DOI:10.13530/j.cnki.jlis.160014

家谱关联数据服务平台的开发实践*

夏翠娟　刘　炜　陈　涛　张　磊

摘　要　数字图书馆对馆藏的揭示,沿袭传统的描述标准(如 MARC),多以文献特征为主,很难直接满足广大读者对文献知识内容进行查询的需求。关联数据技术通过构建关系明确的语义本体,能够很好地提供基于文献知识内容的揭示、导航和检索,通过开放数据重用和与外部数据的互联,丰富了数据的关联性,扩展了数据利用场景,释放了数据的潜能,为基于互联网的数据服务提供了一种基础设施。这是未来数字图书馆进行知识服务的应有之义。上海图书馆以家谱数据作为起点,尝试利用关联开放数据技术重组图书馆传统资源,构建历史文献数据服务平台。该平台经过基于 BIBFRAME 的本体设计,从 RDB 到 RDF 的数据转换,基于关联数据四原则的系统设计和基于语义技术框架的系统开发,支持面向万维网的书目控制,提供针对普通用户的寻根搜索服务和针对专业人士的数据挖掘服务。图 5。参考文献 11。

关键词　家谱　数据服务　关联数据　开放数据

分类号　G254

A Genealogy Data Service Platform Implemented with Linked Data Technology

XIA Cuijuan, LIU Wei, CHEN Tao & ZHANG Lei

ABSTRACT

The description of digital library resources has followed the traditional standard (such as MARC) in the past twenty years. The information, such as title, author, publication information, carrier information, etc, has been well described. However, in this way, it is difficult to directly meet the query requirements of the knowledge implicated in the content. Linked data technologies via building relationships among resources can provide a better way for knowledge organization, description, navigation and retrieval. By reusing and connecting with the open data, linked data technologies can help enrich the relationships among data, expand data using scene, release the potential energy of the data, and build the architecture of data service on the Web. Shanghai Library is trying to use linked data technologies to reorganize the traditional library resources in order to meet the requirements of data sharing, reusing, and also bibliographic control in the internet environment. And at the same time, try to build the historical data services platform which can meet differentiated users service needs. Firstly, we designed an ontology based on Bibliographic Framework (BIBFRAME). Secondly, we extracted the surname, person, place, time, event and other entities from the

* 本文系国家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”(编号:13CTQ008)的研究成果之一。(This article is an outcome of the youth project “The Application of W3C's RDB2RDF Standards in Building Linked Data Services” (No. 13CTQ008) supported by National Social Science Foundation of China.)

通信作者:夏翠娟,Email:cjxia@libnet.sh.cn,ORCID:0000-0002-1859-6979(Correspondence should be addressed to XIA Cuijuan, Email: cjxia@libnet.sh.cn, ORCID:0000-0002-1859-6979)

metadata records according to the ontology. Thirdly, we cleaned the data by merging, disambiguation and standardization, and supplemented information for some important properties (e.g. headstream of the surnames and GIS information of the places). Then, we assigned HTTP URI for each entity and described the entities based on the RDF abstract data model. By using the RDB2RDF data conversion tools which support W3C R2RML standards and the data processing tools called OpenRefine, we transformed the data format from RDB to RDF, and loaded the RDF data into RDF store called Virtuoso. Finally, we designed the system based on the four principles of linked data, and developed the system based on semantic technologies such as Jena, SPARQL, and other data visualization tools. So the system can support bibliographic control in internet environment. That means users can know the genealogy documents location information about nearly 600 organizations all over the world. The open access to all RDF data for the machines is based on simple technologies such as content-negotiation and Restful API. There are easy-to-use search services for those who just want to know about the stories of the surname and family, and advanced search services for those who want professional data mining and knowledge discovering. Most importantly, the platform allows authenticated users to contribute content by submiting comments and suggestions, or modify data directly. After other experts confirm, the modifications would be published openly. All comments and modifications would be recorded automatically. Linked genealogy data is the first project to provide open data services based on linked open data technologies in the area of libraries in China. There are some innovation meanings in the methodology of implementation, the process of development and the usage of technological tools. But it is just a starting point for Shanghai Library. There would be lots of work to do about the authority data, which is still insufficient. And there are more external data sets such as Geonames, DBPedia, VIAF and so on need to mashup with the local data. Finally, there are some unresolved problems such as geographical names authority control in a historical view. 5 figs. 11 refs.

KEY WORDS

Genealogy. Data service. Linked data. Open data.

0 引言

开放数据是互联网发展的一个新趋势,数据作为一种极其重要的资源逐渐在世界范围内形成共识。在开放数据大潮中,政府和公共机构拥有最多的公共数据,是数据开放运动的先锋[1]。2009 年,Data.gov 在美国正式上线,吹响了数据开放运动的号角,澳大利亚的 Data.gov.au,英国的 Data.gov.uk 也紧随其后,到 2010 年 11 月,欧盟委员会首次提出“欧盟开放数据战略”,将数据开放运动推向高潮[2]。《纽约时报》、英国广播公司等媒体已先后成功实施,图书馆行业也是数据开放运动的积极拥护者,瑞典、美国、匈牙利、英国、德国、西班牙、韩国、日本等国的国家图书馆以及 OCLC 陆续将自己的书目数据或规范数据以关联数据的形式发布,美国国会图书馆还牵头开展书目数据格式标准的关联数据化。

上海图书馆(以下简称“上图”)一直非常关注开放数据运动,很早就开始跟踪、研究和尝试开发相关技术,认为这是把数字图书馆带入以数据技术为特征的下一代互联网的新契机。对于上图来说,大量的历史文献资源,如古籍、家谱、尺牍、近代文献、民国文献、档案、照片、笔记、手稿、小报等,虽然从纸质文献到电子文件的数字化工作一直在进行,也一直在提供基本的文献检索服务,然而,想要更好地满足读者需

