol, HTTP）、超文本標記語言（HyperText Markup Language, HTML）及統一資源識別碼（Uniform Resource Identifier, URI）等諸多基礎建設所形成的文件網（Web of Document），是一種分散式資訊系統。全球資訊網協會（The World Wide Web Consortium, W3C）持續開發提供資料網（Web of Data）的標準技術，如資源描述架構（Resource Description Framework, RDF）、網路本體語言（Web Ontology Language, OWL）、SPARQL（SPARQL Protocol and RDF Query Language）及簡單知識組織系統（Simple Knowledge Organization System, SKOS）等（註 1）。這些技術一方面提供語意互用（semantic interoperability）的基礎建設，讓各方得利用共通詞彙，設計在語意層次互相連結的資料，以形成一個儲存在網上的資料庫供大家取用，即像在使用全世界最大的資料庫（Berners, 2006）。建設語意網（Semantic Web）之目標是讓電腦能夠做更多有用工作，發展網路系統間具信用的語意互動。

順著上述文件網到資料網的脈絡，語意網技術試圖在現有異質性（heterogeneous）資料的資訊系統上，經由對映（mapping）、轉換（converting）、包裝（wrapping）等程序，架設以 RDF 為基礎的語意資料層，再以此為基礎開發各種高階智慧型服務（如圖 1）。

本文以國家發展委員會檔案管理局（以下簡稱檔案局）官網（<https://www.archives.gov.tw/>）為研究個案，透過維持現有網站服務運作為基礎，探討利用語意網技術建設 RDF 語意資料層，以提供資料網上的服務介面，研究步驟說明如下：

- 一、首先探討語意網架構及其工作流程，並介紹 RDF 資料模型及語法、本體知識詞彙及 SPARQL 查詢語言，作為設計 RDF 語意資料層的基礎建設。接下來說明鏈結資料（Linked Data）技術以便讓 RDF 語意資料層和資料網接軌。最後從發布模式及工作流程的角度，進一步探討圖 1 概念架構中的具體工作內容，包括資料準備、儲存及發布等工作（Heath & Bizer, 2011）。
- 二、接續前一階段所提具體工作內容，盤點與分析檔案局官網上的異質性資料，分成結構性

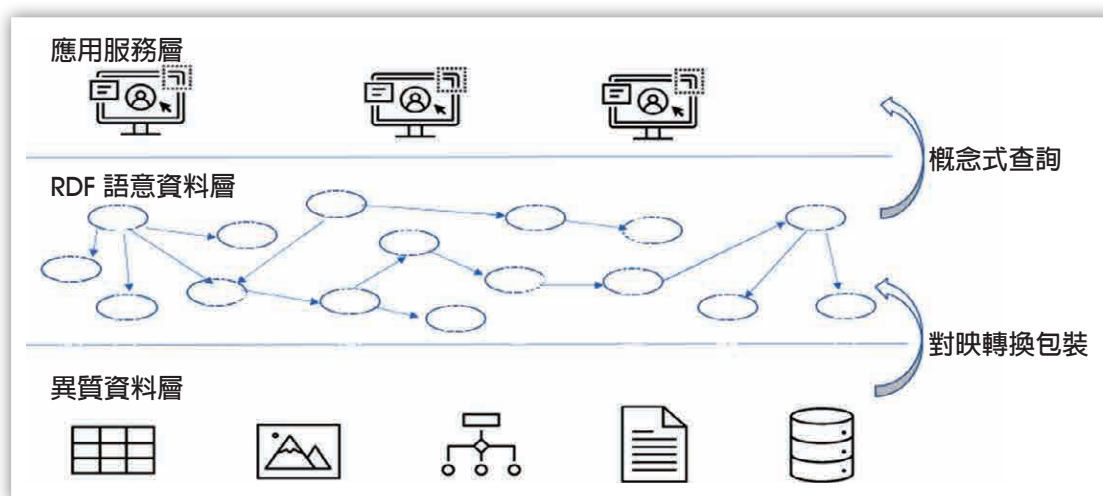


圖 1 語意網概念架構

資料來源：Herman (2012)

資料及非結構性資料 2 種類型，為接下來的
 工作流程準備（Manola, Miller & McBride, 2014）。

三、以前一階段分析結果為基礎，探討結構性資
 料的準備、儲存及發布。聚焦關聯資料庫如
 何對應 RDF 語意資料，以及在維持現有資
 訊服務的基礎上，如何搭建系統架構以完成
 鏈結資料的發布。

四、探討非結構性資料即文件的資料準備、儲
 存及發布。非結構性資料轉成 RDF 資料要
 仰賴自然語言處理的技術，尤其是命名實
 體辨識（Named Entity Recognition, NER）。
 我們先討論「Dbpedia Spotlight」計畫，其
 旨在辨識網頁中命名實體並對應適當資源的
 URI。Spotlight 的目標是鏈結資料的公布，
 本研究試著往知識圖譜（knowledge graph）
 生成的方向討論兩者之異同，以及可用在後
 者的自然語言處理技術（Ji, et al., 2022）。

五、最後探討鏈結資料的未來應用，以知識工
 程方法架構為基礎探討 RDF 語意資料集作為
 知識系統的領域知識。

貳、語意網架構及其工作流程

本節首先介紹語意網技術的基礎建設，其次
 從鏈結資料發布及工作流程角度探討實現語意網
 架構的工作項目。

語意網技術是由一些標準語言組成，包括
 RDF、OWL、SPARQL、SKOS 等，可用來建造資
 料網（Web of Data）。RDF 是用來表示語意資料
 的共通語言，其基本單位為 1 個三元組（Triple）：
 〈主詞〉〈述語〉〈受詞〉，旨在陳述關於某
 個資源（主詞）的事實，譬如以下例子陳述一些
 關於「Bob」的事實：他是人、「Alice」朋友、
 出生日期為 1990 年 7 月 4 日、對「蒙娜麗莎（Mona
 Lisa）」有興趣，最後延伸「蒙娜麗莎（Mona
 Lisa）」的創作者是李奧納多·達文西（Leonardo

da Vinci）：

<Bob> <is a> <person>.

<Bob> <is a friend of> <Alice>.

