

# 以連結開放資料服務為基礎的 數位人文平臺建設方案研究

夏翠娟

上海圖書館高級工程師

E-mail: cjxia@libnet.sh.cn

關鍵詞：數位人文；連結資料；開放資料服務

---

## 【摘要】

計算科學在人文學科的應用促使數位人文的誕生。作為一種跨學科的研究領域，數位人文已經吸引了大量電腦科學、資訊科學與人文研究領域的學者投身其中。圖書館作為一種有著悠久歷史的人類知識收集、保存和傳播的社會機構，積累了大量符合圖書館和資訊標準的、高度結構化的資料，這些資料對於數位人文來說是至關重要的。然而，傳統的館藏資料庫並不是以數位人文學科研究的方式所構建，這使得數位人文研究者難以直接使用。本研究試圖通過使用開放連結資料的方法來改造和規範傳統的數位館藏的格式，建立一系列各自獨立又相互關聯的知識庫來解決這個問題，以為數位人文研究提供更好的支撐。這些知識庫主要包括四種基礎知識庫和一系列文獻知識庫。基礎知識庫即歷史人物規範庫、歷史地理知識庫、歷史紀年知識庫、歷史事件知識庫，分別對應著人、地、時、事這四種與文獻內容密切相關的四個維度，以連結開放資料（LOD）的方式，為多種多樣不同資料結構和資料格式的文獻知識庫提供面向內容的開放資料服務，在資料底層實現多種多媒體文獻知識庫的互聯互通，從而構建面向數位人文研究的資料基礎架構。本文以上海圖書館利用已建立的基礎知識庫提供的連結開放資料服務，實現了兩種文獻知識庫（盛宣懷檔案館知識庫和家譜知識庫）的互聯互通作為案例，詳細介紹設計開發的過程以及異構文獻資源在資料底層實現深度融合的方法和途徑。

## 引言

隨著資訊技術的飛速發展和互聯網的無時不在無處不在，除自然科學以外的人文、社會科學也變成了電腦科學（王彤彤、沈華偉、程學旗譯，2015），資料驅動型研究成為科學研究的第四範式（潘教鋒、張曉林等譯，2012），人文計算化已經成為一種潮流。在這樣的背景下，「數位人文（Digital Humanities）」成為大學和各學術研究機構、各學術期刊的寵兒，倫敦大學學院開設了數位人文專業，史丹佛大學成立了數位人文中心，國內的武漢大學也成立了國

內第一個數位人文中心。數位人文為人文研究提供了全新的方法和技術，為過去難以想像的大規模跨學科研究帶來了可能。「谷歌圖書」項目與圖書館合作，已經將人類有史以來三分之二以上的出版物都進行了數位化，基於這些海量的數位化文本，將語言學、文學、歷史學、人類學、社會學等人文科學研究置於前所未有的長時期、跨地域、多領域的廣闊環境中，以量化計算的方法，結合資料視覺化技術，揭示思想、觀念的變遷，甚至預測未來發展的趨勢。

數位人文也被一些學者認為是圖書館的未來發展趨勢，國外圖書館尤其是研究型圖書館紛紛試水數位人文，美國研究圖書館協會（ACRL）於 2011 年發起了「數位人文討論群組」，於 2012 年對圖書館是否主持數位人文中心或有數位人文專用設施建設進行了調研，調研結果表明，有 30% 的圖書館做出了肯定回答，21% 的圖書館表示正在建設之中（ACRL Digital Humanities Interest Group, 2012）。圖書館與「數位人文」有著天然的聯繫，數據資料和資訊技術是數位人文的兩大重要支柱，而圖書館正是收集、保存數據資料並提供服務的機構。經過過去 20 餘年的數位圖書館建設，圖書館積累了大量的數位化全文和元數據記錄，這是圖書館致力於提供數位人文服務的資料基礎，尤其是遵循一定標準的、高度結構化的元數據記錄，為圖書館所特有，而谷歌等依賴於關鍵字索引的搜尋引擎所不具備。近年來圖書館使用連結資料（Linked Data）技術為這些資料重新編碼，賦予其可被機器理解的語義，並發佈到互聯網上，便於與互聯網資源深度融合，更為數位人文提供了方法和技術基礎。

上海圖書館（以下簡稱「上圖」）近年來致力於研究和探索如何用圖書館領域擅長的知識組織和權威控制方法，結合互聯網時代的開放連結開放資料（Linked Open Data, LOD）技術和資料視覺化技術，對已有的元數據記錄進行重組和豐富，對數位化人文資料的全文進行文本挖掘和資料分析，建立人、地、時、事等基础性知識庫和家譜、古籍、檔案、期刊報紙等文獻知識庫，探索資料開放的模式和技術，提供有著更好用戶體驗的檢索和利用方式，以構建開放的、以資料和知識為基礎的人文研究環境，保證人文研究的持續性、一致性和高效性，實現從「藏書」到「知書」、從查閱借還到資料服務和知識服務的轉型。

## 圖書館數位人文研究現狀調研

經文獻調研發現，圖書館對數位人文的研究主要包括這樣幾個方面：一是探討圖書館在數位人文研究中的扮演的角色和職能；二是以具體的圖書館數位人文專案為案例，探討圖書館建設數位人文項目的方法和技術；三是圖書館如何為研究人員提供數位人文服務。

圖書館作為知識的保存、傳播和服務中心，而人文研究則有強烈的學科學和專業性。圖書館從事數位人文方面的相關工作，是否有其必要性和可行性？圖書館該如何著手支持數位人文？在 Supporting Digital Humanities for Knowledge Acquisition in Modern Libraries（Shepp, 2015）一書中，Shepp 認為圖書館的知識獲取的目的就是支援數位元人文，人文計算化是圖

書館的未來發展方向；McFall 研究了傳統的圖書館編目員和元數據館員在數位人文中的角色，認為他們可以幫助在數位人文專案中設計元數據方案；Aarsvold、Gonnerman、Paul 認為學科館員可以幫助本科畢業生成為人文研究學者；Fortier、James 認為圖書館與數位人文項目之間存在著天然的聯繫，數位人文專案的進行往往需要一個或多個圖書館的支持，圖書館的豐富館藏是數位人文研究的原材料，應該以方便獲取為目的向研究人員提供服務。Keller 認為圖書館的數位人文專案的可持續性發展和長期保存需要圖書館的支援，圖書館及圖書館員能夠在選擇原始資料、資源數位化、編目整理和發佈元數據、數位化存儲、專案推廣等方面發揮明顯的作用。並特別主張支援共用資料的技術架構和元數據品質也是十分重要的，應提供應用程式來支援獲得資料並跨專案使用資料，有可與使用者互動的模型和適合的介面(王寧譯，2014)。

近年來，圖書館發起和主持的數位人文項目為數不少，主要方向是把已有的數位化館藏建設成為可為人文研究服務的數位人文資料庫。然而卻存在以下幾個方面的問題：在數位元人文資料庫的建設方式上各自為政，重複建設，缺少支撐資料庫之間互聯互通的頂層框架設計；內容建設上存在著較強的專業性和學科性，不符合數位人文作為一種跨學科、跨領域研究的特性；所用的技術框架無法與互聯網很好地相容，難以實現在全網域範圍內( Web-scale ) 的開放共用。

馬凱特大學圖書館的「哥特檔案」專案就是一個典型的案例，針對原始資料的圖片掃描品質不高的問題，探索對特殊字體的圖片進行 OCR 實現文本化的技術，針對檔案的描述元數據中使用的術語缺少規範控制導致術語不一致的問題，用圖書館的權威控制方法建立了權威檔，為最終用戶提供上下文語境並為資源之間建立關聯關係提供一個連結框架 (Fortier & James, 2015)。由德國的柏林洪堡大學圖書館資訊學院主持的歐洲數位手稿專案 (Digitized Manuscripts to Europeana, DM2E) 利用連結資料技術整合、重構多個圖書館的元數據，並發佈為連結資料集，開發出支援資料處理、混搭和展示的工具，為數位人文服務 (Baierer et al., 2016)。印地安那大學伯明頓分校主持的「關聯人文專案」也採用了連結資料技術，開發了一個可應用於不同數位人文專案的連結資料平臺 LODÉ，包括資料瀏覽、資料關聯和資料提升三個部分，旨在為研究資料建立關聯關係，並與 DBPedia、Freebase 等外部資料集建立關聯 (Huber et al., 2014)。紐約公共圖書館發起了一個被稱為「紐約公共圖書館實驗室」的數位人文項目，在互聯網上進行一系列探索如何在特色館藏資料中挖掘人文研究所需資料的試驗，主要採用的是用戶協作和眾包技術 (Vershov, 2013)。