求，必须将其所包含的知识内容描述出来，利用新的技术在互联网上提供服务，让更多人在使用的同时，能够参与系统的优化、迭代和内容建设，进而实现系统的平台化，使其成为读者从事相关学习、交流和研究活动的必经之所。

中文家谱是上图最重要的特色文献之一。经过长期的研究和整理，上图已取得了一批具有影响力的成果，例如，编纂（或主持编撰）出版了《上海图书馆馆藏家谱提要》《中国家谱总目》《中国家谱通论》《中国家谱资料选编》等。尤其是《中国家谱总目》，收录了来自港、澳、台地区和日、韩、北美、德国、加拿大、澳大利亚等地近600余家机构收藏的五万四千余种家谱，包含608个姓氏，析出先祖名人七万多个，谱籍地1 600多处，堂号三万余支，不仅是一部华人家谱的联合目录，还是一部中华家谱知识的百科全书。这些宝贵的整理成果目前仅以纸质和影像文件的形态存在，大量的内容研究和标引揭示也只是以出版和提供简单的字段检索为目标，但这些成果正好为开发基于关联数据的知识服务平台提供了一个很好的基础。

经过数年的调研和探索，上图的技术研发团队认为，应用以关联数据为代表的新型数据管理技术，时机已经成熟。这些技术能够帮助图书馆充分利用长期积累的文献研究成果，将其中的数据、事实和其他知识点进行细粒度描述，利用网络知识组织的编码方法和技术手段，对馆藏资源进行重新组织，利用全网域的互联网平台实现图书馆的书目控制理想。对于家谱数据而言，在满足普通用户寻根服务的同时，针对人文研究学者提供分面可视化浏览、语义搜索乃至知识挖掘服务，有助于打破图书馆各类资源库相互隔离的封闭状态，推进数据开放，促进知识流动，在开放利用中充分发挥其多方面的潜在价值。

1 功能需求

1.1 图书馆的书目控制需求

图书馆承担着书目控制的职能和使命。由上图主持、众多家谱专家参与编纂、于2005年出版的《中国家谱总目》，对华人家谱文献在全球范围内的收藏分布情况进行了调查摸底、考证辑录，包含丰富翔实的文献著录信息和内容描述信息，是书目控制手段的一种尝试。如果能借助于网络环境下新的技术手段开发利用起来，将能随时随地提供服务，真正实现全网域范围内的书目控制。

具体而言，家谱总目的书目控制需求重点在以下三个方面。

第一，建立全球家谱联合目录，促进数据重用和共享。书目控制的一个重要职能是厘清某个资源在不同机构的收藏情况，并提供获取资源的线索。互联网技术的发展，使得建立一个全国、甚至全球的家谱联合目录成为可能。《中国家谱总目》已经提供了很好的数据基础。家谱服务平台首先要将已有的数据导入，为读者提供某种家谱现存的版本信息、在世界各地的收藏情况，以及获取文献的途径，同时促进这些由世界各收藏机构和家谱专家整理出来的知识在各机构之间的重用和共享。

第二，基于万维网的规范控制。规范控制是书目控制中重要的一环。“规范控制的本质是实现基于概念的描述和匹配”[3]。需要对各类规范实体，如人、机构、地、事件，从概念的层面进行区分，用统一的文字标签来表征概念。对于家谱来说，需要对姓氏、人、机构、地名、中国历史纪年中的朝代等概念建立模型，用明确的语义来表述这些概念和概念间的关系，并基于概念模型对各类实体进行规范控制，解决同一人名、地名、朝代名不同表述方式的认定、消歧与合并等问题。基于网络资源的规范控制，要求规范词能在全网域范围内被唯一标识和定位，规范词的语义描述信息能被机器获取、识别和理解。

第三，支持书目控制的可持续发展。作为纸质出版物的《中国家谱总目》，是静态的、封闭的，其书目记录截止到2003年。因此，需要有一个开放的平台，进一步补充和完善已有的书目

记录,如某种家谱中先祖名人的个人信息、家族的迁徙地名信息等。更重要的是还要能方便增加新的书目记录,包括新收录和即将收录的家谱书目记录。

1.2 差别化的用户需求

构建一个家谱服务平台,首先要从用户需求出发。上图自1996年开放国内第一个家谱阅览室以来,积累了丰富的用户服务经验,按照用户利用家谱资源的不同目的,可将用户需求分为以下几个层次。

第一,基于有限已知信息的寻根问祖。对于需要寻根问祖的普通用户来说,一般是已知某先祖名人的姓名,或家族的居地,或堂号,去搜寻相关的家谱文献,或相关的先祖名人、亲属的个人信息,如生卒年月、生平大事等。但这类用户想要的往往不仅仅是文献本身,还有文献中包含的内容(数据、事实和知识)。例如,关于家族中某先祖名人或亲属的详细信息、不同家谱文献中人物之间的亲属关系、家族的迁徙路线等。这个层次的需求虽然不复杂,但对查准率的要求较高,已有的基于关键词匹配的检索会导致大量的噪音,需要基于概念及概念间关系的匹配,准确定位到读者想要的结果,不仅要提供方便的文献获取途径,还应直接提供读者想要的内容。