<Bob> <is born on> <the 4th of July 1990>.

<Bob> <is interested in> <the Mona Lisa>.

<the Mona Lisa> <was created by> <Leonardo da
 Vinci>.

上述三元組可表示成數學上的有向標籤圖
 （directed labeled graph），其中〈主詞〉、〈受詞〉
 分別是圖中的節點（vertex），附在加標籤有向
 端線（labeled directed edge）上的〈述語〉表示〈主
 詞〉和〈受詞〉的關係（Manola, Miller & McBride,
 2014）。配合閱讀方便，上述範例中的資源是
 以簡寫方式呈現，實際上這些資源都是表示成
 全球性唯一識別符號 IRI（International Resource
 Identifier），由於 IRI 的唯一性，所有資源的語
 意不會混淆，因此不同資料集的三元組可輕易整
 合（Kyyne, Carroll & McBride, 2014）（註 2）。

前述〈主詞〉、〈受詞〉的 IRI 符號分別
 代表實體或概念上的實體（entity），連接 2 者
 的 IRI 符號則表示彼此的關係，其意思是以電腦
 可理解的 OWL 語言所定義，稱為本體知識框架
 （ontology）或詞彙集（vocabulary）（Hitzler, et
 al., 2012），譬如 schema.org（）、SKOS 詞彙集，
 前者是通用型詞彙，後者表達概念框架的結構
 及內容，如同義詞、分類架構、主題詞表、分
 類方法等（Isaac & Summers, 2009；Google Search
 Central Blog, 2011）。OWL 建立在 RDF 語法上，
 因此各方製作的詞彙集很容易整合。詞彙集內容
 主要是定義資源類型及其性質，讀入 RDF 三元
 組時，電腦程式得據其所引用的詞彙集，做出適
 當的解讀及處理。利用命名空間（namespace）的
 定義，可以簡化每個詞彙 IRI 符號的結構以利閱
 讀（註 3）。

鏈結資料的目的是讓資料集能發布到全球網上，讓個別資料集和其它資料集的項目可相互連結，並避免資料筒倉效應（Data Silo Effect）。柏內茲·李（Berners Lee）在鏈結資料的設計上提出 4 條規則，讓資料可以發布在全球網上以融入全球資料空間中（Berners, 2006）。規則說明如下：

1. 用 URI 為全球網上的資源命名。
2. 用 HTTP URI 以便以名稱找到該資源的相關資訊。
3. 當尋找 URI 時，用標準方法（如 RDF、SPARQL）提供有用的資訊。
4. 加入其他 URI 的連結，以便找到更多資源。

在同一篇文章 Berners Lee 訂出以下的 5 星評等系統：1 星資料是放在全球網上任何格式檔案，採用開放授權條款釋出，而後的星等累加前 1 星等的特徵。2 星資料是加上機器可取用的結構資料方式釋出，如微軟（Microsoft）Excel 格式。3 星資料是使用非專屬的資料格式釋出，如逗號分隔值（Comma-Separated Values, CSV）、可擴展

標記語言（Extensible Markup Language, XML）。

4 星資料是將前一步的非專屬結構格式延伸至以 W3C 的開放標準（如 RDF、SPARQL）來指稱資料內的實體，讓眾人可以指向你的資料。5 星資料是加上資料鏈結，將資料連結到其他人的資料以建立相互關係。

藉由 5 星級區分，讓使用者能夠很快辨識資料集所屬等級，並可透過等級的區分，瞭解不同單位在開放資料上的程度，來鼓勵政府機關提供開放資料（或資料集）採開放格式以利各界查詢與取用，並逐步建立「鏈結開放資料（Linked Open Data, LOD）」，達成機器對機器識別讀取及利用。

上述語意網基礎建設和鏈結資料說明 RDF 語意資料層和高階應用服務的底層結構，接下來從鏈結資料的發布和工作流程角度，進一步探討圖 1 概念結構中的具體工作項目（Heath & Bizer, 2011）（如圖 2）。圖 2 由上而下，第 1 階段工作流程的輸入是不同資料類型，包括結構性資料和非結構性資料，接下來依序是資料準備、儲

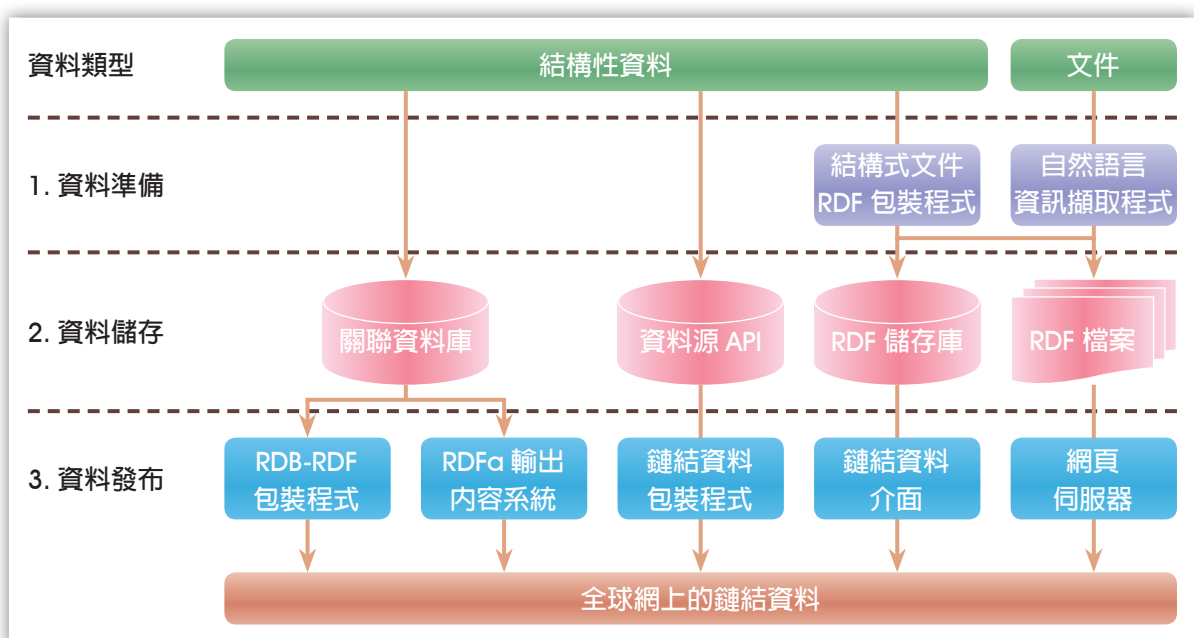


圖 2 鏈結資料發布模式及工作流程架構

資料來源：Heath & Bizer（2011）

選取	序號	全宗名	檔號	案名/案由	檔案產生日期	全文影像
☐	1	外交部	A303000000B/0049/801/0005	案 外交部施政計畫 已數位化	民國 49年 04月	

