# 面向知识服务的图书馆数字人文项目建设：方法、流程与技术

夏翠娟，张　磊，贺晨芝

**摘　要**　数字人文因其结合现代信息技术与传统人文研究的特点，近年来成为各相关领域机构的热门话题。图书馆作为人文研究的资源保存与服务中心，正经历以资源管理和服务为重心向以数据管理和知识服务为重心的转移。图书馆的数字人文项目常以语义万维网、大数据、人工智能等新技术手段为支撑、以互联网时代的知识组织为方法，致力于提供区别于传统文献服务的知识服务。文章以上海图书馆“名人手稿档案库”为例，结合多年来数字人文项目的探索与实践经验，从数字人文项目的建设方法、建设流程以及技术框架三方面梳理总结面向知识服务的图书馆数字人文项目的建设过程。

**关键词**　数字人文　知识组织　知识服务　上海图书馆

**引用本文格式**　夏翠娟，张磊，贺晨芝.面向知识服务的图书馆数字人文项目建设：方法、流程与技术[J].图书馆论坛，2018（0）：1-9.

## Construction of Library Digital Humanities Projects for Knowledge Services：Method，Process and Technology

XIA Cuijuan，ZHENG Lei，HE Chenzhi

**Abstract**　Digital humanities，which is the combination of modern information technology and traditional humanities research method，has been regarded as a hot topic in recent years. As the resource preservation and service centers for humanities research，libraries are experiencing the shift from resources management and document search services to data management and knowledge services. With the support of semantic web，big data，artificial intelligence and other new technologies，plus the methods of knowledge reorganization in the Internet Age，the digital humanities projects of libraries are dedicated to providing knowledge services which are different from traditional document search services. Taking the “Celebrities Manuscript Archives” of Shanghai Library as an example，and with reference to the theory and practice of other digital humanities projects，this paper summarizes the construction process of library digital humanities projects from the aspects of methods，work flow and technical framework.

**Keywords**　digital humanities；knowledge organization；knowledge services；Shanghai Library

## 1 概述

数字人文(Digital Humanities)被认为是数字图书馆发展的必然趋势，以图书馆丰富的馆藏资源和结构化的元数据记录为基础，借助大数据、语义网、数据可视化、GIS(地理信息系统)、UGC(用户贡献内容)等现代信息技术为人文研究者提供新的研究视角、研究方法和研究工具，是图书馆的优势所在，也是使命所在，既是机遇，也是挑战[1]。上海图书馆(以下简称“上图”)自2014年起开始投入人力物力资源，探索图书馆从事数字人文项目建设的方法和路径。从上图的特色资源——家谱开始，利用以关联数据(Linked Data)为主的语义万维网(Semantic Web)技术，融合馆藏元数据记录、专家的研究成果、相关的网络资源，以知识本体为基础的知识组织方法重组数据，以RDF对数据进行编码，以平台化的思维为用户提供差别化服务，吸纳并鼓励专家用户贡献知识，建成了“家谱知识服务平台”，探索基于关联数据技术的数字人文项目建设模式[2]。以关联开放数据(Linked Open Data)形式开展开放数据应用开发竞赛，向全社会征集创意的同时推广馆藏资源，获得了良好的社会影响。

在家谱知识服务平台的基础上，上图以24万余种盛宣怀档案及其他大量近现代名人的手稿和档案为基础，建设面向人文研究的“名人手稿档案库”，集手稿档案的编目与展示于一体，利用社会关系分析、实体关系分析、留言、标注等功能支持人文研究。上图与美国柏克莱东亚图书馆合作建设的“中文古籍联合目录及循证平台”，借鉴“循证实践”(Evidence- based Practise)的概念，试图在收集大量现存或散佚的古籍目录数据的基础上，利用互联网时代的知识组织技术，建设古籍版本知识库、作者知识库、刻工知识库、收藏家及藏印知识库、避讳字知识库等，以支持大数据时代的古籍目录学研究、版本学研究、分类学研究。

上图在建设各种文献知识库的同时，建设“人、地、时、事”等基础知识库，以便于从不同维度探索图书馆的所有资源[3]。比如，将“人名规范库”中的每个人当作一个实体，这样可以从某个人出发，探索所有的手稿、档案、著作、照片、音视频资料，而不用受到传统数字图书馆以资源类型的不同而建设相互独立的数据库系统的限制，真正做到面向内容而非面向文献，为研究者提供精准的知识服务而非仅提供文献查阅服务。