第二,面向特定研究主题的知识发现。对于人文研究者来说,家谱是除正史、方志外重要的研究资料,它的独特价值已得到学界的广泛认同。但宝贵的知识财富隐藏在浩繁的家谱卷轶之中,若仅依赖于目前这种面向文献进行资源组织和基于简单的字段和关键词检索的系统,要穷尽数十万卷的家谱资料,探寻散落在历史角落和时间长河中的数据、事实和知识,是一件费时费力的工作。因而,具有良好用户体验的知识导航和知识发现功能就显得尤为重要。如根据姓氏、谱籍、堂号、朝代的聚类展示,基于概念和实体间关联的关系发现,发现不同家谱中人与人之间的关系、人与地之间的关系、某一家族在地理空间上的迁徙路线、某地域范围内某姓氏的分布情况、某一宗族的散居地覆盖范围,等等。

第三,基于用户贡献内容(User Generated Content,UGC)的知识进化和积累。家谱是同宗共祖的血亲团体记载本族世系和相关事迹、反映本家族繁衍发展过程的历史图籍。民间有大量对本姓、本族家谱有着深入了解和研究的团体和社群,他们既是图书馆的用户,也是家谱方面的专家。他们对某一姓、某一族的家谱了解比图书馆员更为全面、深入。如果能提供一个开放性的平台,不仅为他们提供图书馆收藏的家谱资料,还能为用户与用户、用户与收藏机构之间形成良性互动和交流提供方便,同时对交流过程中产生的知识进行组织、处理和保存,可以达到进一步完善家谱知识库,使知识在交流和传播中增值的目的。随着Web2.0技术的普及,基于UGC的"众包""众筹"理念深入人心,将用户行为纳入图书馆业务流程之中的举措,已屡见不鲜[4]。因此,新建设的家谱知识服务平台不应仅是静态的特色数据库,而应是支持知识不断生长和进化的有机体。

2 设计实现

选择利用关联数据技术来实现家谱知识服务平台,是因为它基于领域概念体系(知识本体)而非文献来组织知识,用"主谓宾"这种普适的数据模型(RDF)来表示和检索知识,借助发展成熟的数据校验和知识挖掘工具支持知识的维护和更新,允许用户访问文献中的部分数据而非整个文献。另一方面,关联数据已在图书馆界得到了广泛而深入的应用[5],形成了一整套基于元数据和知识本体、RDF数据转换、RDF数据存储和查询、数据可视化的实现技术、方法和流程,可以很好地满足书目控制和规范控制、数据重用和共享、知识组织和知识发现的功能。

上图家谱知识服务平台的设计,经历了如

下流程。首先，设计一个向下兼容、易于扩展、便于重用和共享、支持家谱数据重组和知识建模的家谱知识本体，明确定义家谱资源中涉及的人、机构、地、事件等概念及其相互关系。接下来，对已有的家谱元数据进行数据清洗，提取各类概念实体，赋予 HTTP URI。基于 RDF 抽象数据模型，对实体及实体间的关系进行描述，必要时与外部数据关联，丰富数据的语义。数据以机器可读的 RDF 序列化格式编码后，存储于专用的 RDF 数据库中。最后，基于关联数据四原则发布数据，利用语义技术开发框架存取操作数据，利用可视化技术展示数据，利用 Web2.0 技术支持用户贡献知识，实现知识导航、知识发现和知识进化的功能。

2.1 基于知识本体的关联数据模型设计

知识本体是领域知识被抽象后形成的可共享可重用的概念模型，通常表现为一套体系化的术语词表及对相互之间关系的形式化描述，以一定的机器语言编码后可被机器识别和处理的代码体系。知识本体为数据赋予语义，是数据中所含知识的容器。基于尽可能复用已有术语词表的本体设计原则，上图的家谱本体主要基于美国国会图书馆的书目框架 BIBFRAME 2.0，复用了 FOAF、Geonames、Schema.org 等词表的部分术语，之后自定义了一批家谱资源特有的属性。

家谱知识本体以 BIBFRAME 为基础框架，一方面，书目框架（BIBFRAME）是美国国会图书馆牵头开发的下一代书目数据格式标准，用以取代 MARC，并能为图书馆、档案馆、博物馆、美术馆等相关“人类文化记忆机构”共同使用，有良好的包容性、可扩展性和开放性，其词表能很好地描述家谱资源的文献特征；另一方面，BIBFRAME 同时还是一个为书目数据关联数据化而设计的关联书目数据模型，其“作品（Work）—实例（Instance）—单件（Item）”的核心模型是书目记录功能需求（FRBR）的简化[6]，能很好地满足书目控制的需求，其数据模型包含人、机构、家族、事件等概念，也适用于家谱资源内容相关实体的描述，满足规范控制的需求。

家谱本体还复用了 FOAF 中的术语，用来描述家谱中的先祖名人，自定义了“谱名、字、号、谥号”等中国历史人物特有的属性予以补充。复用 Geonames 的术语来描述家谱中的谱籍地，复用 Schema.org 和 W3C Organization 的术语来描述家谱资源涉及的收藏机构，复用 W3C Time Ontology 来描述时间信息，自定义了一些属性来描述家谱资源涉及的中国历史朝代信息。

为了便于家谱本体的共享和重用，上图已将家谱本体以 RDFs 和 OWL 编码在 Web 上公开发布。为了方便业内专家深入了解家谱本体，网站提供三种视图模式供用户浏览。模型视图（Model View）可视化地展示了家谱本体类和属性间的关系；类视图（Class View）通过父类和子类的层级关系浏览类和属性；列表视图（List View）按照类和属性名的首字母顺序排列展示类和属性。网站上也可以打包下载家谱本体的全部 RDF 数据（见图 1）。

2.2 标准开放的数据格式——从 RDB 到 RDF

关联数据的第三个原则是要求数据为 RDF 格式。RDF 数据抽象模型及其各种序列化格式如 RDF/XML、Turtle、JSON-LD 等，是 W3C 的推荐标准规范，是跨平台的、开放的、可被各种程序语言处理的标准数据格式[7]。

