关联数据在家谱数字人文服务中的应用*

夏翠娟 张 磊（上海图书馆）

摘 要 数字人文是数字图书馆建设达到一定规模后的必然发展方向。云计算、大数据、语义技术提供了很好的方法，可以对大规模、多种类的数字资源进行内容分析和基于数据间内在逻辑关联的知识组织，进而提供面向内容和知识的精准服务，挖掘不为人知的知识，节省人力和时间，促进资源的利用，关联数据正是一种已经发展成熟的语义技术实现方式。上海图书馆以家谱开始，利用关联基于语义万维网的规范控制方法和基于知识本体的知识组织方法以及关联数据技术、社会化网络技术（SNS）、可视化技术，实现了面向知识发现的数字人文服务，本文阐述了实现的方法和过程，并以“上川明经胡氏”和“湖广填四川”为例，详细展示了关联数据在数字人文研究中的作用和用法。

关键词 数字人文 关联数据 家谱 知识服务

DOI: 10.13663/j.cnki.lj.2016.10.004

The Application of Linked Data in Shanghai Library's Service of Genealogy Digital Humanities

Xia Cuijuan, Zhang Lei (Shanghai Library)

Abstract Digital Humanities (DH) is the future development direction of modern library. Cloud computing, big data, Semantic Web technologies provide very good methods for content analysis and knowledge organization based on logical relation among the digital resources on a large scale and multi species. In turn, the library could provide accurate service regarding content and knowledge by mining hidden knowledge, while saving manpower and time in organizing resources and promoting the use of resources. Linked Data is a mature technology to realize Semantic Web. Shanghai Library takes genealogy resources as an example and a starting point to practice Digital Humanities by means of knowledge reorganization and provides knowledge discovery service based on Linked Data technologies. This paper expounds the method and process of implementation, and demonstrates how Linked Data can be used in Digital Humanities through the case of “Shangchuan Ming Jing Hu Migration Map”.

Key words Digital Humanities, Linked Data, Genealogy, Knowledge service

1 数字人文的兴起

“数字人文是一个重要的多学科（交叉）领域，指应用数字技术从事人文科学研究。旨在建立应用和模型，不仅是一种以信息技术作为工具的新型研究，为人文科学创建新的应用和新的模型，而且促进计算机科学的进步。同时它也研究信息技术对于文化遗产和人类记忆机构，图书馆档案馆以及数字文化的影响。”[1-2] 典型的应用如对莎士比亚作品和《红楼梦》等经典文学或历史文献的文本分析，结合地理信息系统（GIS）研究社会习俗的演变，利用海量录像研究新生儿的语言形成等。在研究选题的深度广度和规模方面，前所未有地取得了令人瞩目的成果。计算机与网络技术的应用已逐渐深入到人文研究的各个领域，数字技术与人文研究的结合成了学者时下研究的热门

* 本文系国家哲学社会科学基金青年项目“W3C 的 RDB2RDF 标准规范在关联数据服务构建中的应用”（项目编号：13CTQ008）研究成果之一。

主题。近年来，一批数字人文研究机构，如国际数字人文组织联盟（The Alliance of Digital Humanities Organizations）、数字人文学会（The Society for Digital Humanities）相继成立。许多大学还设立了自己的数字人文研究中心，如美国斯坦福人文实验室、伦敦国王学院人文计算研究中心等。

近 20 年来随着数字图书馆建设的持续进行，大量图书、期刊、报纸、档案、古籍、家谱、照片、乐曲、音视频被数字化，同时留下了大量的、遵循一定元数据标准规范的、高度结构化的元数据记录，将使人文学者的研究手段发生根本改变。数字图书馆建设主要是对馆藏资源进行扫描、编目、整理，注重的是资源的数字化形态，目的是利用计算协助文献管理，同时为读者提供查阅服务。但当数字化的建设达到了一定规模的时候，这种基于文献的管理和查阅服务就碰到了瓶颈。当查询结果达到数十页，成千上万条，而读者需要去一页一页地翻阅全文的时候，就很难通过肉眼去发现散落在大量资源中的相关性事实或知识。而现在，各种现代化的数字技术，如云计算、大数据、语义技术提供了很好的方法，可以对大规模、多种类的数字资源进行内容分析和基于数据间内在逻辑关联的知识组织，进而提供面向内容和知识的精准服务，挖掘不为人知的知识，节省人力和时间，促进资源的利用，甚至在资源利用的过程中产生新的知识。