上述文献知识库和基础知识库建成后，都在互联网上提供开放数据服务，试图成为面向人文研究的国家数据基础设施的一部分。本文以“名人手稿档案库”项目的建设为例，阐述上图数字人文项目建设的方法、流程与技术，为图书馆开展数字人文项目建设提供参考。

## 2 图书馆数字人文项目建设定位——从文献服务到知识服务

数字人文利用现代信息技术为人文研究提供了新方法、新手段和新视角，成为人文研究领域的热点和前沿。近年来与之相关的各个领域，如各大高校和科研机构，以及作为人文研究支撑的图书馆、档案馆、博物馆等文化遗产继承机构纷纷成立数字人文中心。武汉大学、北京大学、南京大学作为国内人文研究重镇，对数字人文表现出极大的热情，其中武汉大学成立了我国大陆第一个数字人文研究中心，北京大学成立了数字人文研究小组，发布了数字人文指南，主办了两届广受关注的数字人文论坛[4]。

与高校院系和研究机构对数字人文研究的切入点不同，图书馆作为人文研究所需的资源保存和服务中心，主要任务是利用新的技术手段重组资源，为人文研究者提供更好的服务，辅助人文研究，而不是要深入某一个具体的领域，代替人文研究者从事具体的研究工作。因此，图书馆对数字人文的研究侧重在资源的管理和组织，以及服务提供方面，而这正是图书馆长期以来从事的工作，也是优势所在，更是挑战所在。数字人文向图书馆提出了全新的要求：为人文研究提供大规模高质量的数据、科学的研究方法和计算机辅助研究工具的支持[5]。

为了利用图书馆的已有资源更好地辅助人文学者的研究工作，需要深入了解各个领域人文学者的研究场景、研究方法和研究模式。不仅要了解其当前正在使用的研究方法和工具，还要了解技术发展的现状和趋势，将资源、技术和学者的研究需求结合起来，探索如何为学者提供新的研究方法、研究手段和研究视角。

在数字图书馆时代，图书馆利用元数据来揭示和组织资源，基于元数据的字段检索来提供文献查询和全文阅览服务。而在数字人文时代，随着互联网发展和数字图书馆建设，学者可获得的文献不是太少，而是太多，如何帮助学者在海量文献中找到与研究主题相关的数据、事实和知识是图书馆的首要任务。图书馆需要直接为学者提供文献中蕴含的且与研究主题相关的知识，帮助学者准确而全面地定位到所需文献。更重要的是，数字人文的优势还在于提供不一样的视角，帮助学者发现新的研究问题[6]。因而对图书馆来说，完成从文献服务到知识服务的转型是从事数字人文项目建设的主要目的。

## 3 图书馆数字人文项目建设方法——知识重组

图书馆的文献查阅服务是建立在对文献的规范标引和著录、对标引著录的结果——元数据进行组织和管理的基础上的，而知识服务则建立在对文献中的知识进行组织和管理的基础上。知识组织是揭示知识单元(包括显性知识因子和隐性知识因子)、挖掘知识关联的过程或行为，能最为快捷地为用户提供有效知识或信息。知识组织始见于 1929 年美国图书馆学家布利斯的专著，并在图书馆学、情报学的分类系统和叙词表研究基础上发展起来。当前随着语义万维网、大数据技术和人工智能的兴起，知识组织朝着机器可理解数据的方向发展，成为图书馆学、情报学、计算机科学、知识工程学、现代语言学、认知心理学等领域共同研究的课题。

互联网时代的知识组织与传统知识组织截然不同之处在于：分类系统和叙词表主要是为人使用，便于编目员在文献著录时参考，作为元数据元素的取值，同时在检索系统中进行简单的索引以便于检索；而互联网时代的知识组织是为了让机器能够读取、处理并理解数据中蕴含的语义，归根结底是为机器服务，目的是用机器来帮助学者处理研究所需的大量繁琐、重复的前期工作，如资源的搜集、查询、聚类、统计、分析[7]。因而这种知识组织需要完成以下任务：

(1)知识编码的形式化(Formalization)——机器可读。知识编码的形式化即用规范的机器语言来表达知识，其目的是使机器可读。传统图书馆基于 MARC 的元数据记录以 ISO 2709 格式编码，是一种严格遵循标准的编码格式，在图书馆自动化系统中机器可读，也可用于不同图书馆间的数据交换。但在互联网时代，需要采用更为开放的符合 Web 标准的格式，如 XML、RDF 的各种序列化格式(RDF/XML，Turtle，N3，JSON- LD)。这些数据编码格式是 W3C 的推荐标准，被大部分编程语言支持，有着跨平台跨系统跨领域的特性，因而可使知识变成真正的“(任意)机器可读”的数据。