家谱知识服务平台的 RDF 数据基于已有的元数据生成，除了《中国家谱总目》的元数据外，还有上图新增的馆藏家谱元数据。首先要从元数据中提取作品、实例、单件、人、机构、地名等实体，分别赋予 HTTP URI，用家谱本体定义的类和属性来描述这些实体及实体间的关联关系。笔者在《基于书目框架（BIBFRAME）的家谱本体设计》一文中详细列出了家谱元数据和家谱本体的对应关系。

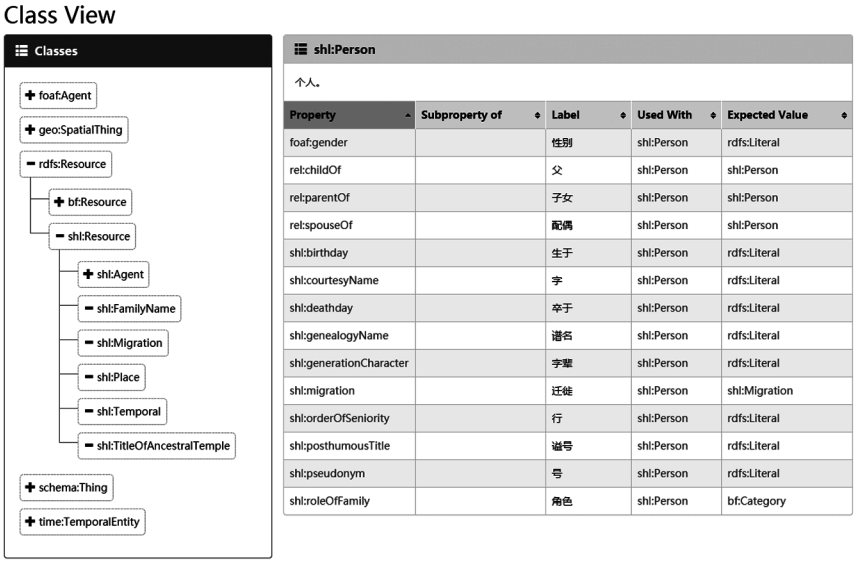


图 1　以关联数据发布的上海图书馆家谱本体

《中国家谱总目》数据存储于 EXCEL 表格中，馆藏家谱数据以 MARC 格式存储于 SQL Server 中，都可以看作是“记录—字段—字段值”的 RDB 数据格式，因此需要将 RDB 格式的数据转换为 RDF 格式（这个过程一般被称为 RDB2RDF）。主要使用两种开源的自动转换工具来完成 RDB2RDF 工作：一个是支持 W3C RDB2RDF 标准规范的 DB2Triples，另一个是 OpenRefine。这两个工具都支持将 RDB 中的表和字段与家谱本体中的类和属性建立映射，定义 URI 的生成规范，自动生成 RDF 数据。不同的是 DB2Triples 应用了 W3C 的 R2RML 标准规范，支持一次性获取多个关系数据库表的数据，生成多类实体的 RDF 数据，缺点是本体的映射需要用 JSON 语言编辑文本格式的配置文件，缺少友好的用户界面[8]。而 OpenRefine 不方便同时操作多个表的数据，但却有“所见即所得”的用户界面。因此，在转换存储于 SQL Server 中多个关系数据库表的馆藏家谱数据时，采用的是 DB2Triples，在转换存储于单个 EXCEL 表格中的《中国家谱总目》数据时，采用的是 OpenRefine。图 2 以家谱中世系表中的先祖名人为例，展示了利用 OpenRefine 将 EXCEL 表格中的数据转换为 Turtle 格式的 RDF 数据的过程。

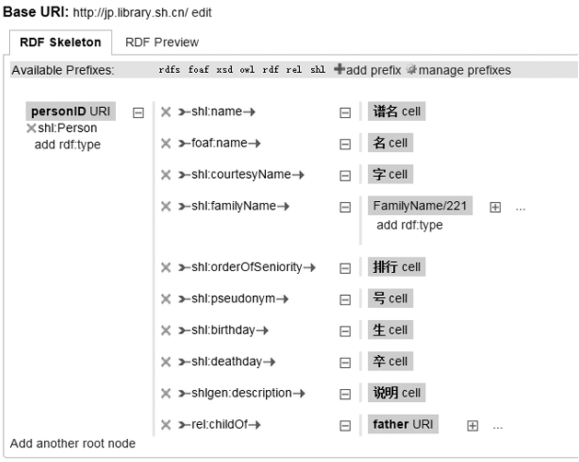
2.3　基于关联数据四原则和语义技术框架的系统实现

系统的设计遵循了关联数据的四原则。调研了 Cool URIs 规范[9]及国际上政府领域和各大图书馆的关联数据项目，根据实际需求，制订《上海图书馆 URI 设计规范》，以此为依据为家谱数据中的各种实体生成 HTTP URI。关于实体的描述信息，以基于 RDF 抽象数据模型来组织，并以标准的序列化格式来编码。访问实体的 HTTP URI 时，可获得关于实体的 RDF 信息。支持内容协商（Content Negotiation）机制，当用普通的浏览器访问时，系统返回供人阅读的 Html 页面，当用语义浏览器或语义代理（程序）访问 URI 时，系统按照请求方通过 Http Header 传送的关于内容格式的请求返回相应格式的 RDF 数据，如 RDF/XML、RDF/Turtle、JSON-LD 等。

系统实现上基于成熟的语义技术和开源框架。利用 RDB2RDF 和 OpenRefine 等工具对原来存储于关系数据库和 EXCEL 表格中的元数据记录进行清洗和转换后，以 Tuttle 格式输出生成

代	personID	father	谱名	名	字	号	排行	生	辛	迁徙	说明
42	16810	33134	洪辞	适	适之		三	光绪辛卯十一月十七未时			考清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