綜上所述，作為研究型圖書館，開展數位人文專案不僅可行而且也是未來發展的方向。圖書館的知識組織、規範控制方法有助於數位人文專案的進行。圖書館界經過近 20 年的數位圖書館建設，積累了大量規範化、結構化的元數據 (Metadata)，以及大量的數位化全文

掃描檔，已逐步成為圖書館構建數位人文平臺的寶貴財富。圖書館可借助連結資料技術、眾包技術、大數據技術、資料視覺化技術、社會化網路技術來構建數位人文服務。在這個過程中，應儘量採用支援資料開放、共建和共用的技術架構，避免重複建設。另一方面，需重視使用者互動，吸引使用者參與貢獻內容。而基於知識本體和語義萬維網技術的連結開放資料（Linked Open Data）技術，可以很好地滿足這些需求。連結資料技術框架具有開放性、靈活性和良好的可擴展性，它以網路技術的基礎設施 HTTP 協議為基礎，以 HTTP URI 作為資源的唯一識別碼和全球定位符，可與互聯網無縫集成，滿足開放與共用的需求；以知識本體為領域知識建模，以 RDF（Resource Description Framework）的三元組為資料的基本單元，可深入文獻內容實現細粒度化的知識組織，並圖書館的傳統規範控制延伸到互聯網環境，為構建可信網路提供內容支撐；以 W3C 的推薦標準 XML，Turtle，Json-LD 等 RDF 的序列化（Serialization）格式編碼，使得機器可理解資料的語義，有助於實現資料的深度挖掘和知識推理。連結開放資料已在圖書館領域得到了深入的應用，自 2008 年以來，英、美、韓、日等各國國家圖書館以及全球最大的圖書館資料提供者 OCLC 紛紛將書目資料發佈為連結資料。著名的「連結開放資料雲圖（Linked Open Data Cloud）」顯示：關聯開放數據技術也在多個領域得到了應用，如出版、生命科學、社會化網路、地理、政府等領域，起到了細化知識組織細微性、促進資料開放共用、支援異構資料融合的作用（Cyganiak & Jentzsch, 2014）。

本節調研的幾個充分利用連結資料技術長處的數位人文專案，很好地解決了自建數位人文資料庫中的資料與互聯網資源深度融合的問題，但仍然缺少全域性、平臺化的思維，缺少支撐資料庫之間互聯互通的頂層框架設計，因而不能很好地解決共建、共用、共享的問題。上海圖書館試圖以「互通互聯、開放共用」為目標，探索了以關聯開放數據服務為基礎的數位人文平臺建設方案，試圖建立面向數位人文研究的資料基礎架構，依託互聯網，形成數位人文研究大環境。

## 建立面向數位元人文研究的資料基礎架構

圖書館積累了大量古籍善本、家譜、私人檔案、手稿、近現代圖書、期刊、報紙等歷史文獻資源的數位化全文，其描述元數據格式大多以圖書館人熟悉的 MARC 格式作為資料檢索和交換格式提供服務。而 MARC 主要是為了描述資源的文獻特徵如題名、作者、出版、裝訂資訊等而設計的，這種面向文獻的揭示方式難以滿足學者進行深入研究的需要。人文研究人員更需要的是文獻的內容包括一些事實、知識資料等，一般表現為人、地、時、事及其相互間錯綜複雜的關係。必須進一步進行基於內容的深度加工和揭示，並提供靈活的、多維度的展示和操控工具，才能使歷史文獻資源得到更好的組織和利用。

近年來，上海圖書館開始從面向內容而非面向文獻作為出發點，對歷史文獻已有的元數據和數位化全文進行重新組織。試圖構建以人、地、時、事為綱，以各類文獻資源為目，基於文獻內容的內在關聯進行知識組織，在資料的底層建立邏輯關聯，使得綱舉而目張，讓現有的異構資源庫深度融合的系統性資料基礎架構。人（Person）、地（Place）、時（Time）、事（Event）與文獻資源（Materials）間關聯關係可如下頁圖 1 所示的抽象模型來表示，在此模型中，文獻、人、地、時、事是從現實世界中存在的文獻資源物件、人物、地點、事件、時間中抽象出來的概念，通過對概念間關係的分析建立關聯模型，以為圖書館構建各種相互關聯而非彼此割裂的知識庫提供指南。

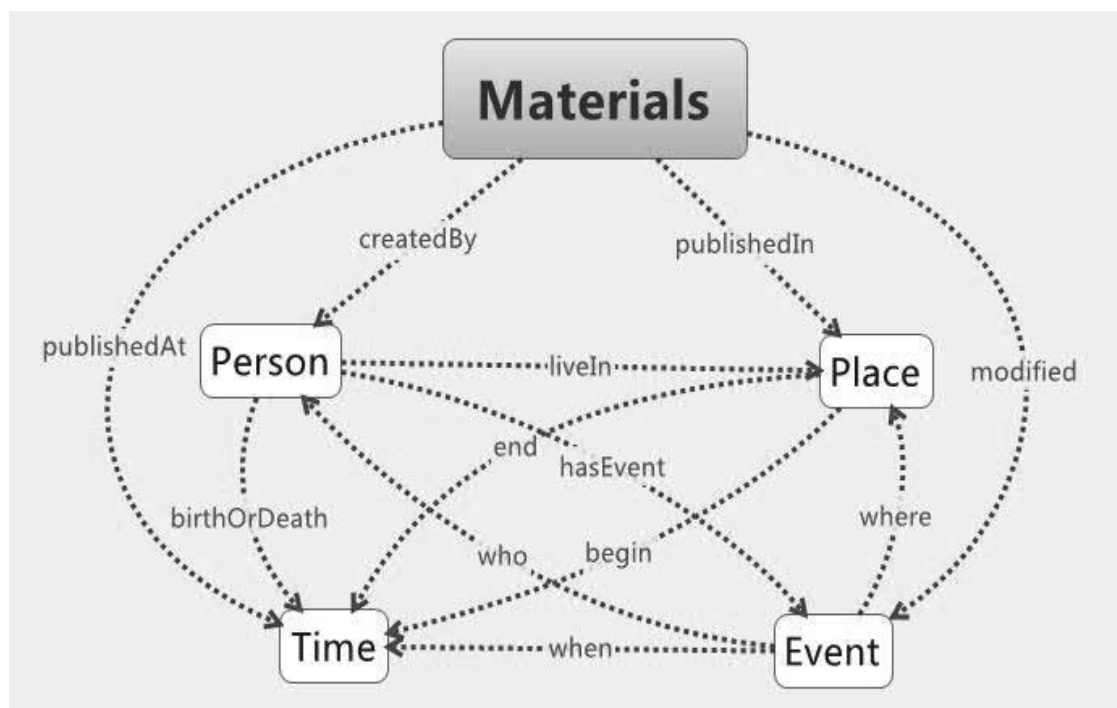


圖 1 人、地、時、事與文獻資源間關聯關係的抽象模型

基於此抽象模型，可以在圖書館的各種資源庫之間建立關聯關係。下頁圖 2 的例子試圖說明：從一個人物出發，可發現與這個人物相關的所有文獻，與文獻類型無關，而在數位圖書館時代，數位資源庫因資源類型不同而形成一個個互不相容的信息孤島。由於資源物件都以 HTTP URI 標識和定位，因而可以突破特定系統和局域網的限制。如此人作為作者創作的作品、私人檔案、筆記，拍過的照片，還是主演過的電影視頻片段，也與文獻是由哪些系統來進行保存或提供服務無關。從此人出發，還可發現其他與之相關的人物、事件以及相關的文獻，或在地理空間上的行動軌跡和分佈情況等等，其他實