(2)知识单元的细粒度化——机器可计算。图书馆的元数据记录描述的对象是文献，以文献为最小单位，主要描述文献的外部特征，目的是让读者能够查询、定位和阅览文献。在数字图书馆时代，虽然纸质文献大多已经被数字化为电子版本，但这种以文献为最小描述单位的情况没有根本改变。数字人文要求深入揭示文献内部的事实、数据和知识，因而描述的知识单元更细致，具体来说就是文献中的人、地、时、事、物等内容特征及其相互之间的关系，目的是使机器可以根据这些知识单元的各项特征属性进行聚类、统计、分析、推理等计算行为。

(3)知识表示的语义化——机器可理解。让机器能够理解人类的信息是计算机科学永恒的主题。语义万维网提出了首先让机器理解机器的有限目标，W3C 推出 RDF(资源描述框架)标准，用“主体(Subject)- 谓词(Predict)- 客体(Object)”三元组模型作为知识表示的基本框架。主体是谓词

描述的对象，其属性特征可通过定义从同类对象中抽象出来的概念来界定；谓词是严格定义的术语，是描述概念特征的属性；客体作为谓词的值，不仅可以是数据，还可以是另一个对象，谓词即是表示主体和客体之间关联关系的规范化术语。这样的三元组模型以简洁、普适、规范的形式，经过以机器理解为目的形式化编码，可用来形式化地表达任何事实、数据和知识，并可超越系统、平台和领域的限制，使得机器与机器之间的相互理解变成了可能。

在不同领域内，概念及其属性的定义不尽相同。某一领域内可共享的概念及其概念间关系的形式化定义被称为知识本体(Ontology)，简称本体。本体是语义万维网环境下知识组织的主要方法和技术之一。

(4)知识组织的关联化——机器可推理。在现实世界中，事物之间的关联是普遍存在的，若将这种关联关系反映到机器世界中，机器便可基于大规模的关联关系推理出新的知识。而知识单元之间的关联关系越多，越有利于推理结果的准确性。关联数据(Linked Data)和知识图谱(Knowledge Graph)就是在知识之间建立可被机器理解的关联关系的技术，这种技术建立在 Web 的 HTTP 协议之上，以 RDF 三元组为最小的知识单元。三元组中的主体、谓词、客体都可由 HTTP URI 来唯一定位和标识。因此，其建立的关联关系是跨网域的，而非只在某一系统内部生效。这种广泛而深刻的、基于 Web 的、植入数据底层的关联关系为大范围、跨领域、大规模数据的机器推理带来了便利。

(5)知识增长的自动化——机器可自学习。在数字图书馆建设时期，图书馆的知识组织工作主要依赖于人工的编目著录，尤其是各种分类法、叙词表、规范档等。随着机器智能时代到来，以及基于神经网络的深度学习(Deep Learning)技术进一步成熟，在知识组织过程中，开始借助机器智能自动地完成知识增长的过程。目前在自然语言处理(NLP)、名称实体识别(NEI)、自动标引和自动分类领域，机器学习大有用武之地。机器学习技术在近年来欧盟最大的数字人文项目“威尼斯时间机器”中得到有效应用[8]。

## 4 图书馆数字人文项目建设流程

依上文所述，图书馆数字人文项目建设的主要目标是提供知识服务，主要方法是互联网时代的知识组织方法。因此，对图书馆已有数据和资源，用知识组织的方法进行知识重组，并利用新技术手段提供知识服务，是数字人文项目建设的主要任务。知识重组的核心任务是数据建模，即根据系统需求和所能获得的数据数量和质量来设计数据模型，定义涉及的概念、概念特征，以及概念与概念间的关系，也就是本体设计，这是基于 RDF 三元组的数据模型的基础。图 1 是上图数字人文项目建设的流程示意图。在项目建设过程中，本体不仅与需求和用户应用场景相关，也与能获得的数据相关，同时受到系统设计开发过程中技术条件的制约，随着项目推进，需要在多次反复中不断完善。