EXCEL 表格中的原始数据



本体映射定义

```
<http://jp.library.sh.cn/Person/16810> a shl:Person ;
        shl:name "洪辞" ;
        foaf:name "胡洪辞" ;
        shl:courteeyName "适之" ;
        shl:familyName <http://jp.library.sh.cn/FamilyName/221> ;
        shl:orderOfSeniority "三" ;
        shl:birthday "光绪辛卯十一月十七未时" ;
        shl:description "考清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" ;
        rel:childOf <http://jp.library.sh.cn/Person/3313> ;
        foaf:name "胡洪辞" .
```

Turtle 格式的 RDF 数据

图 2 RDB 到 RDF 的映射和转换

的 RDF 数据，存储于专用的 RDF 存储库中（Open Link Virtuoso）。RDF 存储库和可视化展示层之间用 RDF 查询语言 SPARQL 实现数据的查询和存取，利用 Jena 作为开发工具来实现对 RDF 数据的处理，并利用 SIMILE Timemap、Baidu Echarts、高德地图等工具实现数据的可视化展示。整个开发框架如图 3 所示。

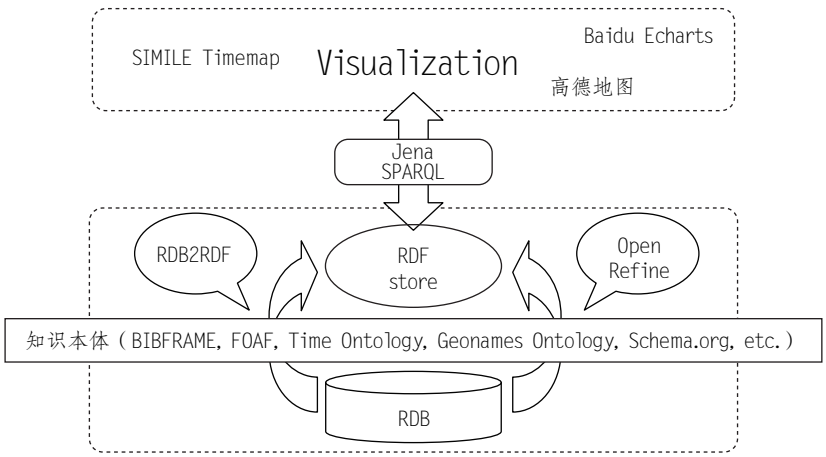


图 3 基于语义技术的开发框架

2.4 面向书目控制、知识发现和知识进化的功能设计

家谱知识服务平台在功能设计上,主要满足三方面的需求:一是满足图书馆的书目控制和数据共享的功能,包括对一种家谱的版本、复本、收藏单位的全面呈现和各类内容实体的规范控制,以及用于数据重用和共享的数据消费接口;二是面向普通大众和人文研究人员的知识发现功能,包括基于概念匹配的检索和基于人、地、机构、时间之间关联关系的可视化浏览;三是面向领域专家的知识进化功能,支持用户对已有的数据进行修正和补充,并对这些知识进行保存、组织和处理。

2.4.1 书目控制

家谱书目控制需求主要由 BIBFRAME 核心数据模型(作品—实例—单件)来支持,但在实际应用中,需要根据家谱描述的具体需求进一步简化。如果严格按照 BIBFRAME 的定义,作品—实例—单件之间的实体关系模型是 1∶n 和 1∶n 的关系,即 1 个作品对应多个实例,1 个实例对应多个单件。这样,当把一种家谱作为一个“作品”,那么内容相同的同一种家谱的不同版本(某一刻本的复印本或影印本)可作为不同的实例,这样可以对每一次制作(出版)的情况进行描述,如“出版”时间、地点等。但对于目前已有的家谱数据来说,虽然保留了木活字本的出版时间、机构,但没有保留复印本或影印本的详细信息,只保留了馆藏信息。如《上官氏四修族谱》五卷有 1936 年天水堂印制的木活字本,保存在“寻源姓氏”和美国犹他家谱学会,还有分别保存在福建省图书馆和漳州市台湾工作办公室林嘉书处的复印本,而馆藏信息正是“单件”的显著特征,因而我们决定将这些复印本都当作“木活字本”这个实例的不同“单件”。因此,家谱的作品—实例—单件的实体关系模型是1∶1∶n的关系,如图 4 所示。

