面向跨语言家谱服务的多源关联数据匹配研究

——上海图书馆开放数据应用比赛作品 Learn Chinese Surnames*

□董行

摘要 随着越来越多非英文关联数据集的发布，语言差异成了数据万维网中资源间相互链接的障碍。对于包括中文家谱在内的文化遗产资源，以关联数据技术为基础的跨语言信息服务有助于保障资源的获取，促进国际交流。此文介绍了关联数据中跨语言本体匹配的概念、代表项目和在文化遗产领域的相关实践，并以 2016 年上海图书馆开放数据应用开发竞赛作品 Learn Chinese Surnames 作为案例，展示了用关联数据的消费技术实现中文家谱数据与 DBpedia、GeoNames、Wiktionary 之间的跨语言匹配，以丰富中文家谱中的相关英文描述的实践，并总结了开发关联数据应用的经验。该研究有助于消除关联数据集之间的语言障碍，实现跨语言的家谱信息服务。

关键词 关联数据 跨语言本体匹配 中文家谱 开放数据应用 信息服务

分类号 G254

DOI 10.16603/j.issn1002－1027.2018.04.008

1 概述

关联数据指国际互联网协会 W3C 推荐的一系列在语义网上发布和联接 RDF 数据的最佳实践。其目的为在现有的互联网基础上，用标准的格式描述和发布数据及其之间的关系，保障数据的有效获取，以形成数据万维网(Web of Data)。随着关联数据规范在众多领域的推广，近年来，互联网上已发布了可观的语义资源，覆盖众多包括政府、学术、媒体、社交网络、图书馆等机构的资源以及大量的用户生成数据[1]。在 2007 年开放关联数据倡议的推动下，关联开放数据云(Linked Open Data Cloud，LOD Cloud，下文简称 LOD)中的数据飞速增长，从 2007 年的 12 种，到 2017 年 8 月的 1163 种①；根据 2014 年的爬取情况分析，关联开放数据云中共包括 90 万余个文档和 800 万余个整体关联的资源[1]。作为一种广泛采用的数据组织规范，关联数据与数字时代的图书馆有密切联系，到 2010 年，已有 20 个图书馆发布了关联数据集[2]。在 2014 和 2015 年，OCLC 研究部以问卷方式采集并分析了来自 20 个国家的 90 个以图书馆机构为主的 112 个关联数据项目，反映了近年来国际图书馆界采用关联数据技术的图景[3]。

关联数据的核心在于资源描述方式的统一和关联。虽然绝大部分用关联数据规范描述的数据集都用纯英文表述，形成了偏向于英文的数据群，但以非英文语种发布的关联数据集越来越多，比如法语的音乐数据集 Dogmazic②，西班牙语的地名库 GeoLinkedData.es③，还有中文的上海图书馆的家谱知识库④及名人手稿档案库项目⑤。这些非英文的数据集不应当因为语言上的壁垒，而成为数据万维网

* 感谢参赛队友西交利物浦大学的 Ilesanmi Olade 和钟坤权在应用界面开发和美工上的贡献。感谢上海图书馆刘炜和夏翠娟老师的支持与指导。感谢上海图书馆举办 2016 年开放数据应用开发竞赛。

通讯作者：董行，ORCID：0000-0001-6828-6891，邮箱：HangDong@liverpool.ac.uk。

① http://lod-cloud.net/。

② http://www.dogmazic.net/。

③ https://datahub.io/dataset/geolinkeddata。

④ http://gen.library.sh.cn:8080/ontology/view。

⑤ http://sg.library.sh.cn/ontology/view。

中的孤岛，相反它们往往具备独特的价值，能反映不同语言和地域独有的文化。应当尽量缩小这些数据在语言上的壁垒，建立多语言的数据万维网[4-5]。

关联数据使得图书馆的知识组织细粒度化，以开展精准的数据服务。上海图书馆使用基于 BIBFRAME 的本体，将传统的用 MARC 方式描述的家谱元数据和《中国家谱总目》中的数据按照关联数据的基本原则加以组织；新建立的家谱数据服务平台支持多种关联数据的消费接口，包括面向专业人士的 SPARQL 端点和以 JSON－LD 格式返回数据的开放数据应用程序接口（API）[6-7]。2016 年 4 月，上海图书馆首次举办了历时两个月的开放数据应用开发竞赛①。本研究介绍开放数据应用竞赛中的作品 Learn Chinese Surnames 的设计和实现方案。作品设计者通过使用主要的关联数据消费技术，将三种多语言关联数据集中的实体匹配并整合到家谱数据中，为中文家谱数据补充了英文的数据层和资源链接；并从应用的层级上验证了这一方案的可行性。为中文家谱关联数据补充西文数据，有利于消除图书馆的文化遗产资源服务的语言限制，实现无差别的公共数据服务，促进中文家谱的国际数字人文方面的研究。