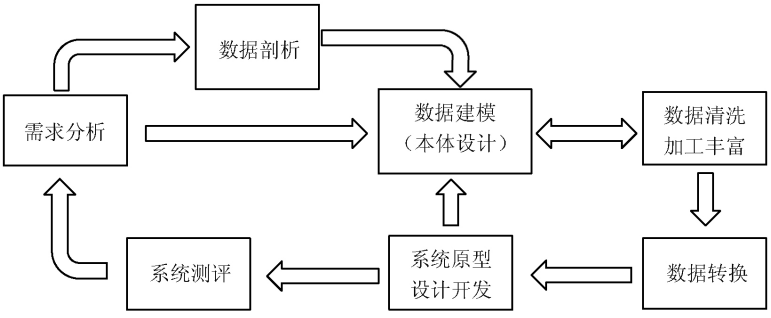


图 1 上图数字人文项目建设流程图

### 4.1 需求分析

需求分析的目的是在充分了解用户应用场景的情况下，界定系统的长、短期目标。上图拥有近 7 万件近现代名人手稿及档案资料，涉及 2 万余人、2000 余个县级及以上地点，时间跨越自晚清、民国至现当代的 200 余年，是研究近现代文学、历史、社会学的第一手宝贵资料。过去

10 余年间，上图成立专门的研究整理小组，负责这些资料的编目，但目前只有供编目用的元数据著录系统，没有基于 Web 提供服务的平台。上图“名人手稿档案库”的首要任务是满足资料的查询、阅览功能，在此基础上逐步实现支持人文研究的动态聚类、时空分析、社会关系分析、数据统计、研究交流等功能。

### 4.2 知识重组

#### 4.2.1 数据剖析

数字图书馆建设为图书馆从事数字人文项目建设奠定了基础，大量规范化、结构化的元数据是数字图书馆建设的宝贵成果，是知识重组的原材料。但是，由于文献服务与知识服务的要求不同，所以需要对这些原材料进行剖析：一方面了解数据的内容结构，为本体设计做准备；另一方面发现不足之处，为进一步的数据清洗加工丰富做准备。

上图名人手稿档案著录系统基于 DCMI 元数据方案设计方法来构建，将资源分成创作手稿、信函、照片、实物、证书等 12 个大类，共用一个核心元数据元素集；每种资源又有自己的特殊元数据元素集，并考虑到人名规范档的建设，在著录过程中对手稿及档案的责任者进行名称规范控制。元数据记录和规范数据记录以表格的形式存储于关系数据库中，并可导出 XML 格式。正如上文所述，这种以文献为知识单元的元数据规范和基于此规范产生的元数据记录存在着一定的缺陷：

(1)由于著录单元的单一性，因此难以充分揭示单个档案与档案集合之间、单个档案与其组成部分之间的关系。例如，当以一封信件为著录单元时，如何揭示一封信与一包信件、一封信与信封信纸之间的关系？如何深入描述包含这一封信件的信件包和一封信的组成部分信封和信纸？这在档案管理中是很常见的问题和需求。

(2)对档案中涉及的人、机构、地、时、事揭示不足。虽然为责任者建立了在本系统范围内进行规范控制的规范档，但没有对机构、地名、时间、事件建立规范档并进行规范控制，只是用自由词作为元数据元素的值，而这些都是能够深入揭示文献内容的知识单元。仅有简单的字符串值(String)是不够的，需要将机构、地名、时间、事件当成现实中真正存在过的对象(Thing)，赋予 HTTP URI 并补充大量结构化的数据，如地名的行政归属地、经纬度，事件的发生时间、地点、人物。

#### 4.2.2 数据建模

在本项目中，数据建模即本体设计。本体设计要解决的问题是厘清深入揭示档案内容的数据中可以抽象出哪些概念，每个概念有哪些特征，以及概念之间有哪些关系。与此同时，用明确规范的术语来表达这些概念——在本体中用“类”(Class)来表示，概念的特征及概念间的关系——在本体中用“属性”(Property)来表示。本体中的类是 RDF 三元组中的主体和客体抽象出来的概念，属性是 RDF 三元组的谓词，类和属性为 RDF 数据赋予了语义，可被机器读取和处理，经过机器的聚类、计算、统计、分析、推理后变成人可理解的知识。