圖 3 「國家檔案檢索」搜尋結果示意圖

資料來源：作者提供

存及發布階段。針對不同資料類型在各個階段各有其工作項目，結構性資料主要是存放在關聯資料庫，透過關聯資料轉 RDF 包裝程式或包裝成 RDFa 格式（註 4），提供鏈結資料發布。另一方面，若結構性資料是透過應用程式介面（Application Programming Interface, API）所提供回應，這種情況就需要透過特殊的包裝程式提供鏈結資料。結構性資料若是結構式文件，像是 CSV 或 XML 等，就須先經過 RDF 包裝程式存入 RDF 儲存庫或是存成 RDF 檔案（Ali, et al., 2022），再提供鏈結資料發布。再來就是文件，也就是非結構性資料，一般都是利用自然語言處理程式擷取文件內的實體及關聯，再存入 RDF 儲存庫或存成 RDF 檔案，最後利用儲存庫介面或網頁伺服器提供鏈結資料發布。

參、網站異質性資料分析

根據圖 2 所示，在開始鏈結資料發布的工作流程前，先要瞭解資料類型及內容，以便進行以下的準備、儲存及發布工作。本節以檔案局官網所列的電子資料和服務項目為研究對象，盤點分析其資料類型及內容。透過官網首頁目錄式瀏覽，網頁項目繁多，本文重點放在國家檔案，其他像是機關、便民服務、國家檔案館建設及認識檔案局等內容，不在本節探討之列。國家檔案目錄提供給使用者的服務介面可分成目錄式瀏覽和

檔案檢索，以下分別從結構性資料及非結構性資料探討：

一、結構性資料

「國家檔案資訊網（<https://aa.archives.gov.tw/>）」之結構性資料來自資料庫，其首頁是國家檔案檢索服務介面，提供簡易搜尋和複合條件的進階搜尋。搜尋結果以表格呈現，有序號、全宗名、檔號、案名／案由、檔案產生日期、全文影像等欄位（如圖 3）。

點選案名／案由可得該筆資料詳細內容，除了前面欄位外還增加媒體型式、數量、內容摘要、主題、目錄公布時間、檔案典藏機關、數位化資訊及全文影像。

「國家檔案資訊網」首頁的「推薦閱讀」將相關檢索結果賦予簡短的主題式說明，提供使用者快速閱讀，點選相關項目後得到類似圖 3 的表列式結果。其他以欄位值為主的搜尋結果索引表，提供使用者各種應用。此外類似的表列式結果有「政治檔案應用專區」下的影像清單、人名索引、內容分析；「參考索引」下的戰役、人名索引。「館藏介紹」下的「本局出版品」也有以類似表格呈現的圖書文宣，每項圖書文宣品下有成對的「屬性—值」詮釋性資料，點選某項圖書文宣品可得該品項的進一步詮釋性資料。簡言之，這裡的表列式結果和「屬性—值」詮釋性資料都是資料庫的各種呈現方式。

接下來有在文件中的表格，如「館藏介紹」下的「典藏數量及內容大要」裡有 25 個典藏類別表格，這些表格採自資料庫總計結果，每一類別表格都有內容文字描述，由於文字內容屬於圖 2 的文件工作流程，因此納入接下來非結構性資料的分析討論。

二、非結構性資料

這裡的非結構性資料是指由數字、文字、語音、音樂、圖形、影像、影片及動畫等數位資料型態所組成的文件。之前提到的「典藏數量及內容大要」，整個文件可看成是資料庫內容的詮釋資料（Metadata），一開始文字說明有 25 個典藏類別，輔以統計圖表（如圖 4）。接下來每個類別以文字說明和表格成對出現，這些表格內容

是自資料庫總計結果，當作各類型典藏的詮釋資料。每個表格有相同的欄位名稱，包含：全宗名稱、內容大要、起迄期間（民國）、長度（公尺）（如圖 5）。簡言之，這種內含表格的文件，具結構性資料內涵，但以文件方式呈現。

其他非結構性資料主要出現在「主題推薦」下的「主題展示」、「探尋國家寶藏」、「影音精華」、「國家檔案精選照片」，以下分別說明。「主題展示」下羅列線上展及檔案展回顧，並提供關鍵字和日期範圍的搜尋服務。點選每一展覽項目是豐富的多媒體文件，以目錄結構方式呈現，如圖 6 臺灣糖業檔案特展。除了展覽資訊及相關商品介紹，展覽主題相關的館藏檔案有以文字描述及影音、動畫方式呈現。「探尋國家寶藏」是以主題式報導探索館藏檔案，和前面「主

國家檔案典藏現況與內容大要

檔案局自民國89年起，即延攬各領域專家學者，進行國家檔案鑑選、徵集工作，本年度亦持續徵集移轉各機關具永久保存價值之檔案，截至111年12月31日止，國家檔案典藏長度總計27,775.12公尺（紙質類檔案26,834.985公尺；非紙質類檔案940.135公尺）；此外，為便捷國家檔案之應用，讓使用者能快速瞭解檔案局所典藏之國家檔案概況，爰訂定國家檔案分類表，共計25個類別，包含府院政策、司法及法務、政治等，各類目典藏情形詳如下表：

項次	分類號	類目名稱	典藏長度(公尺)
1	010	府院政策	447.648
2	020	立法及監察	238.81
3	030	司法及法務	787.549
4	040	考銓及人事	80.575