本文第二部分介绍关联数据中的跨语言本体匹配的概念、代表项目和在文化遗产领域的相关研究。第三部分将上海图书馆开放数据应用比赛作品 Learn Chinese Surnames 作为一个案例，总结用关联数据的消费技术来匹配与整合跨语言多源关联数据的实现方式和基本要求。

2　关联数据中的跨语言本体匹配

关联数据和知识本体密不可分。本体作为领域内概念之间关系的模型，在 RDF 的三元组结构中扮演谓词的身份，并得以不断复用，是关联数据的骨架。在语义网领域，本体可粗略分为逻辑上的形式本体和词汇上的术语本体，因而本体匹配也涉及本体间的逻辑关系或术语概念间的匹配。跨语言本体匹配（Cross-Lingual Ontology Matching）是本体匹配的一个分支。关于单语言、多语言和跨语言本体匹配的概念，施波尔（Spohr），霍林柯（Hollink）和西米亚诺（Cimiano）[8]给出了被学术界较多采用[9]的定义：单语言本体匹配是对两个本体中用同一种语言描述的实体之间的匹配；多语言本体匹配是对两个本体中至少用两种相同语言描述的实体之间的匹配；跨语言本体匹配则是对两个本体中用不同语言描述的实体之间的匹配。这里的本体是包含概念和实体的。其中跨语言本体匹配的实现可以从一种语言翻译至另一种语言，或者将两种语言共同翻译为第三种语言。

跨语言本体匹配被视作建立多语言关联数据的核心任务，即将关联数据中的某一语种的本体和资源与其他语种的本体和资源相链接，从而为现有的语言本体匹配添加一层多语言的信息[5]。依照匹配的对象和方式，这一过程具体可以分为概念层级、实例层级和语言层级[5]的匹配，其中概念层级是对不同本体中的词汇的语义进行关联；实例层级是对不同概念下的实例进行关联，而非关联实例所对应的概念；语言层级不直接建立概念之间的链接，而是通过引入语义层作为新的资源层，来连接概念。

在关联数据规范和相关语义技术产生之前，跨语言本体匹配的实践中一个有代表性的案例是基于英文机读字典 WordNet② 的匹配和多语种扩展，包括 EuroWordNet、Meaning、GlobalWordNet、Kyoto 以及我国的 HowNet[10]。这些项目将最早的普林斯顿英文 WordNet 中结构化的词汇映射到各国语言中，同时也保留本国语言的结构特征（涉及本体本地化的问题）[10]。多语种词典的建设大多需要依照词典的结构进行新的规划，利用已有的词典资源，同时需要专家介入来选择和检查匹配结果，对人力和时间有较高要求[11]。目前关联数据云中囊括了 WordNet 等大量多语言的词典资源③。这些多语言关联数据是实现跨语言本体匹配的重要资源[5]。

较大规模的跨语言关联数据匹配往往采用语义相似度的计量以及机器学习的方法。Ngai 等人[12]通过计算两个单语种本体同义词群间的相似度，完成了跨语言本体英文 WordNet 和中文 HowNet 之间的匹配。王（Wang）姓学者及其团队[13]的研究通过马尔可夫随机场建立了分类器模型，根据知识型本体（即词表）所描述的资源来预测实体间的相似度，用于同质本体（描述图书的词表 GTT 和 Brinkman）之间和异质本体（描述图书的词表和描述多媒

① http://pcrc.library.sh.cn/zt/opendata/。
② http://wordnet.princeton.edu/。
③ http://linguistic-lod.org/llod-cloud。

体资源的词表 GTAA)之间的匹配。清华大学的 Xlore 项目[14-15]将百度百科、互动百科中的实体链接到英文维基百科中,建立了跨语言的知识图谱。通过建立和训练链接因子图模型,Xlore 在英文维基百科和百度百科之间新发现了 20 万余对跨语言匹配链接,解决了维基百科里中文资源缺乏的问题。