数据和技术是数字人文的两大支柱，数字图书馆建设遗留下来的元数据和数字化全文为图书馆提供数字人文服务奠定了数据基础，而关联数据正是一种已经发展成熟的语义技术实现方式，基于关联数据技术来实现数字人文也逐渐成为一种趋势，已在国外众多数字人文项目中得到应用。

2　国内外研究与应用调研

欧美学界涌现出许多古籍数字化、文献数据库建设等数字人文领域的新项目。如斯坦福人文研究中心的“知识界通信地图项目（Mapping the Republic of Letters）”[3]，斯坦福罗马世界的地理空间网络模型等。我国图书馆界也开始意识到数字图书馆建设已经到了一定的规模，需要向数字人文寻求新的机遇和新的突破。2014 年在上海举行的“图书馆前沿论坛（IT4L）”学术会议，其主题即是“数字人文与语义技术”，来自美国、新西兰以及复旦大学、上海图书馆等图书馆的教授及从业者共同探讨数字人文的概念、技术，交流语义技术实现数字人文的案例。

语义技术尤其是关联数据技术作为数字人文的一种技术手段，已得到广泛的瞩目和应用。近年来，几乎所有的数字人文会议至少有一个会场的主题是“数字人文、RDF 和关联数据”，而 2015 年的数字人文大会，是和第三届图书馆、档案馆、博物馆的关联开放数据（LODLAM）年会合办的[4]。一些利用关联数据技术实现数字人文服务的项目也陆续涌现，其中较为成熟的应用是“欧洲数字手稿项目”和“关联人文项目”。

欧洲数字手稿项目（Digitized Manuscripts to Europeana，DM2E）是由德国的柏林洪堡大学图书馆信息学院主持，由曼海姆大学、开放知识基金会等多个研究机构参与的大型研究项目。其建设目的主要有两个：一是将不同来源、不同格式的元数据适配欧洲数据模型（EDM），整合进欧洲数字图书馆；另一个目的是以关联数据技术重构并发布关联开放数据集，为数字人文服务，同时开发出支持数据展示、处理和整合的工具，并可为其他应用项目利用[5]。DM2E 已发布了公开的关联数据集，默认以 CC0 协议开放，但允许数据提供方选择具体的数据开放获取条款。数据集包含的资源种类繁多，包括人文科学家的手稿及相关的人文研究资料，如古籍、信件、图书和旧杂志等。DM2E 的数据模型基于 EDM 设计，EDM 是欧洲数字图书馆用于容纳图、档、博等文化遗产继承机构资源的数据模型，DM2E 数据模型作为 EDM 的一个应用纲要（Application Profile），传承了 EDM 的可扩展性和包容性。该数据集遵循关联数据四原则，符合开放数据的五星标准，可供学术检索，也以 Restful API 的方式提供数据消费，可供编程获取，整合进其他的数字人文服务平台。数据集的支撑平台

支持学者访问内容、添加标注、链接到其他资源；支持版本控制机制，保留学者的操作历史；平台利用关联工具 SILK 自动地为数据集添加外部链接。主要是通过两个自开发的工具来实现这些功能的：一个是分面数据浏览器 OmNom，允许学者从不同的维度探索数据集的内容；另一个是“语义标注工具”Pundit，允许用户为网页添加结构化的语义描述数据，这些用户添加的标注基于“开放注释数据模型（Open Annotation Data Model）”组织，以机器可读的 RDF 格式编码，存储于 RDF 数据库中，以 SPARQL 端点和 REST API 作为数据消费接口，可供程序调用。

“关联人文项目”由美国国家人文研究基金和德国科学基金支持，印第安纳大学伯明顿分校主持，自 2012 年至 2014 年，历时两年。其全称是“数字人文关联与推广（Linking and Populating the Digital Humanities）”，简称“关联人文项目（Linked Humanities Project）”。目的是研发一些用于数字人文资源的数据整合和维护工具，在数据之间建立关联关系，促进数据的共享和重用，为人文研究服务。该项目已经开发了一个可应用于其他数字人文项目的关联数据平台 LODE，包括数据浏览、数据关联和数据提升三个部分。该平台的应用案例有“印第安纳哲学本体项目”和“斯坦福哲学百科全书”，项目开发了一些 Web 服务，使之与 DBPedia、Freebase 等外部知识库关联起来，推广哲学本体的应用[6]。另一个应用是犹太文化历史，在学术参考资料之间建立关联并发布为关联数据，在 Web 上提供数据消费服务，同时可以在 DM2E 项目建立的语义标注环境中对这些资料添加标注，允许数字人文研究学者添加更多的关联，丰富其内容。