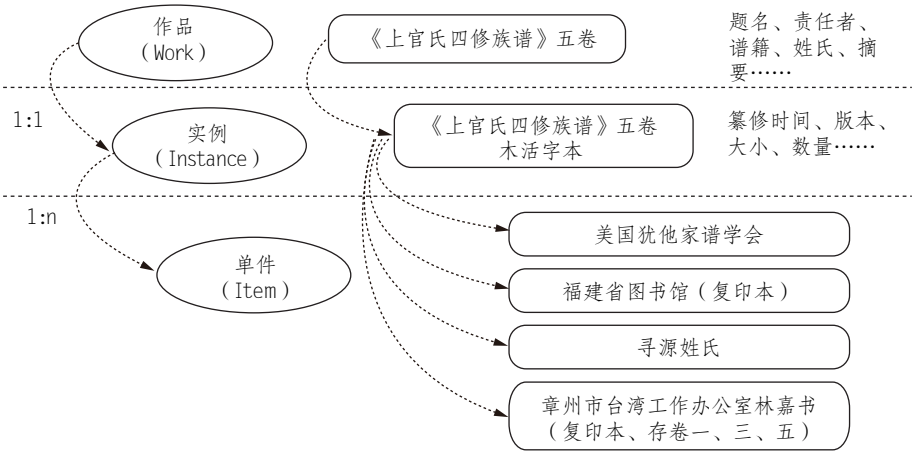


图 4 上海图书馆家谱书目控制模型

2.4.2 知识发现

(1)面向普通大众

为满足普通大众随意浏览和寻根问祖的需求,在界面及功能的设计上尽量考虑到简洁性、新颖性和趣味性。在首页的设计上,采取按姓氏直接呈现相关家谱文献和先祖名人统计数据的方式,吸引用户点击,让用户直接接触到数据和知识,便于发现信息,并吸引用户继续探索。这样,用户无需专深的知识,即可从姓氏开始,了解该姓氏的来龙去脉和历史上出现过的名人,获得认同感,同时发现本平台上收录的先祖名人及所出家谱文献的详情。点击“馆藏信息”,可了解文献的所有馆藏机构及其所在地,如该家谱在上图有馆藏,则会直接显示其全文

影像的链接。首页的右半区域还兼备简单检索入口和高级检索入口的功能，以自动轮播的方式出现。用户也可以直接输入检索词查询。

对于没有明确目标，不知道具体的家谱题名和谱籍地名的文献查询需求，可利用时空图来探索。时间被做成一个旋钮，当用户转动旋钮或输入年份时，即可显示在当前年份纂修的家谱，并依据该家谱的谱籍地在地图的相应位置插上旗标，点击旗标，显示所有符合条件的家谱。

（2）面向专业研究人员

研究人员包括专门研究某一姓氏某一家族的家谱专家，也包括借助家谱文献为特定研究主题提供资料和佐证的其他领域研究人员。针对此类用户，平台设计了基于概念匹配的高级检索和基于时空关联的发现探索功能。

高级检索支持基于姓氏、责任者、谱籍地名、堂号、先祖名人、馆藏机构等概念实体的精确查询。以馆藏机构为例，作为《上川明经胡氏宗谱》馆藏地的“上海图书馆”不仅仅是一个名称，而是作为一个实体，不仅有“上海图书馆”全称，还有“上图”简称，还包括“上海市淮海中路1555号”这个地址信息，更重要的是它有一个全球唯一的标识符：“http://data.library.sh.cn/entity/organization/11v6pvzycw 5419sy”。所以当输入“上海图书馆”或“上图”时，或它们的繁体形式，都能通过唯一标识符定位到这个实体本身，并找到馆藏机构为这个实体的所有家谱文献。因而，无论用户输入该实体的哪个属性值，繁体或简体，检索的结果均保持一致。重要的是，这些描述信息在数据底层持久存在，不依赖于系统的功能和逻辑，可以实现在跨平台跨系统的传输和交换过程中不损失语义。

查询是利用 SPARQL 检索语言来实现的。SPARQL 是 RDF 数据专用查询语言，是直接面向知识的查询，与数据库的物理存储结构无关，只与数据本身的内在知识逻辑有关[10]。

谱籍地名也是如此，不再作为一个以字符串存在的名称，而是与一个真实存在的地点相对应的实体。这个实体不仅有不同的名称，还有经纬度等 GIS 信息，可以在地图上准确定位。因而可以开发出基于地图的发现和探索功能。此外，还有两种方式提供更为精准的浏览。一种是“时间轴—地图”浏览，可以拖动时间轴，发现在某段时间中纂修的家谱在地图上如何分布；另一种是“地图画圈浏览”，在地图上画一个范围，显示所有谱籍地在某地域范围内的家谱。具体如图 5 所示。

2.4.3 平台优化

提供用户交流和互动的平台，支持研究专家、学生、民间团体等相关领域的用户通过留言反馈直接修改数据，贡献知识，可以使数据在使用过程中实现增值与增殖。

（1）反馈交流

具有上图读者证或经上图网上注册系统注册的用户可直接登录平台，登录后通过撰写反馈意见提出关于某种家谱的问题，可以是关于该家谱的著录信息的疑问，如题名、纂修者，也可以是与该家谱的内容如先祖名人、迁徙情况相关的见解与意见。如对家谱题名、纂修者、谱籍地名等数据存在异议并提交反馈后，经专家查证采纳后可在平台上修正。

（2）数据修改

经过认证的专家登录平台后可直接修改数据，经其他专家查证审核通过后发布，系统会记录每一次修改。可修改的数据包括姓氏、谱籍地名、先祖名人等实体的描述信息，也可修改家谱文献的著录信息。修改界面依据本体的定义自动生成，如果取值范围（Range）为另一个实体（如责任者、谱籍地名、姓氏等），只能通过选择已有的实体或新建实体，也就是说，如需修改责任者，不再是修改名字这个字符串，而是去修改这个实体的属性值，如姓名、字、号、所处朝代、生平大事等。数据修改并保存后，系统会记录何人于何时改了哪一个属性，修改前和修改后的值分别是什么。经专家审核通过并发布后，可在前台界面上呈现最新结果。