2.1 在文化遗产领域的跨语言实践

文化遗产领域的数字项目,由于资源的多样性和文化上的共通性,往往需要整合和链接各类资源,以满足用户丰富的信息需求。同时,文化遗产资源往往用本民族的语言描述,由于语言障碍,不易为其他国家或地区的用户获取和利用。对文化遗产资源,目前已经有不少在知识组织系统中跨语言匹配的实践,但基于严格的依照关联数据规范的跨语言项目仍较少。

对于传统知识组织系统,跨语言文化遗产领域的项目主要采用了两种方式。第一种方式是词表间的跨语言匹配和映射。在台湾数位典藏与数位学习国家型科技计划①(TELDAP)下的故宫器物数位典藏子计划中②,为满足习惯英语的用户对我国美术资源的跨语言检索需求,研究人员将中文的台湾故宫博物院词表与英文的艺术与建筑叙词表(Art & Architecture Thesaurus)相映射[16]。该项目采用纯人工的方式,在映射过程中对术语匹配和不同类型概念结构(即概念在两种受控词表中的层级和关联结构)的匹配进行了严格区分。第二种方式相对侧重于实体或元数据层面的翻译和匹配。威斯康星大学密尔沃基分校图书馆对周策纵教授生前捐献的中国书画卷轴和扇面资源③作了英文元数据的丰富[17]。首先,项目组在都柏林核心的基础上重新定义了适用于中英双语的元数据方案。其次,项目组使用人工翻译和词表匹配的方式,提供了基本元数据和描述性记录(主要文本、艺术家生平、题跋与款识)的英文表述。

在语义网的环境下,采用关联数据规范描述文化遗产领域的资源,并整合多语言的关联数据,能提高文化遗产资源的可见度和影响力。Europeana④数字图书馆项目包括了欧洲多语言的 5 千万种文化遗产资源。为了丰富多语言的数据并实现开放检索,该项目建立了关联数据模型和严格的三阶段元数据补充计划(分析—关联—增补),将单语言的元数据匹配到 GeoNames、GEMET、DBpedia 和 Semium Time 等多语言关联数据中[18-19]。Damova 及其团队[20]介绍了 MOLTO⑤ 项目中实现的文化遗产数据的跨语言匹配。该项目按照关联数据规范组织数据,并能满足 15 种语言的检索。项目组人工翻译了关于本体中材质(Material)和颜色(Colour)类下的实体,另外,106 个博物馆的名称则自动匹配到维基百科以获得多语种的标识。

我国的传统家谱作为一种重要的文化遗产资源,在历史和汉学研究上有独特价值。对家谱领域本体进行跨语种扩展,将有助于家谱资源的检索和发现,以及国际化的传播。但目前除美国犹他家谱学会提供的检索系统 FamilySearch.org 之外,尚未发现其他完备的跨语言的家谱发现系统。FamilySearch 是世界上最大的家谱项目,其中包括从 1239 年至今的中文家谱图片约 1300 万张⑥。项目采用众包的方式实现数据的跨语言,让全世界的用户来描述各国的多语言的家谱信息。但该项目尚未用关联数据的规范开放数据,也没有实现数据层面上的跨语言匹配。本文认为,用关联数据的规范来开放我国文化遗产数据,并加以跨语言的匹配和链接,能深化资源的跨语言服务和数字人文方面的研究。

3 案例实现:Learn Chinese Surnames

下面以上海图书馆开放数据应用开发竞赛作品 Learn Chinese Surnames 为例,介绍通过消费现有关联数据[21],实现家谱的跨语言服务的方法。

3.1 多种相关数据源调研

Learn Chinese Surnames⑦ 是一款方便非中文母语的人了解中国姓氏文化和学习汉字的安卓 App。该应用从 2013 年最新的现代四百大姓氏[22]出发,展示了用英文描述的关于中国姓氏的起源、名人、早期家谱(年代、地点、馆藏地),以及姓氏汉字

① http://teldap.tw/index.html。
② http://www.npm.gov.tw/digital/index2_2_8_ch.html。
③ http://collections.lib.uwm.edu/cdm/landingpage/collection/scroll。
④ http://www.europeana.eu/portal/en。
⑤ http://museum.ontotext.com/。
⑥ https://familysearch.org/search/collection/1787988。
⑦ App demo 可在比赛官方网站下载。http://pcrc.library.sh.cn/zt/opendata/apk/Learn%20Chinese%20Surnames%20.apk。