體如時、地、事也是同理。在互聯網時代，一個圖書館的資源畢竟有限，而互聯網上的資源是無限的，因而還需要打通圖書館資源與互聯網資源之間的界限，使得圖書館的資源從封閉的圖書館系統中釋放出來，也能將互聯網資源整合到圖書館的資料和服務中，實現互聯互通。

首先，建設一系列基於人、地、時、事的基礎知識庫，包括「中國歷史地理知識庫」、「中國歷史紀年知識庫」、「近現代中國歷史文化事件知識庫」、「中國近現代名人知識庫」，這些知識庫是以實體為單位的，將人、地、時、事作為現實世界中真實存在或發生的，如「胡適」不再是一個字串，而是對應著歷史中真實存在的人物，也即實體（entity），是「人」這個概念的具體化和產生實體，有各種屬性（Property）來定義和描述，如姓名、性別、籍貫、生卒年、任職經歷、創作的作品等，同時還能與其他人、地、時、事等實體建立關聯。同理，「安徽績溪」則對應著現實世界中存在的地點，是「地點」這個概念的具體化，「1981年」不僅是一個時間類型的字串，而是歷史長河中真實存在的一個時間段，對應著中國歷史紀年的「清光緒十七年」。由於採用了連結資料技術，知識庫中的人、地、時、事、文獻等實體，在 RDF 資料模型中，均作為資源（rdfs: Resource），被賦予可在互聯網上唯一標識和定位的 HTTP URI，因而可在互聯網上隨時隨地被訪問，且可方便地與互聯網資源如 DBpedia 中的資源建立關聯。

其次，建設一系列文獻知識庫，如「古籍聯合目錄」、「古籍版本知識庫」、「家譜知識庫」、「近現代期刊文獻知識庫」、「名人手稿檔案知識庫」等，並使基礎知識庫與文獻知識庫有機結合，構建服務於數位人文研究的資料服務環境，試圖達到如下圖 2 所示功能：從基礎知識庫中的任一實體出發，均能將與之相關的文獻呈現在使用者介面之上，且能發現與之相關的更多的人、地、時、事，從這些人、地、時、事出發再發現更多相關的文獻資源。

同時，引入「眾包」和「眾籌」的理念，聯結使用者與圖書館資料，支援並激勵使用者貢獻知識，推動資料開放，促進知識交流，跳出圖書館領域的限制，與使用者一起共同創建以資料服務和知識服務為基礎的數位人文服務平臺。

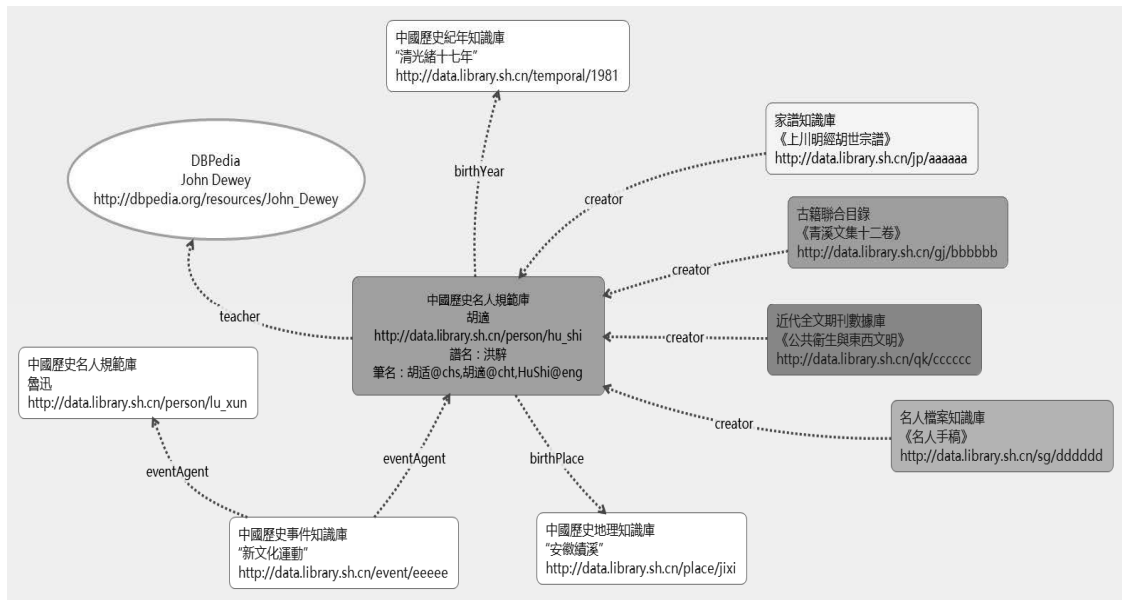


圖 2 人、地、時、事與多種類文獻關聯關係示例

## 基於連結資料的人、地、時、事基礎知識庫建設

### 基礎知識庫的建設流程

採用連結資料技術框架，遵循連結資料四原則來建設人、地、時、事基礎知識庫的建設流程包括本體設計、資料獲取、資料轉換成 RDF、連結資料發佈、開放資料消費介面的設計與開發這幾個步驟。

知識本體設計的目的是建立某個知識領域的資料模型。這是構建連結資料服務的基礎和核心，也是首要解決的問題，而領域本體模型、抽象資料模型、資料編碼格式是這個問題的三個方面。領域本體模型定義應用領域中涉及到的概念及概念間關係，是建立資料間關聯關係的依據，一般表現為適用於特定領域的知識模型，簡稱本體（Ontology）；抽象資料模型定義資料與資料間的邏輯結構，是知識組織方式的體現，根據連結資料的第三原則，資料的抽象模型要使用資源描述架構（RDF）；資料編碼格式（也叫序列化格式）決定了機器如何讀取、處理和理解資料的語義，是知識的最終表達形式。

由於人、地、時、事知識庫的資料來源多樣，包括已有的元數據、百度百科和維基百科等網路資料、CBDB 等專業資料庫、各種正式出版物如地名大辭典、名人大辭典等，因而其原始資料格式包括關聯式資料庫表、EXCEL 表格、TXT 文本等，無法直接應用於構建連結資料服務系統，因而需要轉換為 RDF 格式，如 RDF/XML、RDF/Turtle、JSON-LD 等。在資料轉換的過程中，需要提取人、地、時、事等實體，賦予 HTTP URI，並轉換為 RDF 格式。

關於實體的描述資訊根據本體定義的資料模型和資料結構，以基於 RDF 抽象資料模型來組織，並以標準的編碼格式來編碼。

構建一個連結資料服務系統需要解決資料存儲、資料查詢、資料展示的問題，因而需要掌握各種技術後通過系統開發來實現。

最後是提供服務，分別是為人提供查詢、瀏覽的介面，其中最重要的技術是資料視覺化技術，和為機器提供各種資料消費介面的服務。

## 基礎知識庫的本體概念模型

### 名人規範庫

名人規範庫是以人為實體的集合，其主要目的是實現互聯網環境下人的唯一標識、同名消歧和異名合併，並建立人與人之間的關聯關係，包括親屬關係和社會關係，以實現互聯網環境下的人名規範控制，即基於概念的匹配而非字串的匹配。如無論從人的哪一個稱謂出發，均能到達名人的唯一 HTTP URI，從而對同一個人的不同稱謂如各種筆名、別號進行歸一，對同一個人的所有相關文獻進行聚類，並區分同名不同人的情況。

圖 3 是名人知識庫本體的概念模型，試圖釐清人與檔案、手稿、家譜、書籍、報刊雜誌等文獻之間的關聯關係和人的親屬關係及社會關係；用結構化的資料描述人的不同稱謂和人的出生、婚娶、入仕、升遷、死亡等重要生平大事。通過對各種關係的建模，建立起人與時、地、事、文獻之間的關聯。

