以鏈結資料觀點運用於文書檔案之初探

The Preliminary Research of Documents and Records in Linked Data

蔡政威 Tsai, Cheng-Wei
英福達科技股份有限公司執行長
CEO, Infodoc Technology Corporation

壹、緒論

西元（以下同）1989 年大名鼎鼎的 Tim Berners-Lee 發明了全球資訊網（World Wide Web，簡稱 WWW）。在 Web 1.0 時代，網路形成巨大的文件庫，開始提供使用者讀的功能，瀏覽網站並被動的獲取資訊；Web 2.0 推出後，使用者可以自由地發揮寫的功能，主動將資訊放於網站上與他人分享，並藉由於網路互動方式獲取所需資訊；而現階段發展到 Web 3.0，除了能讀、能寫之外，提供可重複利用的結構化資料集，將資料互相連接，並能依瀏覽經驗自動提供使用者感興趣的資訊。如果將 Web 1.0 比喻為去圖書館，Web 2.0 像是與一群朋友談話，那 Web 3.0 的感受將是把冰冷的網站，變身為人性化的個人助理。

開放資料（Open Data）為 Web 3.0 的首要步驟，Tim Berners-Lee 也對開放資料提出五星評估架構建議（註 1、註 2、註 3）。

一、一星（On the web, open license）：使用者可以透過網路取得開放資料，也就是將檔案放上網路，檔案格式可能為 PDF、JPEG 等檔案格式。使用者並無法直接取得結構化資料，且資料被鎖在檔案中，較難將資料直接加以應用。

二、二星（Machine-readable data）：除達到一星模式外，可以透過程式來讀取結構化資料，而非直接讀取圖檔。但資料仍被鎖在檔案中，要將資料加以應用必須使用專屬軟體將檔案開啓，例如透過 Microsoft Excel 軟體讀取 Excel 檔案。

三、三星（Non-proprietary format）：除達到二星模式外，開放資料雖然仍被鎖在檔案中，但是使用開放的格式，也就是使用非專屬權（Non-proprietary）的格式（如 CSV、XML 等），讓使用者也可以透過開放原始碼或自由軟體之程式或工具來讀取開放資料加以應用。

四、四星（RDF standards）：除達到三星模式外，可使用開放標準規範如全球資訊網協會（World Wide Web Consortium，簡稱 W3C）所建議的統一資源識別碼（Uniform Resource Identifier，簡稱 URI）與資源描述架構（Resource Description Framework，簡稱 RDF）。開放資料終於擺脫只是以檔案方式擺放於網路上，而是直接呈現於網路中，使用固定網址表示開放資料，使

他人可直接連結到資料網路中的位置。

五、五星（Linked RDF）：除達到四星模式外，開放資料本身也可以再鏈結到其他資料，來做為相關內容的延伸。

Tim Berners-Lee 於 2009 年 TED（Technology, Design, Entertainment）大會提到，20 年前所建立的網際網路是將檔案連結起來，下階段將把資料鏈結起來，網際網路的發展與文書檔案的發展不謀而合。現行文書檔案擁有大量文件，並透過檔案編目建立文書關聯，符合 Web 1.0 描繪之情境。Web 3.0 亦可說是語意網（Semantic Web）的時代，語意網並非只是網頁間建立超連結，而是以電腦能理解的方式描述事物，描述事物的屬性以及事物間的關係；簡單來說，Web 3.0 著重於開放資料的同時，希望透過多種形式的資料擷取，將資料加入識別標籤，並定義資料間的關係，將資料鏈結起來，鏈結資料（Linked Data）將網際網路轉型為讓電腦可理解、資料結構化及具有語意的資料網路，並形成人與電腦都可理解的巨大知識庫，進一步方便電腦檢索與探索。現行文書檔案從資料的角度來看，可視為將一群資料經語意敘述後所產生的文件，也就是說文書檔案除詮釋資料（Metadata）外，內容也含有許多資料集合。語意網的目的是將開放資料整合與資料共享，並能重複加值應用，實現語意網最佳實務就是鏈結資料，而現行的文書檔案為結構化文件，因此易將內容轉化為鏈結資料，逐步實踐 Web 3.0。本文將以鏈結資料的觀點進行初步探討。