表 1　Learn Chinese Surnames 移动 App 中使用的以关联数据规范发布的数据集一览

数据源	使用到的信息	获取方式	App 中的使用方式	版权声明
上海图书馆家谱数据	家谱题名、纂修时间、地理位置、馆藏地英文名、姓氏英文名、姓氏人数	SPARQL 语句实时调用 Restful 服务获取	家谱信息的展示页面	(1) CC2.0 协议(署名－非商业性使用－相同方式共享) (2)比赛授权使用
DBpedia 和维基百科(Wikipedia)	姓氏的英文词条	SPARQL 语句离线获取	姓氏和汉字的展示页面	CC BY-SA 1.0－4.0
Wiktionary(维基词典)	汉字的英文词条	直接从 URL 获取	姓氏和汉字的展示页面	CC BY-SA 3.0
GeoNames	中国地理位置的英文元数据	官方 API 在线实时调用	家谱信息的展示页面	CC BY-SA 3.0

(含义、写法、读音)等信息。用户可以根据拼音字母和姓氏的使用人口排行浏览现代四百大姓氏,学习姓氏汉字的写法、含义、对应的姓氏起源,并了解早期家谱的保存情况。

图 1 列出了 App 中所展示的以姓氏或汉字为中心的各类信息。上海图书馆的家谱开放数据中提供了部分可直接在应用中使用的英文数据,包括姓氏的英文名、馆藏机构的英文名称和编纂年代。其他的数据(红色部分)可通过将家谱数据和关联数据集匹配而获得。汉字的动态书法图片主要通过调用 WrittenChinese.Com① 中的 GIF 图片链接来获取。

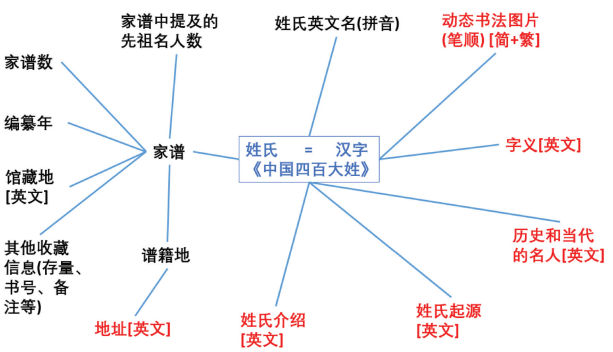

图 1　Learning Chinese Surnames 应用中所用到的信息一览

表 1 列出了本项目使用的在跨语言本体匹配中的数据源、信息、版权声明、获取方式和使用方式。合理地展示这些信息能满足一个外国用户了解中文姓氏和家谱的需求:首先,通过四百大姓氏排行和姓氏的英文名,用户可以浏览和查询姓氏;其次,通过姓氏汉字的笔顺动态图片和维基词典的解释,用户可以学习姓氏的写法和基本含义;再次,通过上海图书馆的家谱信息,用户可以从早期家谱来了解姓氏的起源;最后,维基百科的英文词条为用户提供了延伸的阅读材料,可借此了解姓氏的起源和名人等。

图 2 展示了在安卓系统界面中呈现的对中文家谱数据的跨语言匹配。图 2 左显示了关于中文家谱中"苏"姓所对应的维基词典和维基百科链接。图 2 右显示了中文家谱谱籍地对应的英文名称(从安徽省休宁县到"Xiuning Xian, Anhui China")。

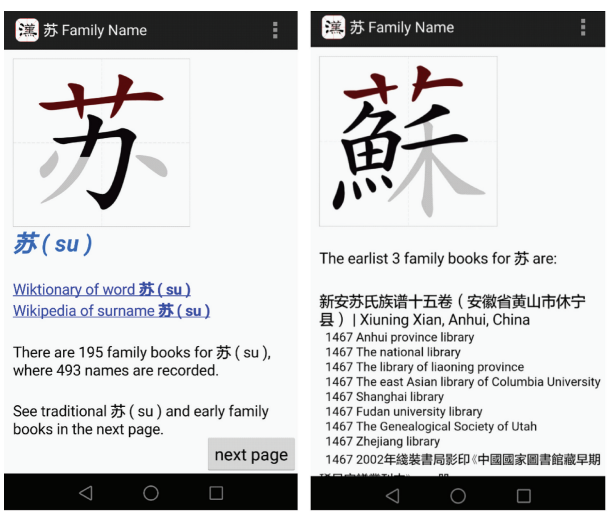

图 2　Learn Chinese Surnames 的安卓 App 界面

3.2　通过关联数据的消费技术实现数据的跨语言匹配

本项目主要为中文家谱的三种信息在关联数据云中匹配到对应的用英文描述的实体,分别为:(1)通过世界地理名库(GeoNames),获得中文家谱谱籍地对应的英文数据;(2)通过 DBpedia 和对应的维基百科(Wikipedia),获得中文姓氏的英文数据;(3)通过 Wiktionary(维基词典),获得姓氏所对应汉字的英文数据。这三种数据的获取正好对应三种主要的关联数据的接口和消费关联数据的方式[21],即 API 调用、SPARQL 端点调用和通过 URL 获取。图 3 反映了如何通过消费关联数据的方式来实现上海图