在家谱资源的组织上，图书馆尤其是国内图书馆习惯于将家谱作为一种文献资源来提供服务，利用 MARC 或 DC 元数据对资源的文献特征和少量内容特征作标引，提供基于题名、纂修者（著者）、版本、谱籍地、堂号、先祖等字段关键词匹配的检索，如日本国立国会图书馆的东洋文库、国家图书馆的“中华寻根网”以及国内各大收藏机构的家谱资源库，基本上与上海图书馆原有的家谱数据库相似，无法提供基于内容和知识的探索。从目前对家谱资源的利用来看，家谱作为一种宝贵的人文研究资料的价值还没有被充分发掘，而数字人文的方法和技术，尤其是应用已经成熟的关联数据技术有助于改善这种状况。

近年来国外出现了利用基于本体的语义技术来重构家谱数据的案例，如美国犹他家谱学会的 FamilySearch.org 网站以及 ancestry.com 网站，不仅仅提供描述文献特征的字段关键词的检索，还可以利用时空关联和亲属关系等内容特征来探索家谱资源和人物关系。较有创新性和以数字人文研究为目的是斯坦福图书馆的大不列颠名人库（Kindred Britain），包含 1500 年历史里的 3 万余英国名人，用可视化技术展示名人之间的关系，包括血缘关系、或因处于同一时空而产生的关联关系，可以帮助学者快速地从海量的数据中发现新的知识。

3 关联数据在家谱知识组织和数据发布中的应用

在数据的层面，数字人文要求知识单元的细粒度化、知识组织的语义化、知识呈现的可视化。关联数据第一、二原则要求以 HTTP URI 来标识和定位一切事物（Thing），为互联网环境下的规范控制奠定了基础。关联数据的第三原则要求使用资源描述框架（RDF）作为抽象数据模型并尽可能多地揭示资源间的关系，是一种科学的知识组织方法。RDF 以“主 - 谓 - 宾”结构的三元组作为基本的数据单位，不再以与一种文献对应的一条元数据记录作为基本的数据单位，一条元数据记录往往可以拆分成多个三元组。一个三元组是关于某个知识点、数据或事实的描述，具有独立描述逻辑，可以实现知识单元的细粒度化。关联数据的第四原则强调数据间内在的关联，并使得这种关联关系可被机器所理解，可以实现知识组织的语义化，因而基于关联数据的四原则来组织和发布家谱数据，可以很好地满足数字人文对数据的要求。

（1）用 HTTP URI 作为一切事物的名称。以下是上海图书馆（下文简称“上图”）家谱数据中的各类实体的 HTTP URI，不仅作为资

源的唯一标识符，也作为全球定位符，这样就为互联网环境下的规范控制奠定了基础。

家谱文献：http://data.library.sh.cn/jp/resource/work/8y9p7s2euppwnerq
人：http://data.library.sh.cn/jp/entity/person/qxj915pkohm96unn
姓氏：http://data.library.sh.cn/authority/familyname/68n959cf8zdfkz3v
地：http://data.library.sh.cn/entity/place/tk5s4pej6linq9tr
时：http://data.library.sh.cn/authority/temporal/4alljneqiivh5691
机构：http://data.library.sh.cn/entity/organization/brvqlrg8y55v1b5q

（2）基于知识本体和 RDF 抽象数据模型来组织数据。RDF 数据由“主体 - 谓词 - 客体”所构成的三元组组成，其主体是一个资源对象，其客体可以是一个字串值，可以是另一个资源对象，而谓词用于表示主体和客体之间的关系，来自于严格定义的知识本体，也是独立的有语义的知识单元。这些知识单元经过一定的 RDF 序列化格式如 RDF/XML，RDF/Turtle 编码后，可被机器识别。以下是对“明洪武”这个历史纪年的 RDF/Turtle 格式的描述：

```
<http: //data.library.sh.cn/authority/temporal/3rwxdjxxfz5bhff9> a shl:Temporal;
    bf:label " 明洪武 ";
    shl:monarch " 太祖 ";
    shl:monarchName  " 朱元璋 ";
    shl:reignTitle  " 洪武 ";
    shl:dynasty <http://data.library.sh.cn/authority/temporal/yex4deivsad41p9q> ;
    time:intervalDuring  <http://data.library.sh.cn/authority/temporal/yex4deivsad41p9q>
    shl:beginYear  1368;
    shl:endYear  1398.
```