貳、所需技術概述

文書檔案轉為鏈結資料運用技術大致可分為資料擷取與鏈結資料兩大階段。

一、資料擷取

資料的擷取技術與來源格式息息相關，多種形式的資料擷取技術，將依文書檔案內容的不同適用於不同技術進行擷取。文書檔案已採用可擴展標記語言（Extensible Markup Language，簡稱 XML）格式進行詮釋資料，著重於標記的結構，使資料具有語法上的結構，使得資料擷取技術可採用 XML 查詢相關技術，使用詮釋資料擷取，例如 XPath 。現行詮釋資料都由人為定義，未來若需自動化產生則需參考內容資料擷取技術，其牽扯到相當複雜的技術，暫不於此探討。以下介紹文書檔案中語意部分之資料擷取技術。

（一）語意化結構文件格式

1. RDFa：RDFa 發展較早，是一個 W3C 推薦標準。它擴充了 XHTML 的幾個屬性，屬於讓電腦可以順利理解網頁內容的標籤格式。由下述例子可看出（如下頁圖 1），一段文字的內容透過語意的標籤可得知語意及內容，可輕易取得人、事、時、地、物等主題項。
2. Microdata：Microdata 是 HTML 5 的一個子集，讓電腦可以順利理解網頁內容的標籤格式，目前除了各大網站搜尋引擎（Yahoo、Google 等）皆支援 Microdata 外，連行動裝置 Siri、Google now、Cortana 等亦能支援。由下述例子可看出（如下頁圖 2），作用同 RDFa，只是語法上有差異，如 itemscope、itemtype 都是 microdata 的語法。語意化結構處理以後，搜索引擎將能夠理解 "http://www.avatarmovie.com" 不只是一個 URL，還可理解更多內容，像是此 URL 是一部科幻電影阿

```
<div xmlns:schemaorg="http://schema.org/Movie">
  <h1 property="schemaorg:name">Avatar</h1>
  <div property="schemaorg:director" itemscope
itemtype="http://schema.org/Person">
  Director: <span itemprop="name">James Cameron</span> (born
<span property ="schemaorg:birthDate">August 16, 1954)</span>
  </div>
  <span property="schemaorg:genre">Science fiction</span>
  <a href="../movies/avatar-theatrical-trailer.html"
property="schemaorg:trailer">Trailer</a>
</div>
```

圖 1　RDFa 標籤格式範例（粗體部分為 RDFa 屬性）

資料來源：作者整理

```
<div itemscope itemtype ="http://schema.org/Movie">
  <h1 itemprop="name">Avatar</h1>
  <div itemprop="director" itemscope
itemtype="http://schema.org/Person">
  Director: <span itemprop="name">James Cameron</span> (born
<span itemprop="birthDate">August 16, 1954)</span>
  </div>
  <span itemprop="genre">Science fiction</span>
  <a href="../movies/avatar-theatrical-trailer.html"
itemprop="trailer">Trailer</a>
</div>
```

圖 2　Microdata 標籤格式範例（粗體部分為 Microdata 屬性）

資料來源："How to mark up your content using microdata," *Getting Started – schema.org Website*,〈https://schema.org/docs/gs.html〉(29 Oct. 2015).