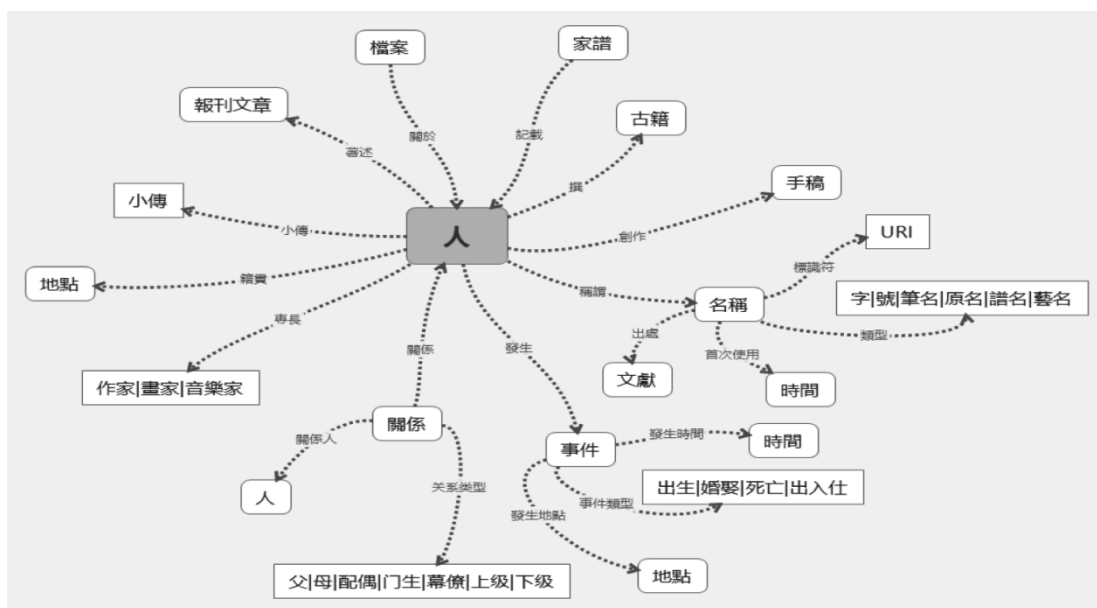


圖 3 名人本體的概念模型



## 歷史紀年知識庫

在各種歷史文獻中，大多以歷史紀年來描述時間資訊，如《龍溪盛氏宗譜》中記載「盛宣懷生於清道光二十四年」，為了時間的可排序和可計算，需要把歷史紀年轉換為以數位為表述方式的西元紀年，清道光二十四年對應的西元紀年是 1844 年，歷史紀年知識庫的主要目的是實現歷史紀年與西元紀年的對照和轉換。

圖 5 為中國歷史紀年的概念模型，主要的概念包括「朝代」和「年號紀年」，在這個概念模型中，朝代被定義為在歷史斷代研究中有明確起止年份的時間段，如「明朝(1368 年－1644 年)」，「年號紀年」被定義為與某個帝王年號相關，隨著朝代的變更、帝王的更換而變化的，有明確起止年份的時間段。以「清」這個「朝代」為例，其起止年為 1644 年至 1911 年，年號紀年「清順治」的起止年為 1644 年至 1661 年，由此可以用簡單的數學方法推算出「清順治 1 年」到「清順治 18 年」之間的任一年號紀年所對應的西元紀年，並實現歷史紀年和西元紀年之間的對照和轉換。至於「朝代」和「年號紀年」這件的關係，存在著「前」、「後」、「期間」、「相接」、「相交」等情況，可由專門的時間本體如 W3C 的 Time Ontology 來定義。

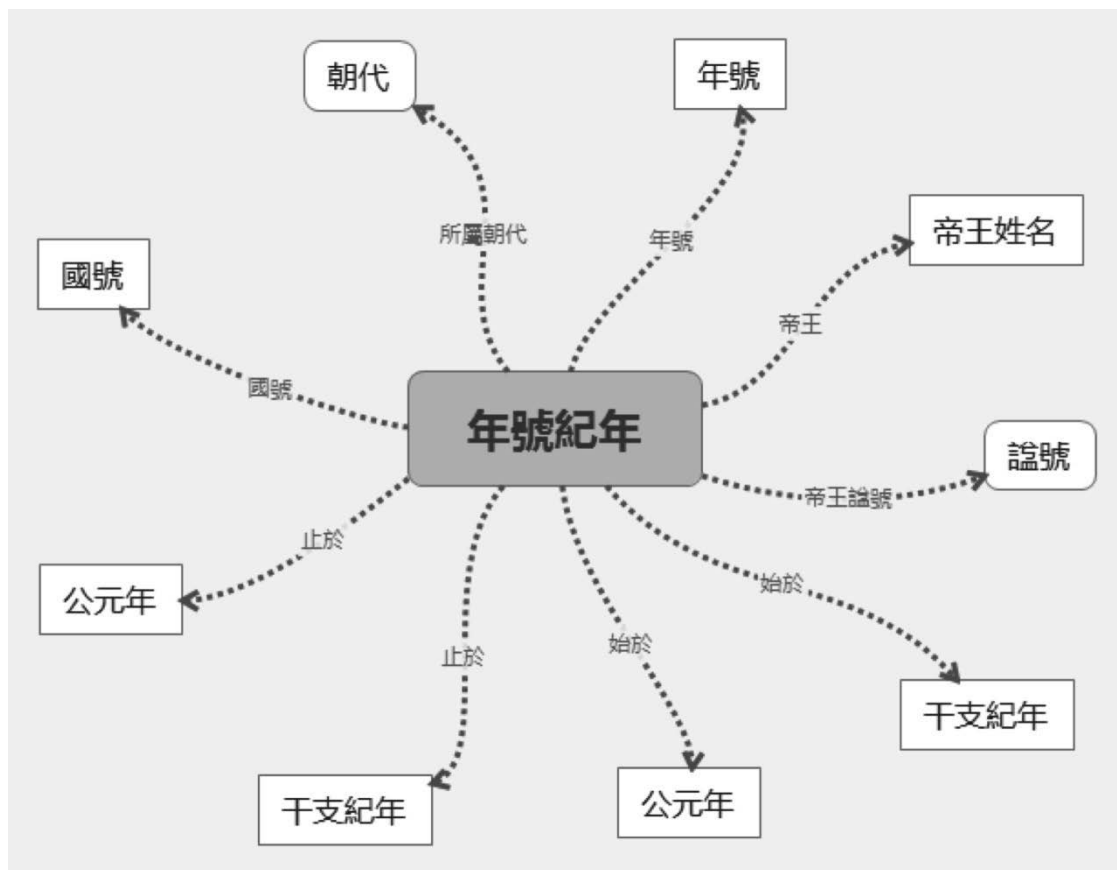


圖 5 中國歷史紀年本體的概念模型

## 歷史事件知識庫

在傳統的文獻編目中，事件往往作為表示文獻內容的關鍵字，用於字串匹配的檢索。往往缺少規範控制，導致對同一個事件使用不同的關鍵字來表示，影響文獻的查全率和查準率。歷史事件知識庫的主要目的是實現事件的規範控制和事件與人、地、時、文獻之間的關聯。歷史事件一般有一個約定俗成的名稱來標識，如「戊戌變法」、「洋務運動」、「中日甲午戰爭」，其中，人物、地點、時間作為事件的三要素，反映在事件名稱中，但不能完整準確地反映這三要素。在圖 6 所示的事件本體的概念模型中，相關人物、事件發生時間、地點作為事件的屬性值與之相關聯，事件本身則作為文獻的主題與之相關聯。建立這樣的關聯關係之後，從事件出發，就能到達與之相關聯的人、地、時和文獻資源。