（3）当访问资源的 URI 时，以 RDF 数据提供有用的信息。家谱数据服务平台支持内容协商机制，当人访问资源的 URI 时，返回 HTML 网页，当程序访问资源的 URI 时，以 JSON-LD 格式返回 RDF 数据。

如程序访问资源的 URI 并用 .json 表示希望返回数据的格式为 JSON 格式：http://data.library.sh.cn/jp/entity/person/qxj915pkohm96unn.json

则返回如下结果：

{"result":{"data":{"@id":"http://data.library.sh.cn/jp/entity/person/qxj915pkohm96unn","@type":"http://www.library.sh.cn/ontology/Person","label":[{"@language":"cht","@value":" 星 華 "},{"@language":"chs","@value":" 星 华 "}],"relatedWork":"http://data.library.sh.cn/jp/resource/work/8y9p7s2euppwnerq","roleOfFamily":"http://data.library.sh.cn/jp/vocab/ancestor/xian-zu","familyName":"http://data.library.sh.cn/authority/familyname/rn3hurvwucnb24pb","@context":{"familyName":{"@id":"http://xmlns.com/foaf/0.1/familyName","@type":"@id"},"relatedWork":{"@id":"http://www.library.sh.cn/ontology/relatedWork","@type":"@id"},"label":"http://bibframe.org/vocab/label","roleOfFamily":{"@id":"http://www.library.sh.cn/ontology/roleOfFamily","@type":"@id"}}}}}

（4）尽可能多地提供指向其他资源的 URI，以发现更多的信息。在家谱文献的 RDF 描述数据中，以 HTTP URI 来指代姓氏，人（作者、先祖名人），地（谱籍地名），时（中国历史纪年中的朝代），机构（收藏机构），用本体中定义的属性来表达它们与文献之间的关系。当获得文献的描述数据的同时，如果访问这些与之相关联的实体的 URI，就能获得更多的关于这些人、地、时、事的详细信息。如访问朝代的 URI http://data.library.sh.cn/authority/temporal/yex4deivsad41p9q，将会获取该朝代的起止公元年为 1368-1644 年。

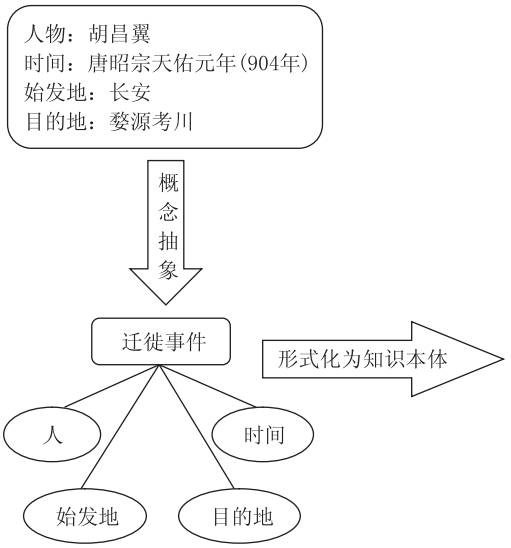
<http://data.library.sh.cn/jp/resource/work/oyz2f36kouez9jdy> a bf:Work;
shl:place <http://data.library.sh.cn/entity/place/t3tec8y1oy2j3kjc>; // 地
shl:temporal <http://data.library.sh.cn/authority/temporal/yex4deivsad41p9q>;// 时
bf:creator < http://data.library.sh.cn/jp/entity/person/etrd44w3m3g1vncn>; // 人
bf:subject <http://data.library.sh.cn/jp/authority/titleofancestraltemple/6biawgp5dbdm9hkm>;
bf:subject <http://data.library.sh.cn/authority/familyname/4feibkhldtiroeu3" // 姓氏
<http://data.library.sh.cn/jp/resource/instance/apz8aio37wj6y524> a bf:Instance;
bf:category <http://data.library.sh.cn/vocab/binding/xian-zhuang>;
bf:edition <http://data.library.sh.cn/vocab/edition/mu-zi-huo-ben>;
bf:extent " 五册 ";
shl:temporalValue 1936, "1936 年 " ;
bf:instanceOf <http://data.library.sh.cn/jp/resource/work/oyz2f36kouez9jdy>;
<http://data.library.sh.cn/jp/resource/item/fgxpi3bc8km3r672> a bf:Item;
bf:heldBy<http://data.library.sh.cn/entity/organization/uoqz22aqnemd3idn>;// 机构
bf: itemOf<http://data.library.sh.cn/jp/resource/instance/apz8aio37wj6y524>;
shl:description " 漳州市臺工辦林嘉書（複印本，存卷一、三、五）" ;