凡達，導演是 James Cameron，出生於 1954 年 8 月 16 日，並作預告片連結等關聯。

（二）非語意化結構文件格式

1. 語意搜尋（Semantic Search）：由已提供的人、事、時、地、物主題項關鍵字進行文件內容關鍵字搜尋已標註該文件的人、事、時、地、物相關資料。
2. 語意分析（Semantic Analysis）：中央研究院多年來已開發成熟的中文斷詞技術（Chinese Knowledge Information Processing，簡稱 CKIP）與中文剖析技術，可將一段文字標註出人、事、時、地、物相關資料。

二、鏈結資料

將資料逐一轉為 RDF 格式，並指使用 URI 呈現，再以 HTTP 作為傳輸管道，形成互相鏈結的資料，此概念即是 Tim Berners-Lee 於 2006 年提倡的鏈結資料，Time Berners-Lee 並提出鏈結資料的 4 項原則[註4]。

（一）使用 URI 為任何事物命名。

（二）使用 HTTP URI，人們可透過網頁檢索這些事物的名稱。

（三）當有人檢索這個 URI 時，利用標準方式（如，RDF*、SPARQL）來提供有用的資訊。

（四）這些 URI 所提供的資訊，應包含連結到其他 URI 資訊，使他們可以發現更多資訊。

以下介紹上述提及之鏈結資料技術：

（一）URI

以 URI 標示任一人、事、時、地、物主題項，並以 HTTP 做為需求者端和伺服器端之間查詢及傳送 URI 的機制，使人或電腦可以查詢特定 URI 所代表的相關資訊。對於特定 URI 所代表的相關資訊中，應包含與其他相關事物的連結的 URI，使得事物間得以鏈結。

（二）RDF

W3C 在 XML 的基礎上訂定 RDF 具有語意的資料編碼方式。RDF 主要包含 3 個部分：

1. 資源（Resources）：以 RDF 表示方式描述的物件結稱為資源，所有的資源都以一個 URI 標示。
2. 屬性（Properties）：描述資源的特性或特徵，內容包含屬性名稱與屬性內容，每個屬性都有一個意義，定義許可的值。
3. 宣告（Statements）：敘述的語句以 RDF 的格式表示。敘述基本分成 3 個部分，分別是 Subject（主詞），可以表示資源；Predicate（述詞），描述資源的屬性和定義關係；Object（受詞），可以是文字、其他的來源或屬性的值。

以下範例（如下頁圖 3、下頁圖 4）是描述 Ralph Swick 說 Ora Lassila 是 http://www.w3.org/Home/Lassila 的建立者。

（三）SPARQL 協定與 RDF 查詢語言（SPARQL Protocol and RDF Query Language，簡稱 SPARQL）

SPARQL 是一種用於 RDF 上的查詢語言，未來有機會取代搜尋引擎的技術，因為 SPARQL 更貼近人們所需的問與答。

以下範例（如下下頁圖 5）是查詢語法表達非洲國家的首都有哪些。

參、參考案例研析

案例一：政府公開文件

在開放政府的風潮下，許許多多政府機關將大量的文件與檔案上網公告，但是缺乏完善詮釋資料定義，使得資料必須由檢索人自行從內容中擷取，尚停留在 Web 1.0 的思維下。以現今社會普遍要求公開政府會議內容為例，一般大多數選擇直接把會議資訊的文件公開，另一個進階選項是公開的詮釋資料及鏈結資料（如下下頁圖 6），以現今政府大多採用結構化文件格式（例如 DI），產生其詮釋資料及鏈結資料並非難事。

```
<rdf:RDF
   xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
   xmlns:a="http://description.org/schema/">
   <rdf:Description>
      <rdf:subject resource="http://www.w3.org/Home/Lassila" />
      <rdf:predicate resource="http://description.org/schema/Creator" />
      <rdf:object>Ora Lassila</rdf:object>
      <rdf:type
resource="http://www.w3.org/1999/02/22-rdf-syntax-ns#Statement" />
      <a:attributedTo>Ralph Swick</a:attributedTo>
   </rdf:Description>
</rdf:RDF>
```

圖 3　RDF 參考範例（一）

資料來源："4.1. Modeling Statements," *Resource Description Framework（RDF）Model and Syntax Specification Website*, 22 Feb. 1999,〈http://www.w3.org/TR/1999/REC-rdf-syntax-19990222/〉(30 Oct. 2015).