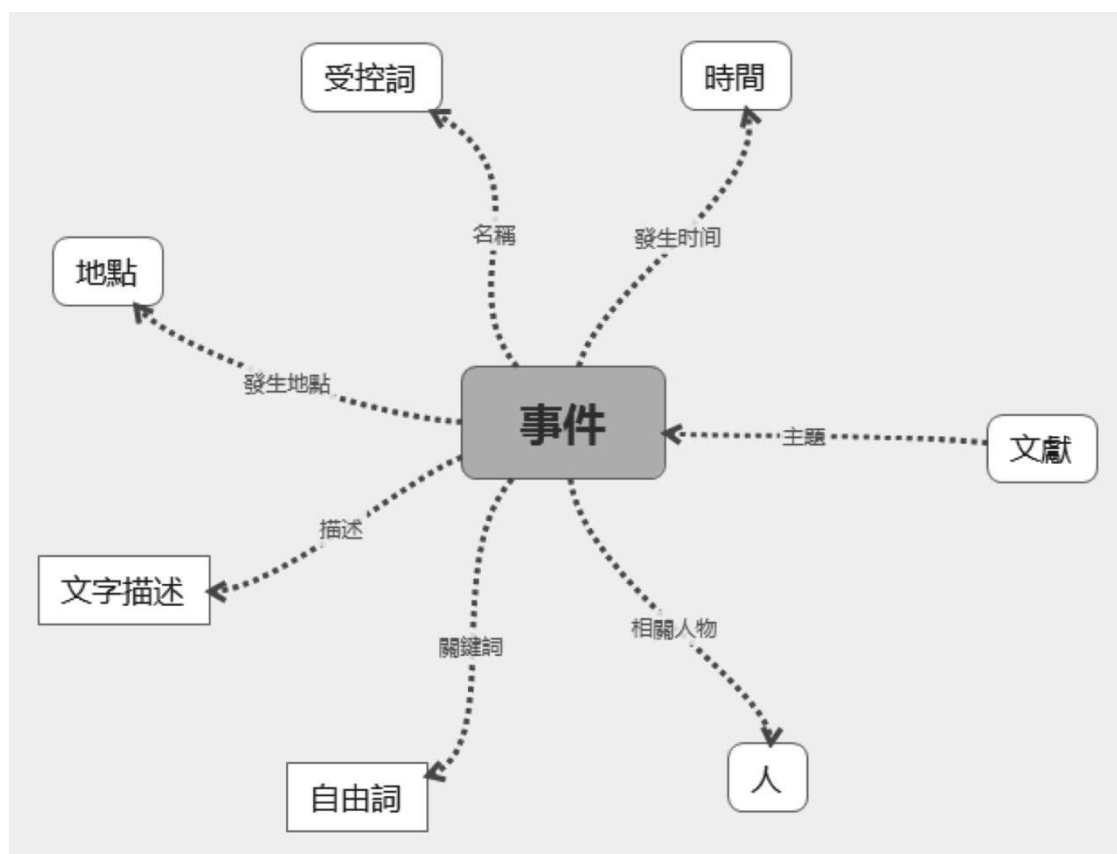


圖 6 事件本體模型

## 基礎知識庫的連接開放資料服務介面設計

作為基礎知識庫，主要目的是為其他知識庫提供資料服務，因而設計可被電腦應用程式調用的資料服務介面就顯得尤為重要，以連結資料為基礎的資料服務介面技術也被稱為連結資料消費技術。連結資料的消費介面有多種方式，如 DBPedia 和 FreeBase 等大型的連結資料集均提供 SPARQL Endpoint，Restful API，開發套件等多種多樣的資料消費介面，基礎知識庫的資料消費介面主要採用三種方式：

第一是內容協商，即訪問資源的 HTTP URI 時，可獲得關於資源的 RDF 資訊。當用普通的瀏覽器訪問時，系統返回供人閱讀的 HTML 頁面，當用語義瀏覽器或語義代理（程式）訪問 URI 時，系統按照請求方通過 HTTP Header 傳送的關於內容格式的請求返回相應格式的 RDF 資料，如 RDF/XML、RDF/Turtle、JSON-LD 等。

第二是 Restful API，是一種輕量級的 Web Service 技術框架，基於 HTTP 協定提供開放應用程式接口供程式調用，一般表現為包含各種輸入參數的 URL。首先，在調用 Restful API 之前，開發人員需在網站上註冊一個 API Key。API Key 是連結資料消費介面常用的授權控制方法，主要作用是對程式調用者進行資料獲取授權。對 API Key 的申請不設條件，申請者只需填寫姓名、聯繫方式、目的等簡單資訊即可獲得。利用 API Key 有助於對資料服務介面的調用資料進行統計和分析，如某個介面的調用頻次，成功與失敗的次數，調用次數最多的開發者等等。這些統計資料為資料服務介面的調整和優化提供了依據，也有利於發現潛在的合作者。註冊了 API Key 後，開發人員在調用時傳入參數，就會獲得該 API 所能返回的資料。參數的數量、調用方法和返回資料的結構和格式由開發人員事先定義。

第三是 SPARQL Endpoint。為熟悉 RDF 專用查詢語言 SPARQL 的開發人員調用，與 Restful API 相比，可為開發人員提供更多的靈活性。其中，Restful API 的功能最為完善，適用範圍最廣，調用也較為方便，因其返回資料格式為 JSON-LD，可被大部分流行的程式開發語言所解析。

以歷史紀年知識庫的資料服務介面為例，其主要目的是實現中國歷史紀年和西元紀年之間的相互轉換。以下為各個介面的調用方法實例：

表 1 歷史紀年知識庫資料服務介面示例

API 介面實例	功能說明
<a href="http://localhost:8080/webapi/data/明?key=YourAPIKey">http://localhost:8080/webapi/data/明?key=YourAPIKey</a>	返回明朝的西元起止年
<a href="http://localhost:8080/webapi/data/明洪武?key=YourAPIKey">http://localhost:8080/webapi/data/明洪武?key=YourAPIKey</a>	返回年號明洪武的西元起止年
<a href="http://localhost:8080/webapi/data/明洪武2年?key=YourAPIKey">http://localhost:8080/webapi/data/明洪武2年?key=YourAPIKey</a>	返回明洪武2年的西元年
<a href="http://localhost:8080/webapi/data/1369?key=YourAPIKey">http://localhost:8080/webapi/data/1369?key=YourAPIKey</a>	返回1369年所在的年號紀年

這三種資料服務介面基本可以滿足一個基礎知識庫為其它基礎知識庫和文獻知識庫提供基礎資料服務的目的。這些介面不依賴於具體的系統和平臺，而是依託互聯網的技術架構，因而其服務對象，不僅僅限於本機構，也向其他圖書館和協力廠商機構或個人開放。

## 基礎知識庫和文獻知識庫的互通互聯——以家譜和盛宣懷檔案為例

### 從人物出發實現文獻資源的互聯

在上圖的家譜知識庫和盛宣懷檔案知識庫中，實現了通過調用「名人規範庫」中的資料消費介面，實現了兩個知識庫中不同種類資源的融合和關聯。首先獲取一個人的 RDF 資料，包括 HTTP URI、姓名、生卒年、字、號、個人簡介等基本資訊，再根據這些資訊，匹配盛宣懷檔案知識庫中與之相關的人和信函、電報等檔案，同時獲取家譜知識庫中記載了該人物的家譜文獻和世系表中記載的親屬關係。

如圖 7 所示，當用戶訪問盛宣懷的 HTTP URI 時，系統返回一個 HTML 頁面，在盛檔知識庫中對信函和電報的寄件者和收件人進行匹配，找出在某個時間段內與之有通信和通電關係的人物，以視覺化的方式展示，點擊人物的頭像，即開始訪問該人的 HTTP URI，並可進入相應的 HTML 頁面。點擊表示的人與人之間的通信關係的線條，即意味著查詢線條兩端所指的人物之間的所有通信通電記錄，並展示文獻的掃描圖片。

同時，在訪問盛宣懷的 HTTP URI 時，系統也會查找家譜知識庫中與盛宣懷相關的家譜文獻，並準確定位到家譜世系表中的相應位置，展示其親屬關係。



## 從事件出發實現文獻資源的互聯

在盛宣懷檔案知識庫中，實現了通過調用歷史事件知識庫的資料消費介面，自動地實現了文獻的動態聚類。首先根據盛宣懷生活的時間範圍，從事件知識庫中獲取與盛宣懷創辦的公司有關的事件，根據該事件的人物、時間、地點等資訊，在盛宣懷檔案知識庫中獲取與之相匹配的信函、電報、合同、公司章程等文獻資源，使之以該事件為主題，聚類在一起展示，便於用戶探索發現。

下頁圖 8 展示了 1870 年代，盛宣懷所創公司之一「輪船招商局創辦」等事件，並展示與該事件相關的檔案。所顯示的檔案根據事件的發生時間 1872 年至 1873 年，事件的相關人物「李鴻章」，以及事件的相關公司「輪船招商局」作為條件來匹配，用來匹配的不是「李鴻章」和「輪船招商局」這兩個字串，而是「李鴻章」和「輪船招商局」這兩個實體的 HTTP URI，與檔案的責任者和主題詞的 HTTP URI 匹配，這樣能做到準確而快捷。