4 家谱数字人文服务案例及实现

家族的迁徙事件是家谱中记载的重要内容，研究某个家族、某个地域或某段时间范围内的迁徙路线是人文研究如人口学、人类学、历史学中的一个重要课题。在上图“家谱知识库原形系统”中，试图以某一个家族的迁徙图为例，探索数字人文的实现路径。

第一种是利用关联数据的知识组织功能，把散落在不同家谱文献中的人、地、时、事关联起来，形成完整的知识图，以可视化的方式展示。

以文化名人“胡适”先生的家族“上川明经胡氏”的迁徙图为例：在上图馆藏家谱中，一共有 6 种跟明经胡氏家族相关的家谱，有两种家谱（即《上川明经胡氏宗谱》）的元数据中记载了胡适是该家族的先祖名人，并记载了始祖是“胡昌翼”。根据家谱的本体模型，始祖、特别是始迁祖有一个属性是迁徙事件（shl:migration），每个迁徙事件包含迁徙人、迁徙时间和迁徙地点，即何人何时从何地迁往何地，首先抽象出所涉及的概念，再将概念形式化为知识本体，即用有意义术语来表示概念所对应的类（Class）和概念间关系所对应的属性（Property），并用机器可理解的语言（RDFs）和格式（RDF/Turtle）来编码。如下图所示：



shl:Migration a rdfs:Class ;
rdfs:label " 迁徙 " ;
rdfs:subClassOf shl:Resource ;
rdfs:comment " 一次迁徙事件。" .
shl:originalLocality a rdf:Property ;
rdfs:label " 原居地 " ;
rdfs:domain shl:Migration ;
rdfs:range bf:Place ;
rdfs:comment " 迁徙之前的居地。" .
shl:locality a rdf:Property ;
rdfs:label " 迁居地 " ;
rdfs:domain shl:Migration ;
rdfs:range bf:Place ;
rdfs:comment " 迁徙完成后的居地。" .

图 1 迁徙信息结构化流程图

将家谱元数据中记载的迁徙事件根据知识本体定义的数据结构，提取其中的人、地、时等实体，以 RDF 数据格式描述，将人、地、时串联成一个个的迁徙事件，可将具有共同先祖的先祖名人及其迁徙事件关联起来。以下是上图家谱知识库原型系统中的一次迁徙事件的 RDF 数据片段。

```
    // 一次迁徙事件
    <http://gen.library.sh.cn/Migration/1> a shl:Migration ;
    bf:eventAgent <http://gen.library.sh.cn/Person/16607> ;
    shl:temporal <http://gen.library.sh.cn/Temporal/14> ;
    shl:locality <http://gen.library.sh.cn/Place/1129> ;
    shl:originalLocality <http://gen.library.sh.cn/Place/203> .

    // 本次迁徙事件的主体（始祖胡昌翼）
    <http://gen.library.sh.cn/Person/16607> a shl:Person ;
    foaf:name " 胡昌翼 " ;
    shl:courtesyName " 宏远 ";
    shl:pseudonym " 眉轩 " ;
    shl:temporal <http://gen.library.sh.cn/Temporal/14> ;
    shl:description " 唐昭宗幼子 " .

    // 本次迁徙事件的始发地
    <http://gen.library.sh.cn/Place/203> a shl:Place ;
    bf:label " 长安 " ;
    shl:city " 西安市 " ;
    shl:province " 陕西省 " ;
    geo:long " 108.93" ;
    geo:lat " 34.27" .

    // 本次迁徙事件的目的地
    <http://gen.library.sh.cn/Place/1129> a shl:Place ;
    bf:label " 考川 " ;
    shl:county " 婺源县 " ;
    shl:city " 上饶市 " ;
    shl:province " 江西省 " ;
    shl:town " 紫阳镇 " ;
    shl:village " 考水村 " ;
    geo:long "117.757787" ;
    geo:lat "29.27642" .

    // 本次迁徙事件发生的时间信息
    <http://gen.library.sh.cn/Temporal/14> a shl:Temporal ;
    shl:dynasty " 五代后唐 " ;
    shl:beginYear "923"^^xsd:int ;
    shl:endYear "936"^^xsd:int .
```