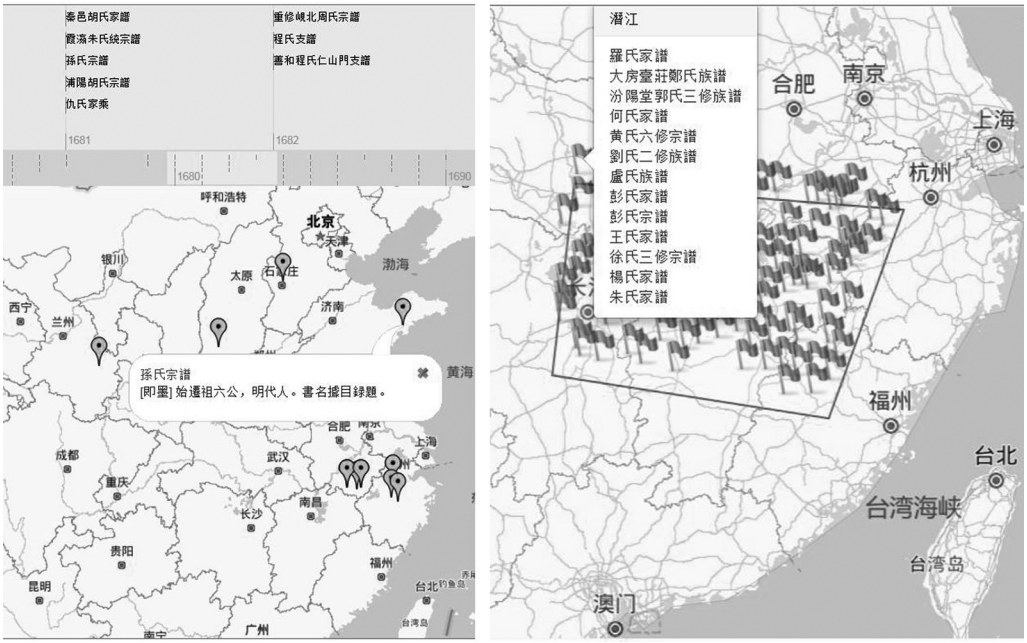


图 5　基于时空关联的家谱知识发现

3　关联开放数据的发布和消费

3.1　公共数据的开放

在家谱数据清洗过程中，首先从已有的家谱元数据中析出词表数据，然后进一步规范和补充，形成规范词表。如，用于描述先祖名人所处时代的时间信息，都以中国历史纪年的方式著录，而用于描述家谱纂修时间的信息大部分以公元纪年著录，少部分残谱只能粗略地以朝代纪年著录。为了能在时间的表述上做到统一，满足检索结果排序或在时间轴上定位的功能，就必须实现中国历史纪年和公元纪年的对照。为此，我们整理了从公元前 841 年到民国时期的“中国历史纪年表”数据。考虑历史纪年和公元纪年对照的功能不仅可用于家谱，还可用于其他历史文献资源，不仅可用于上图，还可用于其他有类似需求的机构，所以将其以关联开放数据的形式在网站 http://data.library.sh.cn 上公开发布，并利用关联数据消费技术中最常用的 Restful API 技术[11]，提供数据开放接口供程序调用。开发人员只需在网站上注册一个 API Key，即可调用该 API 进行中国历史纪年和公元纪年之间的相互转换。以下为输入明朝返回明朝的起止年份的 API 调用方法：

http://data.library.sh.cn/time/data/明？key=yourAPIKey，返回数据为：“1368-1644”

除此之外，还有用于描述谱籍地名的“地理名词表”、用于定位家谱收藏机构的“收藏机构名录”等。这些数据在《中国家谱总目》中作为文后索引，只有词汇，没有关于这个词汇的更多信息。为了更好地利用这些地理信息，对其做了进一步的补充。如“上海”这个词，为其增加了“所属国家”“下辖行政区域”“经纬度”等信息，为收藏机构增加“全称”“简称”“地址”等信息。不仅可以实现在地图上定位的功能，还能丰富数据间在地理维度上的关联关系。为了更好地实现数据在互联网上开放和互联的目的，这些数据集也以关联开放数据的形式在 data.library.sh.cn 上发布。

3.2 家谱数据的开放

家谱知识服务平台中的数据，按照实体的类型区分，包括姓氏、先祖名人、堂号、谱籍地名、收藏机构、家谱书目数据（题名、纂修者、纂修时间、版本信息、馆藏信息等）。其中，谱籍地名和收藏机构从元数据中提取出来，经进一步规范和补充后，已作为公共数据集发布，而姓氏、先祖名人、堂号、书目数据等家谱资源特有的数据也以关联开放数据的形式开放。这些数据依托家谱知识服务平台，以《中国家谱总目》为基础，吸收业界专家的知识，不断更新。这些更新的、经过同行专家审核认可的数据也将在开放数据中实时地反映出来，供社会各界共享和重用，在不侵犯个人隐私以及不违反法律法规的框架下，可利用这些数据创造新的服务。

家谱数据的开放依托于 data.library.sh.cn 网站，以基于 Http URI 的内容协商、Restful API、Sparql Endpoint 等技术手段开放数据。开发人员利用 Http URI 的内容协商和 Restful API 即可获得所有家谱知识服务平台中的家谱数据，以题名、责任者、姓氏、先祖名人姓名、谱籍地名、堂号、馆藏机构名、纂修时间、摘要中的关键词的任意组合作为输入参数，即可返回所有匹配的家谱文献的 RDF 数据，并依据 RDF 数据中关联的姓氏、人、地、时间、机构的 URI，利用内容协商功能，获取更多关于这些实体的 RDF 数据。RDF 数据将以 W3C 的推荐标准 JSON-LD 格式输出，便于开发人员编程处理。以下为获取所有谱籍地为麻城的夏氏家谱数据的 Restful API 调用方法：

http://data.library.sh.cn/jp/data/familyName=夏&place=麻城?key=yourAPIKey

对于能够熟练使用 SPARQL 查询语言的开发人员，可利用 Sparql Endpoint 进行更为复杂的查询和数据获取。

data.library.sh.cn 作为上图的开放数据平台，将陆续向互联网公开发布各种术语词表、规范档、馆藏书目数据等，并提供各种数据消费接口供开发人员调用，以促进基于互联网的书目控制、规范控制和数据共建共享。