### 盛宣怀参与创办公司大事记

1870

輪船招商局创办 ▾

1872年，李鸿章为挽救中国航运业，积极筹划建立新式轮船公司，以“渐分洋商之利”，在得到清廷支持后，改其为轮船招商局，指派专员草拟章程，在上海进行筹办。1873年1月14日，官督商办性质的轮船招商局正式成立，同时标志着中国近代公司制企业的出现。

輪船招商局改名 ▾

1873年8月7日，李鸿章将轮船招商局从上海南永安街（今黄浦区永安路）迁至上海三马路新址，改称其轮船招商总局。并于同年设天津、汉口、长崎、香港等19个分局。

相关档案

輪船招商公司 輪船招商局

**【标题】** 西帮木商行事直文  
**【责任者】** 屈润培  
**【主题词】** 商务 輪船招商局  
**【档号】** SD059068

1/7 上一条 下一条

相关档案

輪船招商总局 輪船招商局

**【标题】** 李鸿章札盛宣怀文  
**【责任者】** 李鸿章  
**【主题词】** 任官 借贷 制钱 唐廷枢 天津练炮局 朱其昂 漕粮 盛宣怀 輪船招商局 运输 采办  
**【档号】** SD059551

1/19 上一条 下一条

圖 8 盛宣懷參與創辦公司大事記及相關檔案聚類展示

## 在時間和空間範圍內實現文獻資源的互聯

在家譜知識庫和盛宣懷檔案知識庫中，都通過調用歷史紀年知識庫和歷史地理知識庫的資料消費介面，實現了在時間和空間範圍內的文獻資源互聯和視覺化展示。根據文獻的修撰時間，與文獻相關的地名，作為參數調用歷史紀年知識庫和歷史地理知識庫的資料消費介面。如果文獻的修撰時間是歷史紀年，就獲取其對應的西元紀年，反之亦然；如果地名是古地名，就獲取其對應的今地名及其經緯度數據。當使用者操作時間軸或地圖時，根據時間和地名的匹配，在地圖上相應的位置展示文獻。

圖 9 展示了 1880 年在中國哪些地方新修了家譜，一個紅色的旗標對應一個地點，點擊旗標，可展示 1880 年此地所有新修家譜文獻的連結，點選連結可查閱該家譜的詳細書目資訊和館藏資訊，如有掃描的全文圖片也可閱覽。用戶可以利用右邊的時間旋鈕來確定時間範圍，通過地圖來確定空間範圍，對於那些不知道自己祖先的明確居住地的用戶起到了導航的作用。

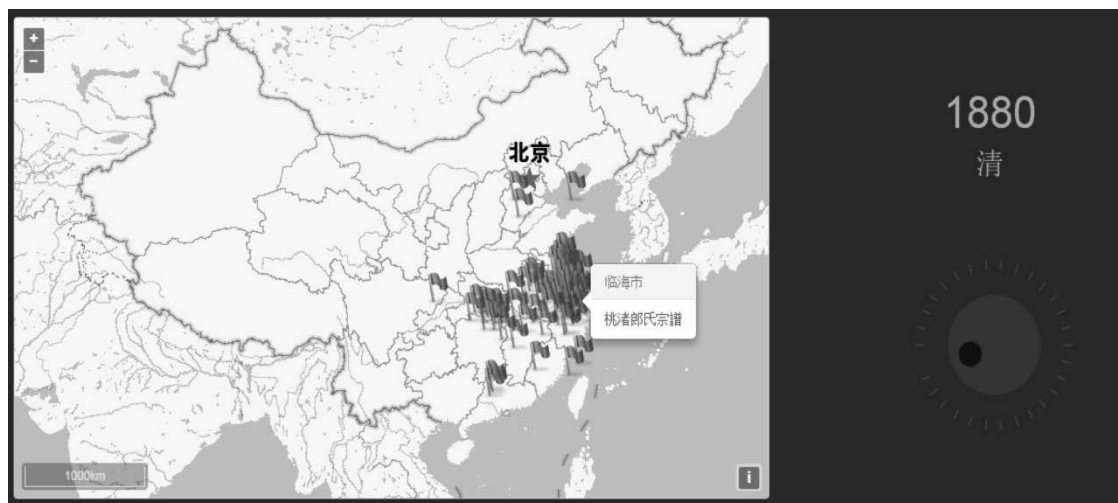


圖 9 家譜文獻在時空範圍內的視覺化展示

## 從文獻資源間的關聯關係實現人物間的互聯

在盛宣懷檔案知識庫中，通過檔案間的關聯關係實現了人物間的互聯。盛宣懷檔案中數量最多的是與當時政界和商界要人之間的來往信函和電報，有 11 萬餘件，通信通電的時間、地點和次數可以在某種程度上反映人物的歷史地位和人物之間的關聯關係。圖 10 展示了盛宣懷的通信通電關係，每一個圓點代表一個人，點擊可查閱該人的基本個人資訊，這些資訊通

過調用「名人規範庫」的介面所得，有向線條代表兩個人之間的通信通電情況，滑鼠懸停在線條上方，系統會顯示兩人之間的通信通電次數和主要內容。這是一個全景圖，可以發現其中與盛宣懷和盛康通信的人是最大的。

### 信函电报收发关系图

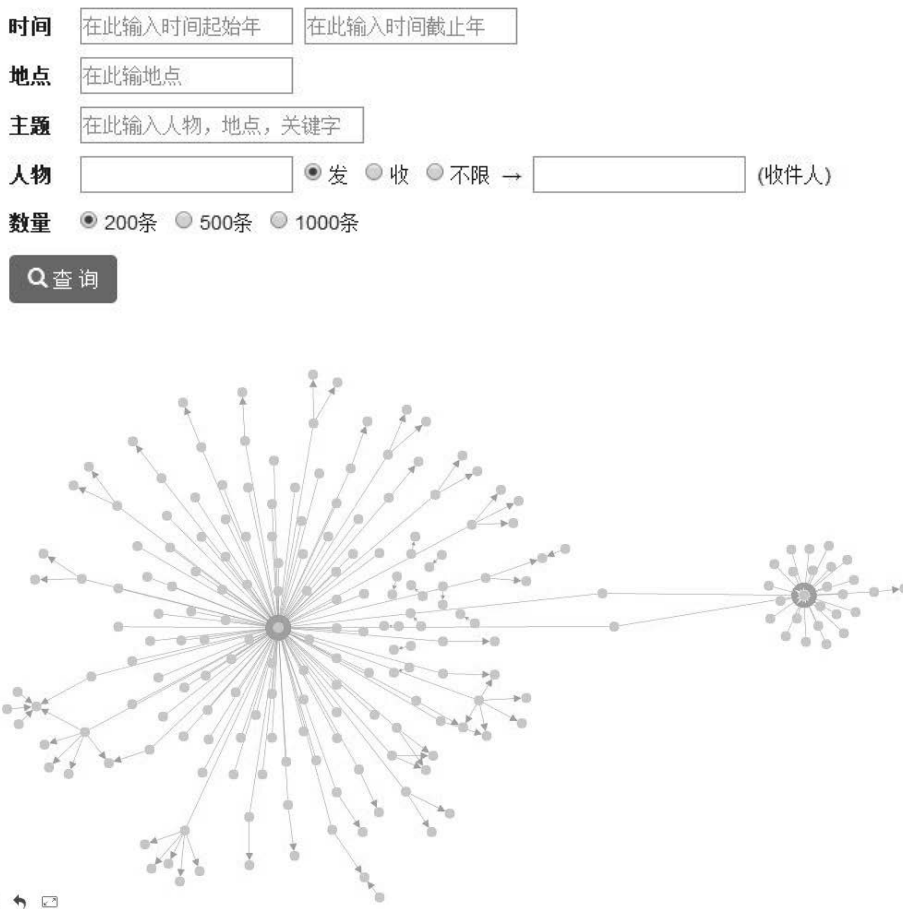


圖 10 盛宣懷檔案中的人物通信通電關係圖

## 結語

本文在對圖書館界的數位人文研究文獻和數位人文項目進行調研的基礎上，認為當圖書館的數位圖書館建設達到了一定規模，當圖書館中的文獻都有了相應的數位化版本，並擁有專業圖書館員參與編制的、基於一定標準規範的、高度結構化的元數據，就為數位人文奠定了基礎。數位人文不僅是大型研究性圖書館的發展趨勢，也是滿足使用者多樣化的資料服務