圖 4 典藏資料庫內容的詮釋資料示意圖

資料來源：國家發展委員會檔案管理局（2023）

二、立法監察類：此類檔案總長度為238.81公尺，詳細資料說明如下表。

全宗名稱	內容大要	起迄期間(民國)	長度(公尺)
國民大會	內容包括：第二、三屆國民大會會議速紀錄及實錄等紙質檔案，以議事類檔案為主，包括議事日程、會議紀錄、會務準備、人民請願等，另有大會開會之攝錄影（音）類等檔案。	2-89	186
監察院	內容包括：民國38年以前檔案，含南京、閩台行署時期之彈劾、糾問、懲戒、建議、糾舉、糾察及行政監察等。	1-109	47.86

圖 5 典藏類型的詮釋資料示意圖

資料來源：國家發展委員會檔案管理局（2023）



圖 6 展覽項目的目錄結構

資料來源：國家發展委員會檔案管理局（n.d.）

題展示」項目不同的是，這裡是以文字加檔案圖片方式呈現。「影音精華」下的項目是影片及摘要文字說明。「國家檔案精選照片」下各主題大項有文字說明，包含精選照片摘要文字說明、檔案來源及管有機關資訊。

根據上面的分析，這裡的非結構性資料可歸納出以下特徵：首先文件中的文字是用以說明表格、圖案和影片內容。其次文件中的表格、圖案和影片內容都來自資料庫，因此文字說明的語意內涵可看成是資料庫內實體語意的延伸，只是各有不同應用領域，前者是實體的衍生闡釋，後者是作為典藏用途。

肆、結構性資料的鏈結資料發布

綜觀前一節所述，檔案局網站結構性資料主要以資料庫為主，本節探討資料庫的鏈結資料發布。根據圖 2 所示，關聯資料庫透過關聯資料轉 RDF 包裝程式或是包裝成 RDFa 格式方式發布鏈結資料，本文主要探討前者。在維持現有資料庫服務運作的情況下，另外提供鏈結資料服務，

本文參考資料庫轉 RDF 伺服器架構（Database to RDF, D2R）（如圖 7）（Heath & Bizer, 2011）。圖 7 伺服器後端由現有資料庫提供關聯資料，透過 D2R 對應檔內的規則轉成 RDF 資料，在前端以 RDF、HTML 瀏覽介面及 SPARQL 客戶介面提供使用。D2R 對應檔是對應規則庫，以宣告式規則提供 D2R 伺服器內的對應引擎（mapping engine）將關聯資料轉成 RDF 資料。

D2RQ 平臺提供其對應語言、引擎及伺服器完整技術，實現圖 7 系統架構的製作（D2RQ, n.d.）。針對關聯資料到 RDF 的對應，W3C 制定了 2 種標準：直接對應（Direct Mapping）和 R2RML（RDB to RDF Mapping Language）（Arenas, Bertails, Prud'hommeaux & Sequeda, 2012；Das, Sundara & Cyganiak, 2012）。前者直接利用關聯資料庫欄位名稱作為對應 RDF 的詞彙，因無法對應其他本體知識詞彙，使用受到限制；後者則是可以對應到其他本體知識，具客製靈活性。除了 D2RQ 平台外，R2RML 有 Virtuoso 工具支援（OpenLink Software, 2023），因此以下探討 R2RML 技術作為

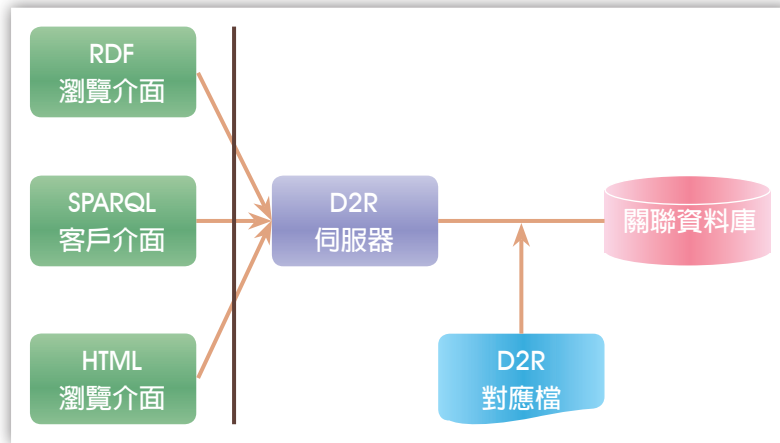


圖 7 資料庫轉 RDF 伺服器架構

資料來源：Heath & Bizer (2011)

關聯資料到鏈結資料的基礎。在介紹 R2RML 之前要先有 RDF 資源（代表網站內事物）制定命名方法，其次選擇適當的本體知識詞彙集，以下分別說明：

一、設計 URI 為事物命名

鏈結資料設計的首要規則是：使用「http://URI」為事物命名（Berners Lee, 2006；Heath & Bizer, 2011）。這可以創造全球網上唯一的識別符號，以識別各個事物，作為建立語意結構的基礎；此外透過 http 協定也可讓網站提供該識別符號的說明文字，讓人理解事物的意涵。因此在設計 URI 時要注意以下準則：不要使用無法控制的命名空間；避免參雜程式製作相關元素；使用自然鍵（natural keys），也就是鍵名要在該領域表現其意義（Heath & Bizer, 2011）。

一個資料集裡所表示的每個事物除要有 1 個 URI 可供識別外，也要以提供該事物相關資訊供人、機器解讀，分別表示成 HTML 及 RDF 文件，所以針對每件事物至少要有 3 個 URI。以 dbpedia 為例，「Computer」的識別、HTML 及 RDF URI 如下所示：

<http://dbpedia.org/resource/Computer>

<http://dbpedia.org/page/Computer>

<http://dbpedia.org/data/Computer>

這種方法的問題是：輸入第 1 個 URI，由於瀏覽器和伺服器間的互動導致重定向（redirect）到第 2 個 URI，使新接觸鏈結資料的開發人員會將 2 者混為一談。上述混淆可透過各自子領域（subdomain）解決，如下所示（註 5）：