① https://www.writtenchinese.com/。

书馆家谱数据的跨语言匹配，图中“L”表示为Literal(值)类型，“@”为语言标记，“@chs”和“@en”分别为中文简体和英文，未标明为“L”的字段均为URI资源。

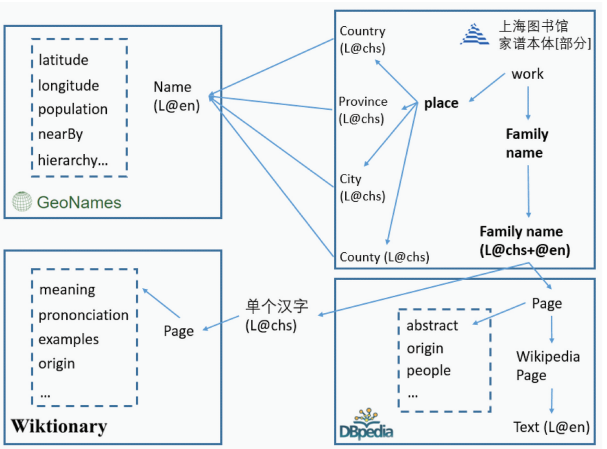

图3　通过消费关联数据来实现中文家谱数据的跨语言匹配

3.2.1　对家谱谱籍地数据的匹配—通过API接口消费GeoNames

GeoNames[①]地理数据库包含全世界1千万余条地理名称，以及9百万余条属性信息，如人口数量和地理名称替代名。GeoNames的数据包含多语言信息，且采用关联数据的原则进行描述，地理名称对应一个资源的URI，同时，用本体连接地名之间以及地名与属性间的关系。GeoNames无官方SPARQL端点，但提供了较为完备的API以及可供直接下载到本地的数据集。也可以使用第三方SPARQL端点FactForge[②]来获取GeoNames数据。

出于开发上的方便和访问速度的考虑，本项目采用了调用GeoNames官方API的方式。上海图书馆家谱数据中包括每个家谱的层级式地理名称(country、province、city、county)。从SPARQL端点获取上海图书馆家谱数据中的地理名称后，可以通过GeoNames中的featureCode(地理特征码)、name和country属性，构造GeoNames API的查询链接。比如获取陕西省西安市户县(1731年家谱《段氏世系》的谱籍地)对应的GeoNames的JSON条目的方式如下，http://api.geonames.org/searchJSON?name_equals=%E6%88%B7%E5%8E%BF&featureCode=ADM3&country=CN&maxRows=10&username=XXX。其中featureCode属性对应的ADM3表示行政上的县级(第三等级)地理位置，country对应要查找的国家代码，username为在GeoNames中注册的用户ID名称，name_equals对应通过完全匹配的方式查询的字符串，%E6%88%B7%E5%8E%BF即“户县”对应的UrlEncode编码。

返回的JSON数据如下，显示了该地理位置的英文名(name)、地理位置的经纬度信息(lng和lat)，以及该地区的人口数量(population)等信息。

{ " totalResultsCount ": 1," geonames ": [{ " adminCode1 ":" 26 "," lng ":" 108. 58764 "," geonameId ": 1806562," toponymName ":" Hu Xian "," countryId ":" 1814991 "," fcl ":" A "," population ": 556377," countryCode":"CN","name":"Hu Xian","fclName":"country, state, region, … ","countryName":"China","fcodeName":" third-order administrative division "," adminName1 ":" Shaanxi","lat":"33.99969","fcode":"ADM3"}]}

对于一个县级地理名称对应多个不同省的地名的情况，可以进一步通过省级地理名称(adminName1)来消歧。通过以上方式，在获得中文家谱谱籍地英文描述的过程中，达到了100%的成功匹配。