需求的需要。資料和技術是數位人文的兩大支柱，圖書館作為千百年來的知識保存、組織和傳播中心，其知識組織和規範控制方法，結合連結資料技術後，在互聯網環境下仍有用武之地，可為數位人文貢獻資料和方法。上圖近年來進行的「數位人文平臺」建設，就是試圖對這個觀點進行的嘗試和驗證。

整個「數位人文平臺」的建設，從較有特色的館藏「家譜」開始，構建了基於連結資料的「家譜知識庫」和「連接開放資料服務平臺」。前者為尋根問祖的普通大眾和研究人員提供基於概念匹配的家譜文獻查閱服務和基於人、地、時、事等實體間關聯關係的資料視覺化服務，後者為機器提供人、地、時、事等基礎資料和家譜文獻資料服務，供程式調用，該平臺在2016年4月舉行的「家譜應用開發競賽」中取得了廣泛的社會影響。「家譜知識庫」和「連接開放資料服務平臺」統稱為「家譜知識服務平臺」，該平臺的推出，為普通大眾、究人員和協力廠商機構提供了更為優質的服務，也為連結資料在圖書館的應用探索了整體技術實現方案，在家譜研究界和圖書館界均取得了較好的影響。

以「家譜知識服務平臺」為基礎在接下來的數年時間中全面展開「數位人文平臺」建設，需要從更為宏觀的層面來考慮，作為圖書館，如何構建面向數位人文的研究環境。本文提出在建設文獻知識庫的同時，建設以人、地、時、事等基礎知識庫的方法和路線。以人、地、時、事為綱，各類文獻知識庫為目，各知識庫以自己的資料為基礎為其他知識庫提供連接開放資料服務，在資料底層實現知識庫之間細粒度化的互通互聯，為人文研究人員提供全景式的知識服務。「盛宣懷檔案知識庫」是繼「家譜知識庫」之後的又一個文獻知識庫，初步實現了人、地、時、事與文獻間的關聯，通過人物之間的關係，實現了「家譜知識庫」和「盛宣懷檔案知識庫」的互通互聯。

在「基礎知識庫」的建設上，借鑒了軟體工程中的「原型法」，在人力、財力、技術條件一步步到位的情況下，逐步豐富資料、調整知識本體、開發更多的資料消費介面，而連結資料技術框架的良好可擴展性和靈活性，足以支援這種方式。目前基礎知識庫仍在完善之中，「名人規範庫」缺少明以前的古代人物，「歷史紀年知識庫」缺少秦以前的資料，將來也會納入日本等歷史紀年資料，「歷史地理知識庫」正在導入與時間相關的地理資料，「事件知識庫」也只涉及到近現代，這些都將在接下來與更多的「文獻知識庫」同步建設。

## 參考文獻

- ACRL Digital Humanities Interest Group. (2012). 2012 dh+lib survey results [Online discussion group]. Retrieved from <http://acrl.ala.org/dh/about/2012-dhlib-survey-results/>
- Baierer, K., Dröge, E., Eckert, K., Goldfarb, D., Iwanowa, J., Morbidoni, C., & Ritze, D. (2016). DM2E: A Linked Data Source of Digitised Manuscripts for the Digital Humanities. *Semantic Web – Interoperability, Usability, Applicability an IOS Press Journal*, 2014(1), 1-13. doi: 10.3233/SW-160234 Retrieved from <http://www.semantic-web-journal.net/system/files/swj831.pdf>
- Cyganiak, R., & Jentzsch, A. (2014). *The Linking Open Data Cloud Diagram*. Retrieved from <http://lod-cloud.net/>
- Fortier, R., & James, H. (2015). Becoming the Gothic Archive: From Digital Collection to Digital Humanities. In Kathleen L. Sacco, Scott S. Richamond, Sara M. Parme & Kerrie Fergen Wilkes (Eds.), *Supporting Digital Humanities for Knowledge Acquisition in Modern Libraries* (pp. 196-213). Oregon, OR: Ringgold Inc.
- Huber, J., Szt Tyler, T., Noessner, J., Murdocket, J., Allen, C., & Niepert, M. (2014, September). *LODE: Linking Digital Humanities Content to the Web of Data*. Proceedings of the 14th ACM/IEEE Joint Conference on Digital Libraries, London, United Kingdom. Retrieved from <http://arxiv.org/pdf/1406.0216v1.pdf>
- Shepp, M. (2015). Digitizing the Humanities: A Future for Libraries. In Kathleen L. Sacco, Scott S. Richamond, Sara M. Parme & Kerrie Fergen Wilkes (Eds.), *Supporting Digital Humanities for Knowledge Acquisition in Modern Libraries* (pp. 1-44). Oregon, OR: Ringgold Inc.
- Vershbow, B. (2013). NYPL Labs: Hacking the Library. *Journal of Library Administration*, 53(1), 79-96. doi: 10.1080/01930826.2013.756701
- 王彤彤、沈華偉、程學旗（譯）（2015）。**可視化未來：數據透視下的人文大趨勢**（原作者：Erez Aiden, Jean-Baptiste Michel）。中國浙江省：浙江人民出版社。（原著出版年：2013）
- 【Aiden, Erez & Michel, Jean-Baptiste (2015). *Uncharted: Big Data as a Lens on Human Culture*. (Wang, Tung-Tung, Shen, Hua-Wei & Cheng, Hsueh-Chi Trans.). Zhejiang, China: People 's Publishing House. (Original work published: 2013).】
- 王寧（譯）（2014）。數字人文和計算化社會科學及其對圖書館的挑戰（原作者 Michael A Keller）。**現代圖書情報技術**，2014（10），1-3。
- 【Keller, Michael A (2014). (Wang, Ning Trans.) *Shutz renwen he jisuanhua shehueikeshiue ji chi duei tushuguan de tiaujan*. *New Technology of Library and Information Service*, 2014(10), 1-3.】
- 潘教峰、張曉林等（譯）（2012）。**第四範式：數據密集型科學發現**（原作者 Tony Hey, Stewart Tansley, Kristin Tolle）。中國北京市：科學出版社。（原著出版年：2009）
- 【Hey, Tony, Tansley, Stewart & Tolle, Kristin (2012). *The Fourth Paradigm: Data-Intensive Scientific Discovery*. (Pan, Chia-Feng & Chang, Hsiao-Lin, et al. Trans.). Beijing, China: China Science Publishing & Media Ltd. (Original work published: 2009).】

# ***Building a Digital Humanities Platform by Using Linked Open Data Services***

**Cuijuan Xia**

Senior Engineer, Shanghai Library, China

E-mail: [cjxia@libnet.sh.cn](mailto:cjxia@libnet.sh.cn)

Keywords: Digital Humanities; Linked Data; Open Data Service

---

## **【Abstract】**

The computational applications in humanities prompted the debut of Digital Humanities. As an interdisciplinary research field, It is drawing an increasing interest from scholars in computer science, information science, and humanities. Libraries as a type of social institution have a long history of collecting, preserving and spreading the knowledge of mankind and have cumulated a vast amount of highly structured data conforming to library and information standards. These data are fundamentally important for digital humanities. However, traditional digital collections in libraries are not built in the way that digital humanities research requires, which makes it difficult for digital humanities researchers to use them directly. This study is to address this problem through using the Linked Data approach to build knowledge bases in transforming and normalizing traditional digital collections into the format that can be easily deployed by digital humanities research. The pilot of four knowledge bases have been developed for people, places, time, and events respectively: Historical People Authority Control Database, Historical Geography Knowledge Base, Historical Chronology Knowledge Base, and Historical Events Knowledge Base. These knowledge bases form the content infrastructure for us to provide Linked Open Data (LOD) services, which enables sophisticated searches and uses of document resources knowledge base with multiple types of documents and multimedia, instead of digital collections with only a keyword search function. Based on the LOD services provided by the knowledge bases at Shanghai Library, two literature resources knowledge bases were developed to demonstrate the feasibility of LOD services: the Genealogy Knowledge Service Platform of Shanghai Library and the Sheng Xuanhuai Archives Knowledge Base of Shanghai Library. The process of design and development as well as the way through which resources are interlinked are described in detail in this paper as a case study.