從系統架構角度看，「data」子領域的機器負責提供 RDF 資料，「pages」子領域的機器負責提供 HTML 文件，而「id」子領域的機器處理與顧客端的重定向工作。另外在所屬概念類型下以事物名稱加附檔名也有被採用，如下所示：

http://dbpedia.org/people/Alan_Turing

http://dbpedia.org/people/Alan_Turing.html

http://dbpedia.org/people/Alan_Turing.rdf

選擇第 2 種子領域做法，只要在系統架構中增加 1 台機器扮演 D2R 伺服器角色跟「aa.archives.org」子領域的機器互動，便可從資料庫中取得資料並包裝成 RDF 格式，另一方面也取

得查詢事物的 HTML 文件描述供使用者取用，此外也處理「id」子領域重定向工作。使用資料庫中檔案的檔號當作識別符號所得 URI 範例如下所示：

<http://id.archives.gov.tw/A303000000B-0049-801-0005>

<http://pages.archives.gov.tw/A303000000B-0049-801-0005>

<http://data.archives.gov.tw/A303000000B-0049-801-0005>

二、用詞彙集豐富事物的語意

RDF 本身僅是主詞－述詞－受詞三元組所組成的概念模型，必須附加特定領域詞彙才得以彰顯三元組意涵，作為概念式查詢及相關推理知識基礎。W3C 制訂 SKOS、RDFS、OWL 3 種定義詞彙集的語言，其中 SKOS 用來定義同義詞、分類法、主題詞系統及主題架構。

RDFS 是 W3C 最先的本體知識語言，主要是定義類型（class）、性質（property）及其包含關係。OWL 是繼 RDFS 後的加強版本體知識語言，強化了 RDFS 的不足之處，有更好的語意模型表達力及推理能力。本文探討的鏈結資料用途，是以 RDFS 再加上一些 OWL 常用的敘述像是「owl:sameAs」，表示 2 個 URI 是指稱相同的資源（Heath & Bizer, 2011）。RDFS 的子類型、子性質述詞「rdfs:subClassOf」、「rdfs:subPropertyOf」配合 OWL 述詞「owl:equivalentClass」、「owl:equivalentProperty」可定義不同詞彙間的對應關係，因此提升資料集之間的語意互通性。

在語意互通性及推廣應用的考量下，儘可能使用現有詞彙集來定義 RDF 語意，常見的詞彙集如：通用型詮釋資料 DCMI、人脈網路 FOAF、社群網站 SIOC、軟體專案 DOAP、音樂 Music Ontology、節目單 Programmes Ontology、電商應

用 Good Relations Ontology、版權許可 CC、書目 BIBO、網路資源聚合 OAI ORE、評論及評級 Review Vocabulary 及地理資訊 Basic Geo（WSG84）等（註 6）（Heath & Bizer, 2011）。如果描述資料集因特定所需而開發的新詞彙，也要跟常用的詞彙集中概念相近的詞彙連結。開發新詞彙時可以使用 Protege 本體知識編輯框架（Musen, 2015）。上述詞彙集中的網路資源連結 OAI ORE 值得進一步探討，可延伸至數位策展（digital curation）之用（Abbott, 2008）。將檔案詮釋資料提升至語意模型，利於未來策展選擇、組織和管理展示項目。

相較於上述詞彙集各有其特定領域，結構化資料標記（Schema）是一個涵蓋廣泛領域並活躍的詞彙集，除了創始公司谷歌（Google）、Microsoft、雅虎（Yahoo）、央捷科斯（Yandex）的開發人員，社群也有相當積極的參與和貢獻。Google 使用 Schema 作為加結構化資料標籤的詞彙集（Google Search Central Blog, 2011）。因此可以納入此類詞彙集強化其應用性。

三、關聯資料到 RDF 的映射

R2RML 語言制定法和直接映射法主要差別在於前者可以選用詞彙，後者只能使用 RDF 和 XML Schema 詞彙（Das, Sundara & Cyganiak, 2012）。相較於直接映射法只有一套公用的對應規則，R2RML 定義詞彙提供客製化映射，其中「TriplesMap」定義一組映射規則，將邏輯表（logical table）中各筆資料映射到數個 RDF 三元組。邏輯表是針對實質資料表不同觀點所得結果，可以是一個基礎表（base table）、檢視表（view）或 SQL 查詢句。除了邏輯表當作映射源，還定義映射源的資料欄位如何形成三元組的主詞、述詞、受詞。以圖 8 為例，中間的「TriplesMaps」拿「EMP 表」當映射源。主詞映射部分是透過模板（template）的定義，將邏輯表中的主鍵值（primary key value）結合到 URI 基底當作主詞的 URI，並且指明其類型是

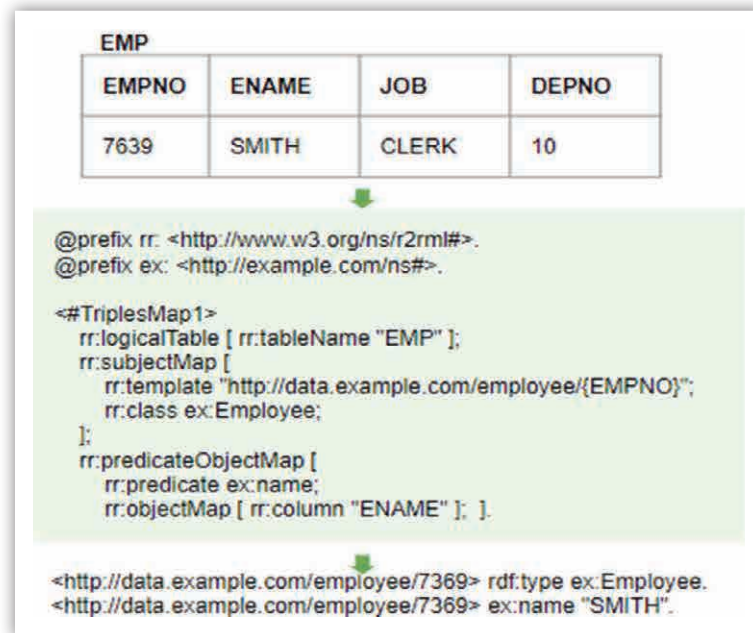


圖 8 制定 R2RML 規則將關聯資料表映射至 RDF

資料來源：Das, Sundara & Cyganiak (2012)