本体设计中常常存在的困惑是哪些作为类、哪些作为属性。比如，对一封信的信封的处理，是否需要定义“信封”这个类，主要依据还是需求和数据情况。在上图名人手稿档案本体中，是将信封作为类来处理的，主要原因是：在已有的元数据记录中，已经将信封的各项特征如收件人地址、姓名，发件人地址、姓名，邮戳，信封书写文字的颜色和字体等分别作为不同的子元素，是高度结构化的数据。如果在本体中将信封不作为类来处理，会损失这些结构化数据，而作为类处理后，就可以将上述子元素定义为这个类的属性。此外，当信件和信封作为类来定义后，还能以面向对象的思维灵活地定义它们之间的关系。在本体中，每一个类都有不只一个属性来描述，每个类的实体都是描述的对象，而不仅仅是文献。更重要的是，信封的各项特征(人、地、时、事)也是人文研究的宝贵资料，结构化后便于机器计算。

上图名人手稿档案本体包含 44 个类和 195 个属性，已发布在 Web 上，见图 2；主要类的关系见图 3。

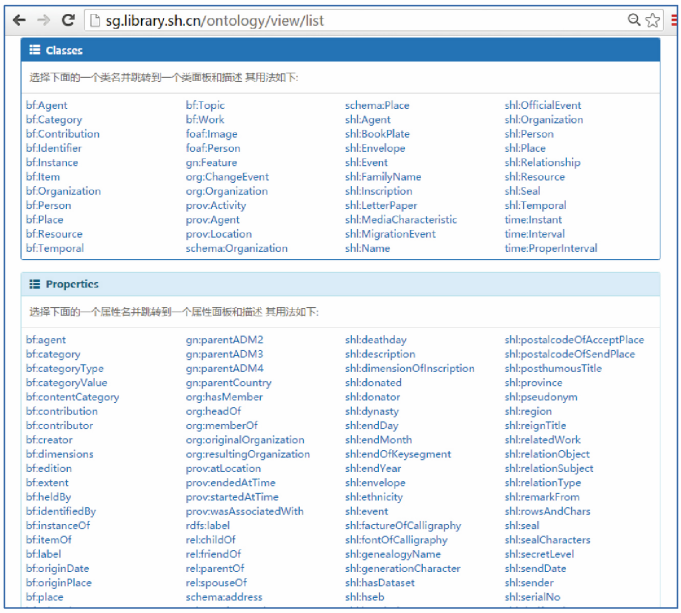


图 2　上图名人手稿档案本体网站

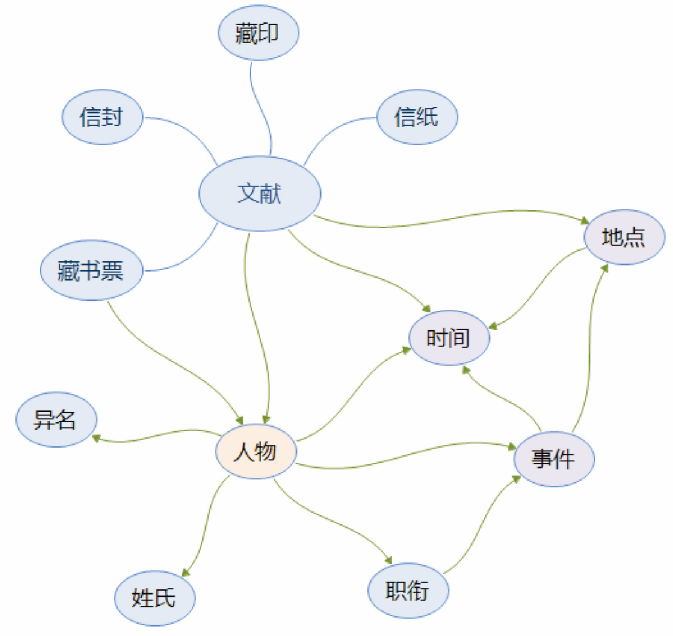


图 3　上图名人手稿档案本体模型

#### 4.2.3　数据清洗、加工、丰富

这一步的主要任务是根据本体设计的结果，从已有的数据中提取所有本体中定义的类和属性，进一步发现数据中的错漏和不规范之处并加以修正，对缺失的数据进行补充和丰富，对属性的取值进行规范控制，必要时定义一些取值词表，如档案类型、责任方式。在这一步中还可能会发现本体设计的不合理之处，也可能会根据新获得的数据对本体进行补充修正。这正是用本体的方法来进行数据建模的一个好处，因为本体的类与类之间既相互联系又彼此独立，修改或增加一个类的属性不会对其他类造成影响。