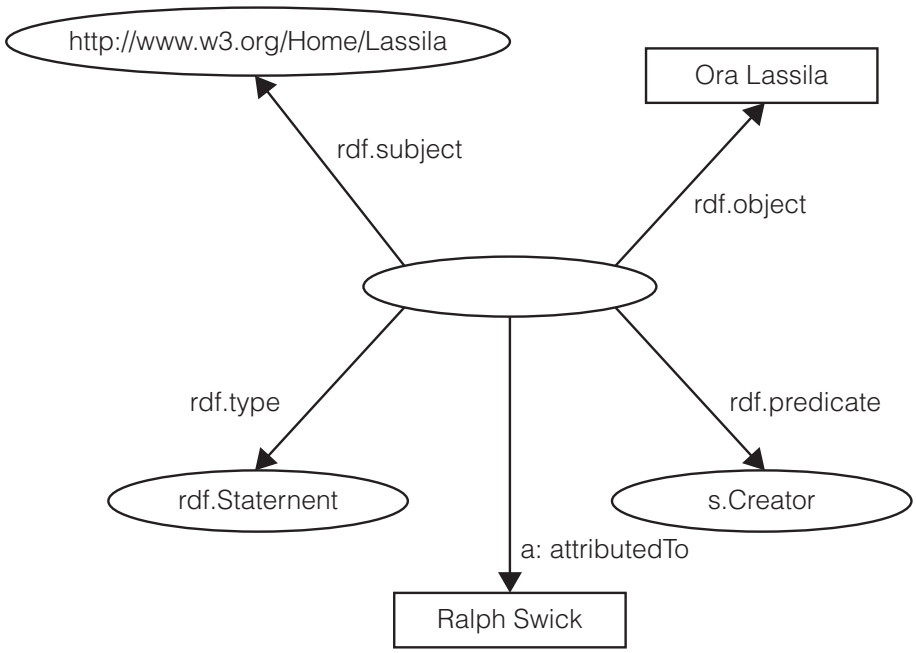


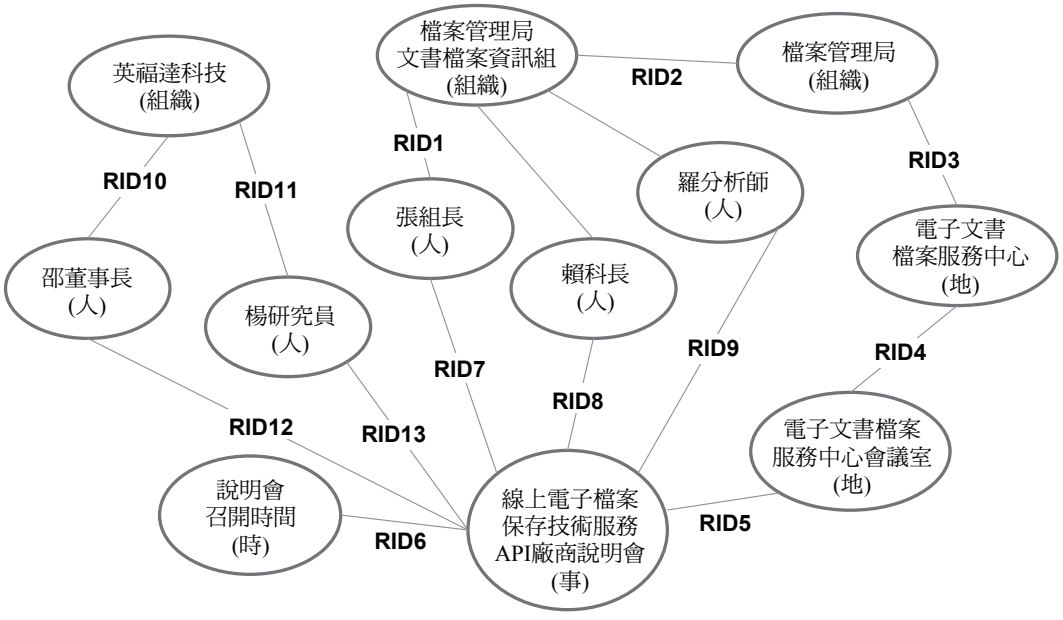
圖 4　RDF 參考範例（二）

資料來源："4.1. Modeling Statements," *Resource Description Framework（RDF）Model and Syntax Specification Website*, 22 Feb. 1999,〈http://www.w3.org/TR/1999/REC-rdf-syntax-19990222/〉(30 Oct. 2015).

```
PREFIX abc: <http://example.com/exampleOntology#>
SELECT ?capital ?country
WHERE {
?x abc:cityname ?capital ;
abc:isCapitalOf ?y .
?y abc:countryname ?country ;
abc:isInContinent abc:Africa .
}
```

圖 5　SPARQL 參考範例

資料來源：維基百科，自由的百科全書，〈SPARQL〉〈https://zh.wikipedia.org/wiki/SPARQL〉(30 Oct. 2015).



RID1：組織主管　RID2：組織內單位　RID3：會議地點　RID4：會議地點(會議室)
RID5：會議名稱　RID6：會議時間　RID7：會議主持人　RID8：出席人員