「ex:Employee」。接下來是述詞、物件映射規則，「EMP 表」的 ENAME 欄位名稱對應到「ex:name」。

伍、非結構性資料到鏈結資料

第參節分析網站的非結構性資料是由表格、圖案和影片及其說明文字所組成，依照圖 2 工作流程，這裡的非結構性資料要透過自然語言處理技術，擷取當中的實體、關聯以存成 RDF 檔案或存入 RDF 儲存庫，再透過服務介面提供使用。接下來探討自然語言技術應用在文件內容探勘（text mining）。

作為最早的鏈結開放資料集，「Dbpedia Spotlight」計畫開發工具自動標註出文件裡的命名實體（named entity）並鏈結到 Dbpedia 相對應的資源，換句話說將文件中辨識出來的實體標上 Dbpedia 詞彙，賦予實體語意，以便與鏈結資料結合（Mendes, et al., 2011；Daiber, Jakob & Mendes, 2013）。其做法分成 5 個步驟，以下說明之：

1. 建立辭典：Spotlight 抽取 Dbpedia 為資源建立表面形式（surface form）、同義詞、歧義形式作為後續步驟所需的辭典。另外也善加利用 wikilink，即維基百科中頁面互相鏈結的關係，根據每個資源出現的表面形式及所處的段落情境，提供混淆消除階段所需的證據來源。
2. 發現詞串：利用以辭典為基礎的分詞器，基本上這是建立在最長詞的字串比對演算法上，分詞結果也附上詞性標記（part-of-speech tagging），在這階段僅由動詞、形容詞、副詞及介系詞部分忽略不計（註 7）。
3. 挑選候選資源：前一階段分出句子中有用的詞及其詞性，接下來為這些詞標上候選資源。候選資源太多會影響效率，太少會影響涵蓋率。可以用辭典階段裡每個詞對應到資源的機率，排序候選資源，當作混淆消除的預處理。

4. 混淆消除：前階段每個詞被附上 1 組候選資源，這個階段就要應用和詞相關的情境資訊，如詞所處的段落來消除混淆，以過濾最可能的資源。Dbpedia 裡每個資源都塑造成向量空間模型（Vector Space Model, VSM）多維度詞空間的一個點。應用在資訊檢索中常用的詞頻－反向文本頻率（Term Frequency-Inverse Document Frequency, TF / IDF）比重，給候選資源的比重排序，挑選比重最高者。

5. 配置：為了讓使用者針對不同的使用案例，配置最佳的加註策略，Spotlight 提供以下指標：限定欲標註的資源集；給定資源的顯著性；設定資源的顯著性；語境歧義；歧義消除的信心指數。

原先 Spotlight 計畫僅針對英文、荷蘭文，至新版本改進上述方法步驟，以提升效率及精確度並引進多語言。技術細節可以參考原論文（Daiber, Jakob & Mendes, 2013），本文不再細究。要把 Spotlight 的經驗用在本研究上，主要困難在辭典建立。由於 Spotlight 有 Dbpedia 的辭典基礎及 Wikipedia 的資源頁面鏈結關係，為了引用上述演算法而建立類似辭典，對本研究而言是極大的挑戰，甚至不實用，因此需要尋求替代方案。

知識圖譜是 Google 2012 年提出來的概念（Singhal, 2012; Ji, et al., 2022），類似於鏈結資料。兩者都是建立在 RDF 及 URI 技術之上，以帶標籤的方式連結不同實體形成物聯網（Web of Things），其內容是關於事物的知識表示法。兩者有不同的範圍及用法，後者著重鏈結資料的發布及資料集的連結，提供使用者以 SPARQL 語法取用鏈結資料內容，可以看成是全世界最大的資料庫；前者則透過特定的資料模型來表示及組織知識，以提供各種不同的應用，如概念式搜尋、推薦等。根據上面的比較，我們可以在原先文字

資料轉成鏈結資料的路徑，納入輸出知識圖譜作為另一種選擇，好處是可以擴展原先以 SPARQL 查詢為主的鏈結資料，增加更豐富的應用。

近年結合深度學習的自然語言處理正蓬勃發展，成功的應用有文件分類、命名實體識別、機器翻譯、文件生成、文件總結、問答、聊天機器人、情感分析、語音辨識、語音轉文字等。當中命名實體識別的任務是從輸入的文件中識別並提取命名實體。譬如以 Google 雲端自然語言應用程式介面（Cloud Natural Language API）為例，當輸入檔案瑰寶（<https://www.archives.gov.tw/ALohas/ALohasColumn.aspx?c=2688>）的一個段落，會得到以下結果（如圖 9）。這個 API 辨認出段落中的實體，表面上分成 2 種：專有名詞、普通名詞，並賦予每個實體所屬的類型，如「LOCATION」，對於不清楚類型為何就傳回「OTHER」，此外也以量化方式提供實體在段落中的顯著性。

根據上述 Spotlight 的演算法架構，NER 處理好第 2 步發現詞串的部分，接下來要做的是挑選候選資源、消除混淆及配置。維基百科（Wikipedia）的 wiki 文章都是在說明特定資源，裡面也有豐富的鏈結資訊。檔案局網站上的文件和 wiki 文章不同，除了文件間的鏈結資訊不明顯外，文章的結構也大部分是反映某個目的或任務，如主題展示、主題報導等，並在文章結構中安插、引用相關資料庫中實體。所以本研究做完 NER 後，混淆消除的任務是要把文章中識別及提取出來的實體和資料庫實體連結。最後結果存成 RDF 檔案形式或者存入 RDF 儲存庫，提供使用者概念式查詢和推薦服務。除了上述商業型 NER API 外，也有其它開放源 NER 工具可用，如史坦福大學 NER、spaCy NER、NLTK NER、Allen NLP NER、Hugging Face NER 等（註 8）。