#### 4.2.4　数据转换

数据转换的主要目的是得到以本体作为数据模型的 RDF 数据。首先，需要定义已有数据(一般是 RDB、Excel、CSV 格式)各个字段与本体的映射。其次，为每一个数据实例(实体、对象)赋予合适的类，生成 HTTP URI，将 URI 作为 RDF 三元组的主体，将描述该类的属性作为三元组的谓词，谓词的取值可以是一个字串值、数值、日期等，也可以是另一个数据实例的 URI。这样，每个数据实例就有多个三元组来描述，每个三元组都是一个知识单元，多个互相关联的三元组就构成了一个知识图谱。

### 4.3　系统设计、开发与测评

如果说本体设计和数据清洗加工转换得到的 RDF 数据是原材料，那么系统设计就是菜谱，而系统开发则是将原材料和菜谱生产出可为用户提供服务的产品，系统测评则是对产品进行检验，看是否符合既定需求。在系统设计、开发和测评的过程中，受到技术条件的制约，仍然会对本体提出进一步的修改需求，因为本体既是知识结构，也是数据结构，本体设计得过于复杂或过于简单，会对需求的实现和系统的性能产生影响。

## 5　图书馆数字人文项目的技术框架

图 4 是上图数字人文项目建设的技术框架，主要分为知识重组、数据存储和知识服务三个部分。知识重组的技术要求是对图书馆的元数据、目录索引数据和用以补充这两者不足的外部数据进行清洗加工后，完成从 String 到 Thing 的转

变，即从二维表格数据(RDB/EXCEL/CSV/TXT)到多维网状 RDF 图数据的转变。数据存储要解决的问题是用图数据库来存储 RDF 数据，并支持高性能的数据存取。知识服务的目的主要是将图数据库中的 RDF 数据用多样化的数据可视化技术展示、呈现给用户，方便用户查询、发现、探索数据中存在的事实、数据和知识。此外，作为一个大型研究型公共图书馆，上图还承担着为本地区中小型图书馆提供数据支撑服务的责任，也希望能将经过知识重组后的数据开放给社会大众和第三方机构，所以还提供面向机器的开放数据服务，便于开发人员调用并整合到自己的应用系统之中。

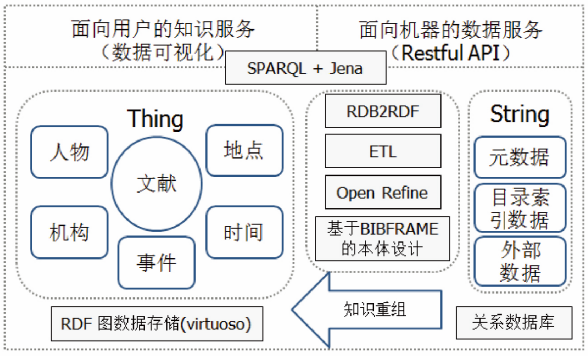


图 4　上图数字人文项目建设的技术框架

## 5.1　数据处理

数据处理主要包括数据清洗、加工、丰富(采集)三个方面。B/S 架构的开源软件 Open Refine 在数据清洗加工方面有着广泛的应用，被评价为“看起来像表格，用起来像数据库”，有良好的用户界面和强大的数据处理功能。它首先是一个强大的数据剖析工具，可以将半结构化的数据(维基、XML、TXT 等)根据自定义的规则结构化，变成二维表格数据，或直接导入 EXCEL 或 CSV 格式的结构化数据，生成表格后，对表格的列提供分面统计的功能，可以清晰地看出哪些数据有问题，如明显的数据错漏和格式不规范，并可在 Web 界面上直接批量修改；或者转换数据类型，如将数值型的数据转换成文本型的数据。此外，该软件还支持用 GREL 语言编程的方式对数据进行复杂操作，如逻辑运算、数学运算、字符串处理等，并可实现循环、嵌套等流程控制功能。Open Refine 支持通过外部数据的 API 来获取数据，并直接在字段层面与本地数据融合。

为补充数据的不足，上图开发了一个 ETL (Extract- Transformation- Load)工具，支持Web 数据的自动采集、对采集的数据进行过滤、配置本体映射并转换成 RDF 数据后导入 RDF 存储库；还开发了一个支持 W3C 的 R2RML 语言的 RDB2RDF 工具，可直接将关系数据库中的数据与名人手稿档案本体建立映射，导出 RDF 数据。

## 5.2　数据建模

名人手稿档案本体基于美国国会图书馆的书目框架(BIBFRAME2.0)设计，复用了其Work- Instance- Item 三层模型作为图 3 中“文献”类的数据模型，在此基础上扩展适用于名人手稿档案的类和属性。BIBFRAME2.0 的三层模型可以较好地解决作品与不同版本，版本与不同复本之间的关系，见图 5。