RID9：出席人員　RID10：組織人員　RID11：組織人員　RID12：廠商出席人員
RID13：廠商出席人員

圖 6　公開的詮釋資料及鏈結資料範例

資料來源：作者整理

經由上述語意網資料儲存方式，可大幅增加資料加值應用範圍，藉由知識庫關聯性強度分析技術可開發多種加值應用。

案例二：機關檔案目錄

文書檔案轉為開放資料運用，一般直覺性是將現行被動檢索的資料轉為開放資料，例

如機關檔案目錄。機關檔案目錄以詮釋資料為主，部分資料為檔案管理所需詮釋資料相關，若以鏈結資料應用為主要考量時，分類、編案及主題項最為重要，「分類」是指將檔案依其內容性質，歸入檔案分類表中最適當的類目。「編案」是將同一分類號性質相同或案情關聯之檔案，編整於同一案卷，並賦予簡要案名。「主題項」係依檔案內容，擇取足以表達案卷內容的檔案關鍵詞彙，並非文句之敘述，以提供檔案資訊多重檢索途徑，包括人、事、時、地及物等項目。上述三大詮釋資料可視作人為從內容擷取資料，並產生關聯度鏈結資料，但是因為未定義共用詞庫，表達詞彙的概念及詞彙之間的關係未統一，相關內容由檔案人員自由意識訂定，因此難以建構資料語意關聯。離鏈結資料尚有一段距離需努力。

以下 2 個範例（圖 7、下頁圖 8）將簡單呈現文書檔案標籤格式轉換成鏈結資料之範例。一般文書檔案都已有既定的標籤，將既定的標籤藉由都柏林核心集（Dublin Core）對照並轉換，例如：案由、案名及主要來文者；而內文部分使用中文斷詞技術，像是案由中的慈濟醫院及高級心臟救命術，分別鏈結至谷歌地圖（Google Map）及維基百科（Wikipeida）。