圖 9 文章段落經 NER 應用程式介面跑出的結果

資料來源：作者提供

陸、以知識工程的角度看知識圖譜的應用

前面提到以 RDF 資料層為基礎的資訊系統建構可分為領域無關及特定領域 2 方向，前者不牽涉專業領域，以 RDF 儲存庫系統及其提供的 RDF 資料管理操作、查詢為主，如 Wikidata、Dbpedia、Google 知識圖譜等；後者則要納入應用領域所需知識以建立知識系統，譬如從網路安全的非結構性資料建立知識圖譜以應用在教育方面（Agrawal, 2022）。這樣的工作首先面臨的挑戰是制定專為網路安全教育所需的本體知識框架，其次為剖析網路安全教育相關非結構性資料，以便根據框架建立知識圖譜，再來就是如何利用知識圖譜以遂行網路安全教育工作流程。這是極為複雜的知識系統，未來可以試圖利用知識工程方法論 CommonKADS，以循序漸進方式在 RDF 儲存庫系統上建立以領域知識框架為基礎的推理機制，提供具專業知識的服務（Schreiber, 2000）。

CommonKADS 是一個以模型為基礎的結構化物件導向知識工程方法論。一開始在完成問題描述後，針對問題描述進行情境層的分析，並建

立組織、任務及代理人模型，首先從組織角度，描述企業組織脈絡的營運理念、目標、實施策略與外在因素等，並列出察覺的問題癥結與改善之道。接下來針對問題工作程序，展開成一序列的任務及運用知識資產。最後對風險、成本和利益、資訊科技的支援性、專案等方面進行可行性評估。接下來依任務模型構成要素，針對前面所列任務加以細分，尤其是知識項目，如代理人模型描述任務的執行者所屬部門、負責事項和訊息之間的傳遞、會運用到的知識資產及應具備什麼特殊才能來處理某種特殊工作等項目。

對問題本質及結構瞭解透徹後，接下來進行概念層的建構，包括知識和溝通模型。前者是一個知識體系，由上而下分別是任務（task）、推理（inference）和領域（domain）3 種不同形式知識。最底層是跟某應用相關的領域知識、資訊及關聯，如醫學應用會有相關的疾病、症狀和測試及彼此之間關聯性，這相當於軟體工程的資料、物件模型。中間的推理知識說明針對領域知識的推理步驟，這是推理機器的組件，如醫學應用推理範例：假設和驗證。前者是根據症狀推導出可能的疾病，後者再經由測試結果確認觀察到的症

狀是導致疾病之原因。最上層是在探討我們的應用要達成某個目標，而這個目標可以拆解成哪些任務。總結如下，知識模型提供目標、任務、推理結構、領域知識由上往下的知識體系，以上的模型與程式設計無涉。溝通模型（communication model）是透過不同代理人進行工作的轉換和交接，此模型呈現不同代理人需共同完成某項工作時，什麼樣的資訊必須被傳遞交換或是代理人之間該如何協調運作。

最後，設計模型（design model）根據知識及溝通模型繼續建立，這是以 MVC（Model-View-Control）為基礎的系統藍圖，最終運用各種技術，將藍圖製作成可運行的系統。上述模型可視為建立一套知識管理系統所分解出來的活動，因此，CommonKADS 知識工程與管理方法論可以完成有關軟體模組與系統架構等問題。

上述步驟提供我們漸進式的方法，透過模型建立及收集分析所得特徵，作為系統架構設計的依據。有了這些系統性的方法，可以考慮為現有工作提升自動化程度。譬如前述檔案局網站之「主題展示」、「探尋國家寶藏」、「影音精華」、「國家檔案精選照片」等工作，透過知識系統的建立，循著上述步驟分析現行程序、盤點知識資產、建立知識模型及系統架構設計藍圖，一方面提升這些工作的產出，另一方面完成這些工作所

需的知識亦可保存在知識模型，促進知識的再利用及傳承。

柒、結論

本文以資料庫為基礎的資訊系統和文件管理為主的網站，從工作流程及系統架構 2 方面，探討如何透過轉換、擷取等方式建立網站內容的語意架構。一方面可作為發展高階智慧型使用者服務介面的基礎，另一方面也可以使網站加入語意網鏈結資料，與其它語意資料集建立互通的連結。除了促進語意互通外，標準架構也有利於發揮使用工具的槓桿作用，更多標準架構開發的工具可用來發展系統。

工作流程分時序和問題 2 個正交軸，從時序軸來看，我們可以清楚瞭解如何按部就班、循序解答問題。問題軸分 2 種，一種是結構性資料生成鏈結資料；另一種是將非結構性資料轉成鏈結資料。這樣的軸線交織地圖，技術上提供開發團隊方法。完成工作流程後可得語意框架及內容。未來可引進知識工程方法論，從問題分析開始，設計推理結構及所需知識框架，最後開發製作知識系統。此外，也可以嘗試利用圖神經網路（Graph Neural Network, GNN）方面的發展，提高知識圖譜的用途。

參考文獻

- 國家發展委員會檔案管理局 (2023)。典藏數量及內容大要。檢自 <https://www.archives.gov.tw/Publish.aspx?cnid=1466> (Mar. 24, 2023)
- 國家發展委員會檔案管理局 (n.d.)。臺灣糖業再出發。檢自 https://atc.archives.gov.tw/sugar/document_1.html (Mar. 24, 2023)
- Abbott, D. (2008). What is digital curation?. Retrieved from <https://www.dcc.ac.uk/sites/default/files/documents/resource/briefing-papers/what-is-digital-curation.pdf> (Mar. 24, 2023)
- Agrawal, G., Deng, Y., Park, J., Liu, H. & Chen, Y.-C. (2022). Building knowledge graphs from unstructured texts: applications and impact analyses in cybersecurity education. Retrieved from <https://doi.org/10.3390/info13110526> (Mar. 24, 2023)