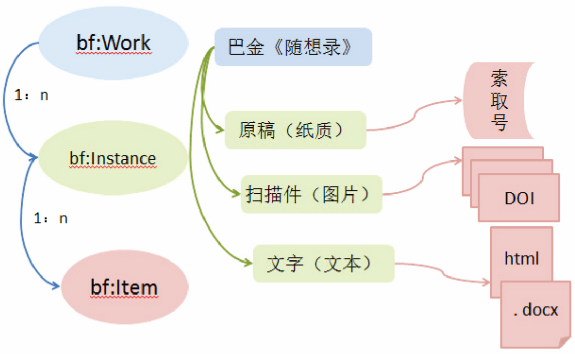


图 5　名人手稿档案的书目数据模型

## 5.3　系统开发

“名人手稿档案库”系统开发时，在数据存储方案选择上，主要用到了图数据存储技术。这是因为 RDF 数据本身是多维网状的图数据格式，由节点和边组成。随着大数据技术的兴起，近年各种 NoSQL 数据库快速发展，可用于存储 RDF 数据的 NoSQL 解决方案越来越多。图数据库是 NoSQL 数据库的一种，最大优点是可以直接导入 RDF 数据，设计者只需考虑数据本身的内在知识逻辑，无需像关系数据库那样设计大量的表

和字段，将知识逻辑与数据存储结构紧紧地捆绑在一起。RDF 数据的结构由本体决定并反映在 RDF 数据底层，与数据库无关。这种特性让图数据库拥有一个重要优点：有着强大的灵活性和可扩展性，当本体有所变化或有新的数据增加时，只是新节点与边的增加，可以随时更新数据而不会对原有数据产生影响。

由于本体、RDF、图数据库从数据结构到数据编码，再到数据存储的整个周期，都有着极大的灵活性和可扩展性，所以项目建设的流程就可以是一个不断反复、迭代的流程，而无需一蹴而就。例如，在建设“名人手稿档案知识库”中的人名规范档时，对名人的各种基本信息、社会关系、职衔数据分步骤加工和导入，一部分人只有生卒年、籍贯等数据，缺少职衔、社会关系数据，但并不影响项目的总体进程，可以在流程的任何阶段进行补充，甚至在系统开发完成后还可以对本体进行微调，或导入新的数据。图 6 是本体属性增加后实例属性增加的示意图，试图说明在项目流程的任意阶段，数据结构的修改和数据本身的修改不影响原有数据，只是节点的增加和节点之间边的增加。

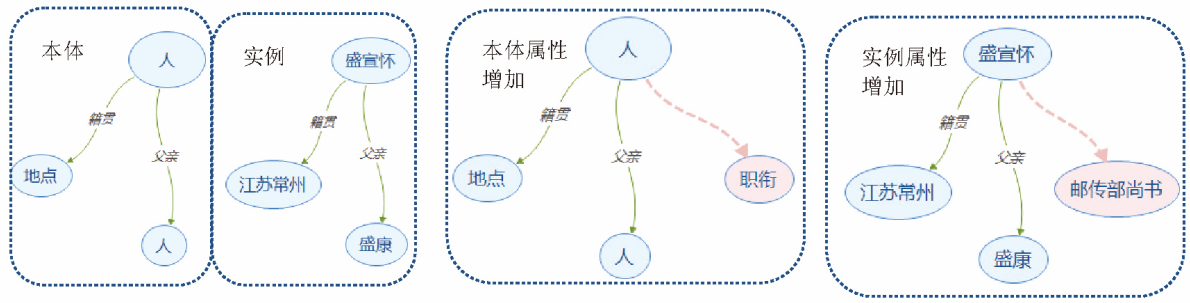


图 6　本体属性增加后实例属性增加示意图

在调研众多 RDF Store 后，项目采用 Open Link Software 公司的开源版 Virtuoso 来存储与名人手稿档案资源直接相关的 RDF 数据，同时利用关系数据库来存储系统的流程控制、业务逻辑、用户访问数据。在开发方面，采用语义万维网的相关技术，如 RDF 数据查询语言 SPARQL 和Jena 开发框架，实现对 RDF 数据的查询和增删改操作。