下图是利用百度 E-charts 这个可视化工具，根据上述 RDF 数据自动生成的上川明经胡氏家族迁徙图，可视化地展示自五代后唐至清末民初的家族迁徙路线，挖掘掩埋在历史角落和时间长河中的故事。这个故事完整地讲述了上川明经胡氏的始祖“胡昌翼”在唐末自都城长安迁往婺源考川，其后代又自考川迁往安徽绩溪等不同散居地的迁徙历程。

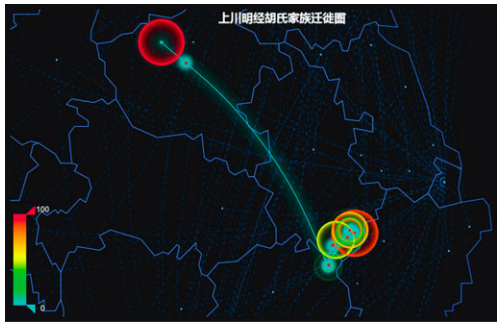


图 2　上川明经胡氏家族迁徙图

这样，系统自动地将尊奉同一个始祖的多种家谱中的先祖名人以迁徙事件为链接点关联在一起，根据其迁徙时间、地点等数据，家族的迁徙路线及后代散居地由系统自动生成可视化的迁徙图。

这样的方式有利于读者直观获得隐藏在不同文献中的知识，并将不同的文献按照某一主题有机地组织起来，提供知识导航。另一方

面，还可以帮助用户发现问题，提出问题。例如上述迁徙图直观地告诉我们，在五代后唐时期，上川明经胡氏的始祖胡昌翼经历了一次从长安到婺源考川之间的远距离迁徙事件，在那种交通极不方便的时代，为何要进行如此远距离的迁徙？这个事件背后的原因是什么？如果用户按照上述迁徙图指引的方向找到相关家谱文献的描述元数据，会在摘要字段中看到“始祖昌翼，谱称本唐昭宗幼子，母何后，天祐元年生，避祸来婺源考川，遂冒义父胡三公姓”。这句话意为胡适始祖胡昌翼本为唐朝最后一位皇帝昭宗之子，胡适为李唐皇室后代。这样有明显戏剧性的故事会引发用户进一步探索家谱全文的兴趣。

家谱最重要的内容是世系表，世系表详细记载了从始祖到家谱修纂时家族所有男性成员及其配偶子女的生卒、婚娶、迁徙、生平大事等，其中存在着丰富的数据、事实和知识，但目前却以扫描图片的形式供人阅览。上图也开始尝试将图片形式的世系表转换为结构化的数据，以下是“胡适”这个人物的 RDF 描述数据，描述了“胡适”的基本信息以及与其他资源之间的关系，如人与人之间的关系：父子关系、夫妻关系，人与文献间的关系等，这些关系均以由知识本体严格定义的属性作为谓词来表达，以人机均可理解的 Turtle 格式编码。

```
<http://data.library.sh.cn/jp/entity/person/
      uqwn21pa7iah9p87> a shl:Person ;
      shl:name " 洪骅 " ;
      foaf:name " 胡适 " ;
      shl:courtesyName " 适之 " ;
      foaf:familyName <http://data.library.sh.cn/
authority/familyname/rvmgzfsec8os93mv> ;
      shl:orderOfSeniority "3" ;
      shl:birthday " 光绪辛卯十一月十七未时 " ;
      shl:description " 考派美国留学生名适字
适之事迹见学了生光绪辛卯十一月十七未时
娶江氏生光绪庚寅十一月初八辰时 " ;
      shl:relatedWork <http://data.library.sh.cn/
jp/resource/work/jklhb5c3ga1rvxe3>;
      shl:roleOfFamily<http://data.library.sh.cn/
jp/vocab/ancestor/xian-zu>;
      rel:spouseOf <http://data.library.sh.cn/jp/
entity/person/kvop3qdif1okxhoa>;
      rel:childOf <http://data.library.sh.cn/jp/
entity/person/mc7khkqgql3rpceb> .
```

这样的结构化数据还可用于统计和分析，如从《上川明经胡氏宗谱》世系表中析出 8915 名男性及其配偶 4733 人，对配偶的姓氏进行统计后，发现“李”姓虽为国人大姓之一，却没有出现在这 4733 名上川明经胡氏的配偶之中，而曹姓最多，有 982 人，如图 3，原因在于同姓不通婚是中国人婚娶中的一个重要原则。