4 效果、问题与展望

家谱知识服务平台利用现代信息技术手段，把专家的成果和大脑中的知识呈现于平台上，让大家分享、共享。通过利用资源和各类统计分析、可视化研究工具，使资源发挥更大的价值。作为国内率先推出关联数据技术开放数据的图书馆案例，平台一经上线，就得到了广泛的关注，《中国文化报》《文汇报》《新民晚报》等媒体纷纷报道。一些家谱爱好者和各民间家谱研究团体，先后百余人参与了平台的测评，积极反馈结果。

家谱知识服务平台构建过程中遇到了不少困难，其中最大的困难在于从已有的家谱元数据记录到家谱本体框架下 RDF 数据的映射和转换。已有的家谱元数据是图书馆行业按照传统的面向文献的标引方式著录产生的，而家谱本体强调文献属性的同时也注重内容属性，这就需要将原来的元数据记录打散，一一映射到家谱本体的框架下。虽然《中国家谱总目》的数据记录已经足够规范，但仍然存在数据不一致的问题，这就为数据转换工作带来了困难和障碍。例如，馆藏机构简称的不一致，“香港中文大学图书馆”在《中国家谱总目》的家谱文献馆藏信息中，有的被简称为“香港中大”，有的被简称为“香港中文大学”。这样的问题不是个别现象，会严重影响查全率。我们的解决办法是根据《中国家谱总目》的机构索引，将数据中所有“香港中文大学”修改为“香港中大”，这样无论用户输入“香港中大”还是“香港中文大学图书馆”，都能找到所有香港中文大学图书馆的馆藏家谱。但这是一种事后的补救措施，另一种解决方法是事先了解已有数据中存在的所有的不一致问题（显见是一项极为费时费力的工作），将“香港中大”和“香港中文大学”都作为“香港中文大学图书馆”的简称，在数据转换过程中或数据查询过程中，将馆藏地为“香港中大”和“香港中文大学”的家谱馆藏地都关联到“香港中文大学图书馆”这个实体。

另外一个更难处理的问题是古今地名著录

不一致的问题,比如“苏州”有时会按照文献记载如实著录为“吴县”,而在当前地图上,只能定位到“苏州”,没法定位“吴县”,这就需要将“吴县”作为今地名为“苏州”这个实体的另一个属性。但这样的情况普遍存在,需要依赖具有古今地名对照功能的“地理名词表”,这样的地名词表的建设和维护需要依赖更为专业的机构。厘清一个地名在历史上不同时间段的变化情况,对于图书馆来说也是一项艰巨的任务。

数据开放在国际上已成为一种常态,在国内也正在得到越来越多的重视。家谱作为上图最成熟和有一定影响力的资源,是率先提供数据开放实践的馆藏,未来还将在“上图历史文献数据服务平台”上开放、发布更多的资源。上图基于家谱等历史文献资源的数据开放尝试,一个重要的动因是希望能带动更多的图书馆和相关行业,促进数据开放和共享,避免重复建设和资源浪费,提升图书馆资源的价值。近期,上图还将基于开放数据接口开展数据应用开发竞赛。我们相信,数据的价值在于开放,用户是平台的真正主人,只有充分树立开放的理念,紧紧依靠用户,图书馆的创新才会永无止境。

参考文献

[1] 吴建中. 知识是流动的:出版界与图书馆界的新课题[J]. 图书馆杂志,2015,34(3):4-11.(Wu Jianzhong. Knowledge is fluid: new challenges to the publishing and library circles[J].Library Journal,2015,34(3):4-11.)

[2] Judie A,Fabrizio O,Simon S,et al. A systematic review of open government data initiatives[J]. Government Information Quarterly,2015,32(4):399-418.

[3] 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):22-33.(Liu Wei,Zhang Chunjing,Xia Cuijuan. Authority control for the Web[J].Journal of Library Science in China,2015(3):22-33.)

[4] 范丽娟.众包对图书馆的影响及其运用[J].图书馆建设,2011(1):89-92.(Fan lijuan. Influence and utilization of crowdsourcing in libraries[J].Library Development,2011(1):89-92.)

[5] Mitchell E T.Library linked data: research and adoption[R].Library Technology Reports,2013:10-15.

[6] Library of Congress. BIBFRAME 2.0 items[EB/OL].(2015-10-20)[2016-04-12]. http://www.loc.gov/bibframe/docs/pdf/bf2-draftspecitems-10-29-2015.pdf.

[7] Klyne G,Carroll J J,McBride B. RDF 1.1 concepts and abstract syntax[EB/OL].(2014-02-05)[2016-04-12]. https://www.w3.org/TR/2014/REC-rdf11-concepts-20140225.

[8] 夏翠娟,金家琴. 从关系数据库到关联数据:W3C 标准应用探析[J].图书馆杂志,2015(5):85-94.(Xia Cuijuan,Jin Jiaqin. On the application of W3C's RDB2RDF standards[J].Library Journal,2015(5):85-94.)

[9] Ayers D,Volker M.Cool URIs for the semantic Web[EB/OL]. (2008-10-03)[2016-04-12].https://www.w3.org/TR/cooluris.

[10] Harris S,Seaborne A,Prud'hommeaux E. SPARQL 1.1 query language[EB/OL].(2013-03-21)[2016-04-12].https://www.w3.org/TR/sparql11-query.

[11] 夏翠娟,刘炜. 关联数据的消费技术及实现[J]. 大学图书馆学报,2013(3):29-37.(Xia Cuijuan,Liu Wei. Technologies and implementation of consuming linked data[J]. Journal of Academic Libraries,2013(3):29-37.)

夏翠娟 上海图书馆系统网络中心高级工程师。上海 200031。

刘 炜 上海图书馆研究员。上海 200031。

陈 涛 中国科学院上海生命科学信息中心工程师。上海 200031。

张 磊 上海图书馆系统网络中心高级工程师。上海 200031。

(收稿日期:2016-03-14)