- Ali, W., Saleem, M., Yao, B., Hogan, A. & Ngomo, A-C, N. (2022). A survey of RDF stores & SPARQL engines for querying knowledge graphs. *The VLDB Journal*, 31, 1-26.
- Arenas, M., Bertails, A., Prud'hommeaux, E. & Sequeda, J. (2012). A Direct mapping of relational data to RDF. Retrieved from <https://www.w3.org/TR/rdb-direct-mapping/> (Mar. 24, 2023)
- Berners, L. T. (2006). Linked data [Blog post]. Retrieved from <https://www.w3.org/DesignIssues/LinkedData.html> (Mar. 24, 2023)
- D2RQ. (n.d.). Accessing Relational Databases as Virtual RDF Graphs. Retrieved from <http://d2rq.org/> (Mar. 24, 2023)
- Daiber, J., Jakob, M., & Mendes, P. (2013). Improving efficiency and accuracy in multilingual entity extraction. Retrieved from <https://doi.org/10.1145/2506182.2506198> (Mar. 24, 2023)
- Das, S., Sundara, S., & Cyganiak, R. (2012). R2RML: RDB to RDF mapping language. Retrieved from <https://www.w3.org/TR/2012/REC-r2rml-20120927/> (Mar. 24, 2023)
- Google Search Central Blog. (2011). Introducing schema.org: Search engines come together for a richer web [Blog post]. Retrieved from <https://developers.google.com/search/blog/2011/06/introducing-schemaorg-search-engines> (Mar. 24, 2023)
- Heath, T., & Bizer, C. (2011). *Linked Data: Evolving the Web into a Global Data Space*. Retrieved from <http://linkeddatabook.com/editions/1.0/#htoc69> (Mar. 24, 2023)
- Herman, I. (2012). Tutorial on Semantic Web. Retrieved from <https://www.w3.org/People/Ivan/CorePresentations/SWTutorial/Slides.pdf> (Mar. 24, 2023)
- Hitzler, P., Krötzsch, M., Parsia, B., Patel, S. P., & Rudolph, S. (2012). *OWL 2 web ontology language primer (Second Edition)*. Retrieved from <https://www.w3.org/TR/owl2-primer/> (Mar. 24, 2023)
- Isaac, A., & Summers, E. (2009). SKOS Simple Knowledge Organization System Primer. Retrieved from <https://www.w3.org/TR/skos-primer/> (Mar. 24, 2023)
- Ji, S., Pan, S., Cambria, E., Marttinen, P., & Yu, P. (2022). A survey on knowledge graphs: representation, acquisition, and applications. *IEEE Transactions on Neural Networks and Learning Systems*, 33(2), 494-514.
- Kyryne, G., Carroll, J. J., & McBride, B. (2014). RDF 1.1 Concepts and Abstract Syntax. Retrieved from <https://www.w3.org/TR/rdf11-concepts/#section-IRIs> (Mar. 24, 2023)
- Manola, F., Miller, E., & McBride, B (2014). RDF 1.1 Primer. Retrieved from <https://www.w3.org/TR/rdf11-primer/> (Mar. 24, 2023)
- Mendes, P., Jakob, M., García-Silva, A., & Bizer, C. (2011). DBpedia spotlight: shedding light on the web of documents. Retrieved from <https://doi.org/10.1145/2063518.2063519> (Mar. 24, 2023)
- Musen, M. (2015). The Protégé project: A look back and a look forward. *AI Matters. Association of Computing Machinery Specific Interest Group in Artificial Intelligence*, 1(4), 4-12.
- OpenLink Software. (n.d.). Virtuoso R2RML Support. Retrieved from <https://vos.openlinksw.com/owiki/wiki/VOS/VirtR2RML> (Mar. 24, 2023)
- Schreiber, G., Akkermans, H., Anjewierden, A., de Hoog, R., Shadbol, N., Van de Velde, W., & Wielinga, B. (2000). *Knowledge Engineering and Management: The CommonKADS Methodology*. Retrieved from 10.7551/mitpress/4073.001.0001 (Mar. 24, 2023)
- Singhal, A. (2012). Introducing the Knowledge Graph: things, not strings [Blog post]. Retrieved from <https://blog.google/products/search/introducing-knowledge-graph-things-not/> (Mar. 24, 2023)

註釋

- 註 1 W3C，全名 World-Wide Web Consortium 即全球網協會，官網：www.w3.org，致力於開發全球開放標準，以確保全球網長期發展。RDF、OWL、SPARQL、SKOS 是 W3C 建設語意網的標準語言，其相關的活動如鏈結資料、資料網、物聯網。
- 註 2 IRI 是 URI 的擴充版，主要是加入 Unicode 字元碼。URI 是 IRI 的子集合，本文交互使用 2 者，並不影響其表示的語意。
- 註 3 RDFS 是 W3C 最早用來制定詞彙集的語言，用以定義資源類型及性質，後來被豐富的詞彙定義語言 OWL 取代。
- 註 4 這是 W3C 制訂將 RDF 資料鑲在 HTML 文件的「屬性－值」構造中，讓 HTML 文件成為裝有 RDF 資料的容器。
- 註 5 以下 URI 是本文為說明需要所虛構，此並非真實 URI。
- 註 6 這裡所列詞彙集網址如下。DCMI：<http://dublincore.org/documents/dcmi-terms/>、人脈網路 FOAF：<http://xmains.com/foaf/spec/>、社群網站 SIOC：<http://rdfs.org/sioc/spec/>、軟體專案 DOAP：<http://trac.usefulinc.com/doap>、音樂 Music Ontology：<http://musicontology.com/>、節目單 Programmes Ontology：<http://purl.org/goodrelations/>、電商應用 Good Relations Ontology：<http://purl.org/goodrelations/>、版權許可 CC：<http://creativecommons.org/ns#>、書目 BIBO：<http://bibliontology.com/>、網路資源聚合 OAI ORE：<http://www.openarchives.org/ore/>、評論及評級 Review Vocabulary：<http://purl.org/stuff/rev#>、地理資訊 Basic Geo(WSG84)：<http://www.w3.org/2003/01/geo/>。
- 註 7 原先使用 LingPipe 工具，但現已不再維護。可以改採 NLTK 的分詞及詞標記工具：<https://www.nltk.org/book/ch05.html>。
- 註 8 這裡所列開放源 NER 工具網址如下。史坦福大學 NER：<https://nlp.stanford.edu/software/CRF-NER.shtml>、spaCy NER：<https://spacy.io/api/entityrecognizer>、NLTK NER：<https://www.nltk.org/book/ch07.html>、Allen NLP NER：<https://demo.allennlp.org/named-entity-recognition/named-entity-recognition>、Hugging Face NER：<https://huggingface.co/dslim/bert-base-NER>。

「

「