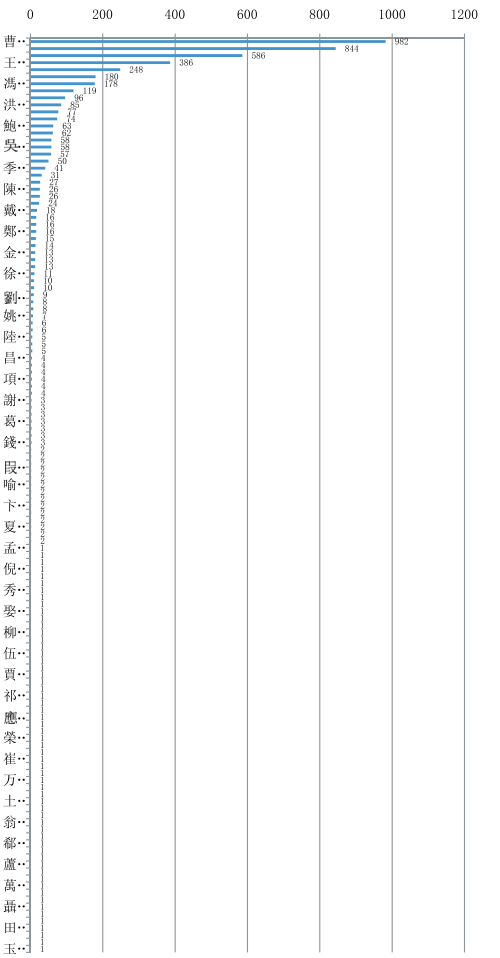


图 3 《上川明经胡氏》世系表中配偶姓氏出现次数统计

这从数据的层面进一步为“始祖昌翼，谱称本唐昭宗幼子，母何后，天祐元年生，避祸来婺源考川，遂冒义父胡三公姓”这段描述提供了证据，这样的证据显然比文字更有说服力。

上述例子呈现了一个家族的迁徙情况，系统还可以支持某个地域或时间范围内的迁徙情况分析。在描述家谱文献的元数据中，有一个“居地”(也叫“谱籍”) 元素，意为修谱时该家族的居住地，一般著录到县一级别，在上图馆藏家谱元数据中，依据照实著录的原则，使得很多居地以古地名著录，比如“苏州”，古称“吴县”，而在《中国家谱总目》中，则以今地名“苏州”著录。还有的情况是，有一些家谱的居地为同一个县名，但这个县名却存在于不同的省份中，这就涉及地名的规范控制问题。对于大部分普通用户来说，很难知晓一个地方的所有不同名称或一个县级地名所属的省份。而规范控制要解决的问题是，当读者需要了解某个地方的家谱时，合并同地不同名的情况，区分同名不同地的情况。

在互联网环境下，规范控制的本质是基于概念的匹配而非基于字符的匹配，而基于HTTP URI、RDF、知识本体的关联数据技术正是这样一种解决方案[7]。当一个地点被HTTP URI所唯一标识，并被赋予一个本体所定义的概念，且以RDF数据描述后，这个地点就不再是一个以字符串表示的地名，而是一个与实际存在的地方对应的资源对象，包含丰富的RDF描述数据，不仅可以在互联网上访问，其由本体定义RDF描述数据（包括今地名、古地名、别名、所属省市、经纬度等）可被机器读取和处理，因而可以处理同地不同名，同名不同地的问题，也可以分析一个地名所辖的所有下级地名的情况。

以“四川省”的数据为例，当所有四川省所辖的下级县都通过“gn:parentADM2”这个属性与“四川省”产生关联，系统即可方便地将所有居地属于四川省这个地域范围内的所有家谱进行动态地聚类，得到的家谱文献数量是956种，还能对四川的迁徙情况进行聚类，发现有记载从“湖北省麻城市”迁徙到“四川省”的家谱有221种，接近总数的四分之一。这样的比例是较为引人注意的，也会引发用户问出为什么。实际上，这一数据可作为研究中国人口迁移史上“湖广填四川”这一论题的佐证。著名的历史地理研究专家葛剑雄教授在《中国移民史》一书中提到，家谱中对明清时期大量移民自“麻城孝感乡”迁往“四川省”这一史实有大量的记载，而却少见于正史和方志。家谱作为与正史、方志一起作为历史研究的重要史料，得到进一步证明。

```
<http://data.library.sh.cn/entity/place/
    topq2nlfiugtg964>
 a shl:Place;
    bf:label          “四川"@cht" 四川"@chs;
    shl:abbreviateName          "川";
    shl:country "中國"@cht" 中国"@chs;
     owl:sameAs          <http://www.
geonames.org/1815285>;
    gn:lat     30.67;
    gn:long          104.07.
    <http://data.library.sh.cn/entity/
place/1v1smmtsmqg29rlf>
    a shl:Place;
     bf:label        "樂山"@cht" 乐山"@chs;
       shl:province "四川省";
        gn:parentADM2        <http://data.
library.sh.cn/entity/place/topq2nlfiugtg964>;
        ……
```