案例三：資料庫百科全書（DataBase Pedia，簡稱 DBpedia）

DBpedia 係柏林自由大學及萊比錫大學的研究人員所開啓之專案計畫，其注意到維基百科有很多資料，所以寫一個程式將資料從維基百科結構化資料提取出來，並將其放到鏈擷資料中，在網路上可以連結，稱之為 DBpedia 。DBpedia 3.9 版已含括 400 萬筆實體（Entities），其中有 322 萬筆由連貫知識本體（Ontology）進行分類，包含 832,000 位人物、639,000 個地點、116,000 份音樂專輯、78,000 部影片、18,500 個電動遊戲、209,000 個組織、226,000 個物種及 5,600 種疾病。DBpedia 共有 2,460 萬個圖片連結及 2,760 萬筆外部網頁連結、4,500 萬筆連結到其他 RDF 格式的開放資料集合、6,700 萬筆連結到維基百科的分類頁。DBpedia 使用 RDF 來呈現擷取的資料。

```
<ROW>
  <檔案目錄傳送名稱>D</檔案目錄傳送名稱>
  <功能>N</功能>
  <案由>慈濟醫院訂於 95.5.13 辦理「高級心臟救命術再認證課程」,
請踴躍報名參加</案由>
  <案名>其他</案名>
  <主要來文者>花蓮縣衛生局</主要來文者>
</ROW>
```

圖 7　文書檔案標籤格式範例

資料來源：作者整理

```
<?xml version="1.0" encoding="UTF-8"?>
<rdf:RDF
   xmlns:rdf=http://www.w3.org/1999/02/22-rdf-syntax-ns#
   xmlns:archives=http://www.archives.gov.tw/rdf/
   xmlns:dc="http://purl.org/dc/elements/1.1/">
   <rdf:Description  rdf:aboout="https://zh.wikipedia.org/wiki/高級心臟血管救命術">
      <dc:title>慈濟醫院訂於 95.5.13 辦理「高級心臟救命術再認證課程」,請踴躍報名參加</dc:title>
      <archives:案名>其他</archives:案名>
      <dc:publisher rdf:resource="http://www.gov.tw/OrgInfo/ORPF-GOV-02.aspx?oid=2.16.886.101.90021.20002.20003">花蓮縣衛生局</dc:publisher>
      <dc:relation  rdf:resource="https://zh.wikipedia.org/wiki/高級心臟血管救命術">高級心臟救命術</dc:relation>
      <dc:relation rdf:resource="https://www.google.com.tw/maps/place/%E4%BD%9B%E6%95%99%E6%85%88%E6%BF%9F%E7%B6%9C%E5%90%88%E9%86%AB%E9%99%A2%E8%8A%B1%E8%93%AE%E6%85%88%E6%BF%9F%E9%86%AB%E5%AD%B8%E4%B8%AD%E5%BF%83/@23.9957044,121.5904775,17z/data=!3m1!4b1!4m2!3m1!1s0x34689fae71fafafb:0x75db4eeadbc19057?hl=zh-TW">慈濟醫院</dc:relation>
   </rdf:Description>
</rdf:RDF>
```

圖 8　文書檔案鏈結資料範例

資料來源：作者整理

案例四：圖書館資料

圖書館目錄結構與機關檔案目錄結構擁有許多相似性，許多圖書館早已將書目公開，並提供外界資訊連結到書目紀錄，例如維基百科的相關資訊已可連結至特定書目，但是並非採有鏈結資料的方式，而是採用 Web 1.0 超鏈結的方式。但還是有少部分的圖書館扮演 Web 3.0 的先驅。

瑞典的 LIBRIS 聯合目錄利用 RDF 及 URI 連結內部的資源及外界資源（如下頁圖 9），瑞典國家圖書館自 2008 年起已可連結至維基百科的鏈結資料資料庫 DBpedia 。