### 5.4　知识服务与开放数据服务

知识服务是用数据可视化技术将机器可读的数据以用户喜闻乐见的形式展示出来，变成能被人辨识、理解、发现、探索和勘误的知识。数字人文领域常见的数据可视化方式有时空分析、社会关系分析、实体关系分析、文本统计分析等。上图“名人手稿档案库”采用这几种数据可视化方式。比如，在地图上展示名人的籍贯分布情况，并支持用户在地图上点选一个区域来发现籍贯为该区域范围内的名人的手稿档案文献；以名人的出生时间作为时间轴，用户可拖动时间轴，实时展示在某段时间内出生的名人的手稿档案文献；利用名人间的信函、电报的数据，用户可以发现人与人之间的社会关系和联系的强弱度，或者发现任意两个人之间的通信通电情况。目前有大量优秀的可以集成到数字人文平台中的数据可视化的工具，如 D3.js、Data.js、Tableau、Gephi、Baidu Echarts、Zoom Charts，本项目主要采用 Zoom Charts。Zoom Charts 支持各种统计图、关系图、时空图的生成，时空可视化中用高德地图作为地图服务工具。

Web2.0 技术是 2000 年后蓬勃发展的一项技术，支持用户贡献内容(UGC)。在本项目中，这项技术被应用于支持用户之间的交流和在线研究。用户可以针对一件档案发表留言，贡献自己的疑问、观点和研究心得，还可以在手稿的扫描图片上针对任何区域做标记，留下自己的笔记，并决定是否和他人共享笔记的内容。

如果说知识服务是面向人的服务，那么开放数据服务则是面向机器的服务。关联开放数据(LOD)是近年来图书馆、档案馆、博物馆等文化机构采用得较多的数据开放技术。上图的数据开放建立在关联开放数据的基础上，主要提供开放数据接口而非以数据打包下载的方式来支持在Web上调用数据。开放数据接口有三种：Sparql Endpoint、Restful API、Content Negotiation。Sparql Endpoint是RDF存储库Virtuoso本身提供的接口，只要开放相关的端口即可在Web上被访问，但要求开发人员熟练掌握SPARQL语言；Restful API是建立在HTTP协议上的数据接口，返回JSON-LD格式的数据，可被大部分程序语言调用和解析；Content Negotiation允许程序访问资源的HTTP URI，并在http header中指明返回数据的内容类型来获取相应格式的数据。

## 6 结语

“名人手稿档案库”是上图整个数字人文平台的一部分，其建设方法、流程、技术也和数字人文平台的建设基本一致。经过3年多的探索，这套方法、流程和所采用的技术逐步趋于成熟，其特点是：数据架构和技术平台的灵活性和可扩展性，支持在不断的迭代中完善数据、模型和功能，减轻了项目建设和系统实施的压力。该项目并非尽善尽美，仍然需要进一步摸索，尤其是对各个具体的人文领域的研究方法、研究资料、应用场景的了解和把握；需要不断优化知识组织方法和知识服务功能，做到真正从实处帮助用户更好地利用图书馆的资源，提高研究的效率，扩展研究视野，发现新的研究课题；继续发挥互联网时代图书馆的职能，加快加强数据加工和知识重组工作，将知识组织的成果开放给社会，成为互联网时代人文研究数据基础设施的一部分。

## 参考文献


[1][6]冯晴，陈惠兰. 国外图书馆参与数字人文研究述评[J]. 图书馆杂志，2016（0）：14-19.

[2] 夏翠娟，刘炜，陈涛，等. 家谱关联数据服务平台的开发实践[J]. 中国图书馆学报，2016（0）：27-38.

[3] 夏翠娟. 以连接开放资料服务为基础的数位人文平台建设方案研究[J]. 图书馆学与资讯科学，2017（0）：47-70.

[4] 朱本军，聂华. 互动与共生：数字人文与史学研究——第二届“北京大学数字人文论坛”综述[J]. 大学图书馆学报，2017（0）：18-22.

[5] 周晨. 国际数字人文研究特征与知识结构[J]. 图书馆论坛，2017（0）：1-8.

[7] 明海英. 在困惑中追寻前沿研究方向——记武汉大学出版科学系教授王晓光[N]. 中国社会科学报，2014-02-21.

[8] Alison Abbott. The time machine’reconstructing ancient Venice’s social networks[J]. Nature，2017，546（7658）：341-344.


作者简介 夏翠娟，上海图书馆系统网络中心研发部高级工程师；张磊，上海图书馆系统网络中心研发部高级工程师；贺晨芝，上海图书馆系统网络中心研发部助理工程师。

收稿日期 2017-10-10

责任编辑：刘洪；英文编辑：杨继贤