图书馆作为数字人文服务的提供者，不是要代替用户回答问题，而是启发用户提出问题，并为用户深入探寻大量史料提供更为方便的方法和工具支持。这样有望提高资源的利用率，也有助于吸引更多用户，提高系统的黏度。

5　结语

上图的“家谱数字人文服务”是利用关联数据技术实现数字人文的尝试和探索。利用图书馆已有的元数据，从中挖掘事实、知识和数据，基于数据间的关联关系，形成知识地图，利用可视化技术和数据分析工具，帮助用户探索难以用肉眼和人脑完成的大规模海量资源，

从中找到自己所需。

本文所述的案例仅仅是一家的迁徙图和世系表，如果将上图所有馆藏家谱的迁徙信息和世系表结构化数据化，将为家谱资源的利用带来全新的革命——基于大量数据、结合时间、空间，对姓氏、人物及人物间的相互关系进行全景式的可视化展示和统计分析。然而，迁徙信息和世系表的数据化和结构化是一项大规模的工程，迁徙信息中包含大量的古地名和村镇名，需要历史地理名称的对应和转换来支持，目前上图也正在建设《历史地理名词表》这样的数据集和工具。而世系表则需要借助社会的力量以“众包”等形式共同参与，目前正在考虑在平台的层面支持这一功能。

另外，家谱中存在着“攀附权贵”“扬善避恶”“重男轻女”等情况，所以基于家谱数据得出的结论是否可靠还存在一定的疑问。但数字人文是一种技术和方法，而不是最终目的，可用来辅助人文研究而非代替人文研究；技术提供平台和工具，把决定权交给用户，用户生成内容，帮助用户发现问题，找到研究问题的证据，用户自己得出结论；除此之外，数字人文可提升用户服务体验，吸引用户利用图书馆的资源。基于关联数据来实现数字人文，可帮助更多的第三方数据服务发现资源，提高图书馆资源的利用率。

参考文献

[1] Claire Warwick, Melissa Terras, Julianne Nyhan. Digital humanities in practice[M]. London: Published in association with UCL Centre for Digital Humanities, 2012.

[2] 刘炜. 数字人文与关联数据[EB/OL]. 2014年图书馆前沿技术论坛, (2014-06-20)[2016-02-09]. http://society. library. sh. cn/IT4L2014.

[3] Nicole Coleman. et al. Digging into the Enlightenment: Mapping the Republic of Letters[EB/OL]. [2016-02-08]. http://www. clir. org/pubs/reports/pub151/case-studies/enlightenment.

[4] Kevin Page. A Humanities Web of Data: Publishing, Linking and Querying on the Semantic Web[EB/OL]. [2016-02-10]. http://digital. humanities. ox. ac. uk/dhoxss/2014/HumData. html.

[5] Konstantin Baierer, Evelyn Dröge, Kai Eckert, et al. DM2E: A Linked Data Source of Digitised Manuscripts for the Digital Humanities[EB/OL]. [2016-02-10]. http://www. semantic-web-journal. net/system/files/swj831. pdf.

[6] Jakob Huber, Timo Sztyler, Jan Noessner, et al. LODE: Linking Digital Humannities Content to the Web of Data[EB/OL]. [2016-02-12]. http://arxiv. org/pdf/1406. 0216v1. pdf.

[7] 刘炜, 张春景, 夏翠娟. 万维网时代的规范控制[J]. 中国图书馆学报, 2015(3).

夏翠娟 女，上海图书馆（上海科学技术情报研究所）系统网络中心研究开发部，高级工程师。研究方向：元数据、知识本体、关联数据、数字人文。E-mail:cjxia@libnet.sh.cn 上海 200031

张 磊 上海图书馆（上海科学技术情报研究所）系统网络中心研究开发部，高级工程师。研究方向：移动服务、数字阅读、数字人文。 上海 200031

（收稿日期：2016-07-28 修回日期：2016-08-05）