參考上圖，瑞典國家圖書館將圖書館的目錄擴展到圖書館之外去。除了圖書館本身的訊

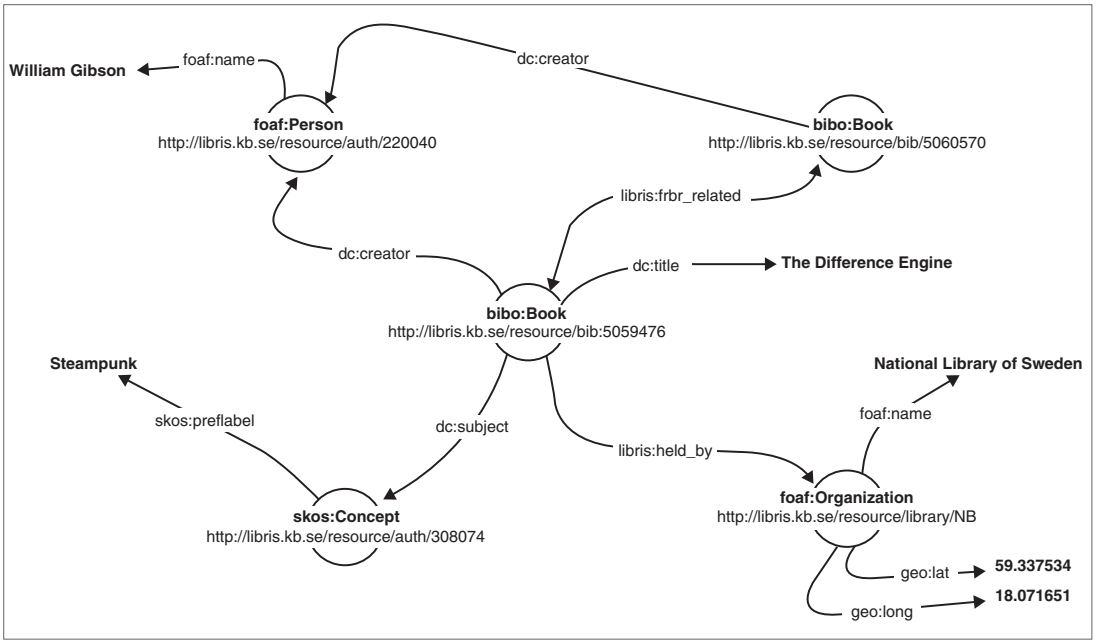


圖 9　瑞典國家圖書館公開目錄鏈結範例

資料來源："LIBRIS available as Linked Data," *LIBRIS BLOGGEN Website,* 3 Dec. 2008,〈http://librisbloggen.kb.se/2008/12/03/libris-available-as-linked-data/〉(29 Oct. 2015).

息，還鏈結至其他公開資訊，提供「關於…」（aboutness）的資訊，包含所描寫到的人、地的資訊。這可幫助讀者決定他們所需，也提供他們一個探索的起點。

肆、結語

文書檔案高度資訊化超過 10 年，內容經層層審核具高度可信賴性，文書檔案內容遠高於網路上的內容，並已產生大量結構化之文件及相關詮釋資料，若僅將詮釋資料轉為開放資料，將不足以彰顯文書檔案內容的價值。文書檔案轉為鏈結資料技術核心著重於內容資料擷取及產生鏈結資料，10 年前已有人提及文檔知識網，經過 10 年相關技術也相對成熟，因此未形成高技術門檻。相信後續再深入探究相關技術的發展，例如網絡本體語言（Web Ontology Language，簡稱 OWL）及知識本體（Ontology）包含控制語彙（Vocabulary），語意上的相互連結（Semantic Interconnection），以及在推論（Inference）和邏輯（Logic）上的規則等相關技術，提供更專業及具體之深入解析，探討如何結合本文所提之鏈結資料技術，將可逐步使文書檔案建構成完整之知識網。

註　釋

註 1："Linked Data–Design Issues," Is your Linked Open Data 5 Star?, 2010,〈http://www.w3.org/DesignIssues/LinkedData.html〉(29 Oct. 2015).

註 2：電子治理研究中心，〈政府開放資料加值營運模式之研究（委託研究報告）〉(民國 103 年 3 月)。〈http://www.teg.org.tw/web_zh/research/view.do?id=1362457511825〉(29 Oct. 2015).

註 3：開放資料的五顆星，〈範例〉(民國 104 年 8 月 31 日)。〈http://5stardata.info/zh-TW/〉(29 Oct. 2015).

註 4："Linked Data–Design Issues," *Linked Data,* 2009,〈http://www.w3.org/DesignIssues/LinkedData.html〉(30 Oct. 2015).