## **【Long Abstract】**

### **Research Background**

In recent years, Shanghai Library (SL) has devoted itself to the research and exploration of how to combine the knowledge organization and normative control methods that the library field is good at with Linked Open Data (LOD) technologies and data visualization technologies in the Internet age. The purposes and motivations include: (1) to reorganize and enrich existing metadata, perform text mining and data analysis on the full texts of digitized humanities data, (2) to establish literature databases of basic knowledge bases, genealogy, ancient books, files, and journals & newspapers of person, place, time, and event, (3) to explore the models and technologies of data openness, (4) to provide better retrieval and utilization methods of user experiences by developing an open data and knowledge-based humanities research environment, (5) to guarantee the continuity, consistency, and high efficiency of humanities research, and (6) to achieve transformations from “collecting books” to “knowing books” and from accessing and lending books to data services and knowledge services. Besides, Linked Open Data technologies based on knowledge ontology and Semantic Web technologies can help achieve the in-depth data mining and knowledge reasoning and to meet said needs.

### **Methodology**

This study used LOD-based methods to transform and regulate the formats of traditional digital archives and develop a series of mutually independent and associated knowledge bases, including four basic knowledge bases and a series of literature knowledge bases. Basic knowledge bases include historical figures and norms knowledge base, historical geography knowledge base, historical chronology knowledge base, and historical event knowledge base, which correspond to the four dimensions that are closely related literature content – person, place, time, and event. This study used LOD-based methods to develop the dimensions and basic data structure of digital humanities research. The abstract model of associative relationships among Person, Place, Time, Event and Materials is shown in Figure 1.

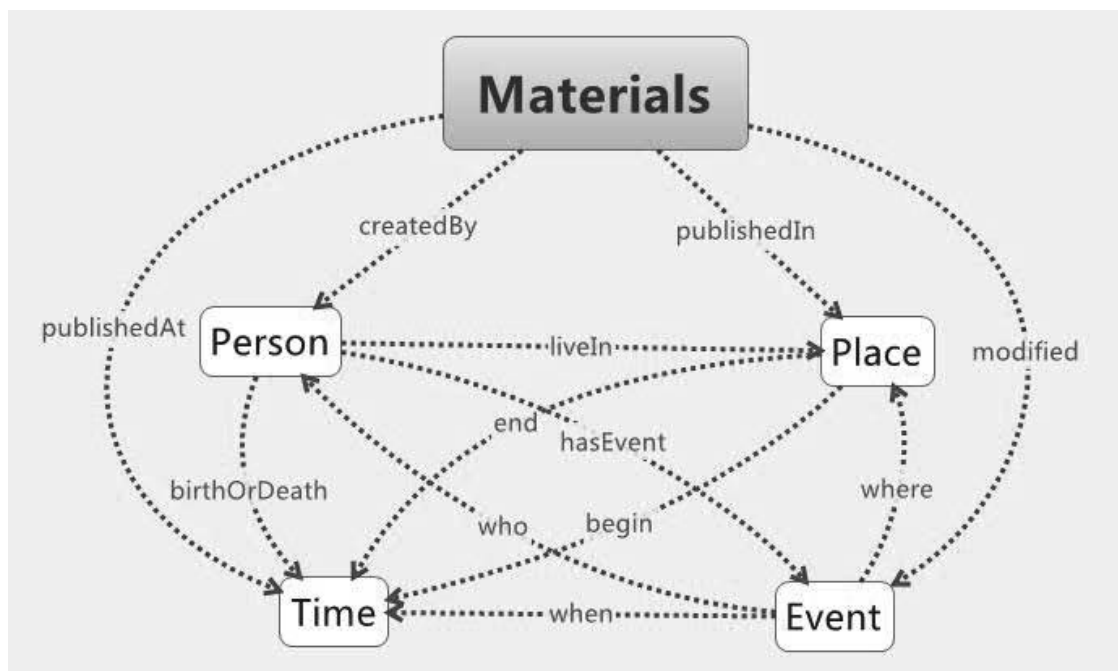


Figure 1 **Abstract Model of Associative Relationships among Person, Place, Time, Event, and Materials**

First, this study developed a series of basic knowledge bases based on person, place, time, and event, including “Chinese historical geography knowledge base,” “Chinese historical chronology knowledge base,” “modern Chinese historical and cultural event knowledge base,” and “modern Chinese celebrity knowledge base.” These knowledge bases use entities as units, and view person, place, time, and event as objects actually existing or taking place in the real world. For example, “Hu Shih” is no longer a character string, but corresponds to a figure that actually exists in history. In other words, the entity is the specification of the concept of “person” and generates an entity with various properties as definitions and descriptions, such as name, gender, native place, year of birth and death, work experience, and works created, which can establish an association with other entities, such as person, place, time, and event.

Secondly, this study developed a series of literature knowledge bases, such as “ancient books Joint Catalogue,” “ancient books version knowledge base,” “genealogy knowledge base,” “modern journals and literature knowledge base,” and “celebrity file knowledge base,” organically combined basic knowledge bases with literature knowledge bases, and established a service environment of digital humanities research data.

## Results

This study took Shanghai Library, who used the LOD-based services provided by the developed basic knowledge bases to achieve the interconnection between 2 literature knowledge bases (Sheng Xuan Huai Archives Knowledge Base and Genealogy Knowledge Base), as an example, to introduce in detail the design and development processes, as well as the methods and approaches for processing heterogeneous literature resources, namely, achieving in-depth integration.

### Ontology concept model of basic knowledge base

#### 1. Celebrity names knowledge base

Celebrity names knowledge base is the collection containing persons as entities. Its main purpose is to achieve the unique identification, disambiguation, and synonym of person in the Internet environment, as well as to develop the associative relationships between person and person, including kinship and social relationships, in order to control celebrity names in the Internet environment. In other words, concept-based matching, instead of character string-based matching, can develop the associations among person, time, place, event, and literature through the modeling of various relationships.

#### 2. Historical geography knowledge base

Historical geography knowledge base is the collection of places and place names. Its main purpose is to achieve the normative control of place names and comparison between ancient place names and current place names and provide geographical space data services, such as latitude and longitude, to other basic knowledge bases and literature knowledge bases, in order to facilitate the positioning on maps and provide the service of data visualization of geographical spaces.

#### 3. Historical chronology knowledge base

Among various types of historical literature, historical chronology is mainly used to describe time information. To make time sortable and computable, historical chronology has to be converted into Common Era where numbers are used as a way of expression. The 24th year of the reign of Qing Emperor Daoguang corresponds to 1844 A.D. The main purpose of historical chronology knowledge base is to achieve the comparison and conversion between historical chronology and Common Era.

#### 4. Historical event knowledge base

The main purpose of historical event knowledge base is to achieve the normative control of events and development of associations among event, person, place, time, and literature. In general, a common name is used to mark a certain historical event, such as “Hundred Days' Reform,” “Self-Strengthening Movement,” and “First Sino-Japanese War” where person, place, and time are used as the 3 essential

elements of an event and are reflected in the name of event.

### Design of LOD Service Interface of Basic Knowledge Bases

Linked Data-based data service interface technologies are also called Linked Data consumer technologies. Many methods are applied to Linked Data consumer interface. For example, large-scale associated data sets, for example, DBpedia and FreeBase, both provide multiple and diversified consumer interfaces, such as SPARQL Endpoint, Restful API, and development packages. 3 methods are mainly used for the data consumer interfaces of basic knowledge bases: content negotiation, Restful API, and SPARQL Endpoint.

### Application/Value

This platform achieved a comprehensive social effect in the “Genealogy Application and Development Competition” held in April 2016. On the one hand, it provides better services to the general public, researchers, and cooperative suppliers and organizations. On the other hand, it also explores the overall technology achievement program of application of Linked Data in libraries and achieves being a good effect on the genealogy research and library field.

**【Romanization of Chinese references is offered in the paper.】**