3.2.2　对家谱姓氏数据的匹配—通过SPARQL端点获取DBpedia数据

DBpedia项目抽取维基百科中的结构的和多语言的知识，并以关联数据的标准免费开放[③]。DBpedia共包含125种语言的资源、3800万余条事物，其中英文版的DBpedia知识库包含458万个事物。至2014年，DBpedia共发布30亿对RDF三元组的关联数据集，其中英文信息占24.6亿条，并根据维基百科的数据实时更新。DBpedia是整个关联数据的核心中枢，在数据云中与其他资源有5千万条RDF链接。

对于本案例，当前DBpedia数据集中包含了大量用英文描述的关于我国姓氏文化的信息。在RDF图里，foaf:isPrimaryTopicOf属性连接了一个DBpedia事物和一个单独的维基百科页面。DBpedia的资源URI非常规则，但由于DBpedia页面的名称和提供的信息的规范程度和完备程度不

① http://www.geonames.org/about.html，https://datahub.io/dataset/geonames-semantic-web。
② http://factforge.net/。
③ http://wiki.dbpedia.org/about。

高,因此无法直接采用 URI 方式来获取 DBpedia 或者维基百科的数据,需要使用 SPARQL 语句查询和人工核对等更加精准的方式。DBpedia 数据中质量不完善的问题主要有以下几点:(1)DBpedia 词条页面名称命名不规范,比如姓氏"刘"对应的页面名属性 dbp:name 的值是"Liu"①,"欧"和"区"对应的页面名是"Ou_(surname)"②;(2)DBpedia 中不同姓氏的页面中的本体属性不一致,无法全部结构化地提取维基百科中的最主要文本信息;(3)相比维基百科,DBpedia 提供的信息不全,一个维基百科的链接可以对应多个姓氏,如维基百科页面 Wu_(surname)③对应吴、武、伍、邬、巫、乌,但 DBpedia 的相应页面④中只显示了姓氏吴。

因此,项目采用了自动匹配和人工检查共同进行的方式。为了将姓氏自动匹配到维基百科词条,由于 DBpedia 的 SPARQL 端点的网络访问不畅,项目采用离线的方式调用 SPARQL 语句获取 DBpedia 的数据,并将四百大姓氏名、使用排名、维基百科链接存储在安卓 App 中。

使用 SPARQL 的前提在于对本体结构和含义有充足的了解。通过 DBpedia 本体中的 dct:subject 属性,可以把搜索范围缩小到某一主题的资源;通过 dbo:abstract 属性,可以获取某一资源对应的摘要,如果某一资源存在多语种词条,就获取多条不同语种的摘要;通过 foaf:isPrimaryTopicOf 属性,可以获取某一 DBpedia 资源对应的维基百科词条链接。以下的 SPARQL 语句通过匹配出现"赵"或"趙"的次数,并将链接资源降序排列,设定输出条数为 1,获取了赵姓在维基百科中的对应词条。

```
#DBpedia SPARQL 端点
#方法一:汉字匹配 DBpedia 资源摘要
select distinct ?url ?ex count(?res) as ?count
where{
?res dct:subject <http://dbpedia.org/resource/Category:Chinese-language_surnames>.
?res dbo:abstract ?a.
filter(contains(str(?a), "赵") || contains(str(?a), "趙")).
?res foaf:isPrimaryTopicOf ?url.
optional {?res dbo:wikiPageExternalLink ?ex}
}
order bydesc(?count)
limit 1
```

也可以通过直接将姓氏的英文名匹配资源 URI 的方式来获取姓氏在英文维基百科中的对应词条。在实践中按照字符串长度升序排列,并将输出结果数设为 1 条,这样有助于程序进行简单的批量处理。以下的 SPARQL 语句获取了曾姓(Zeng)在维基百科中的对应词条。

```
#DBpedia SPARQL 端点
#方法二:姓氏英文名匹配 DBpedia 资源 URI
select distinct ?u
where{
?m dct:subject <http://dbpedia.org/resource/Category:Chinese-language_surnames>.
filter(contains(str(?m),"Zeng")).
?m foaf:isPrimaryTopicOf ?u.
}
order by asc(fn:string-length(?u))
limit 1
```

以上两种方法都只能正确匹配一部分词条,由于是离线处理,而且仅有 400 条需要匹配的实体,项目综合采用了以上两种方法,并加以人工筛选和核对,完全确保了匹配的正确性。

3.2.3 对汉字相关信息的匹配—通过 URI 获取 Wiktionary 的数据

Wiktionary(维基词典)项目⑤用 wiki 的方式创建了一个免费的在线词典,于 2002 年上线,致力包含"所有语言的所有词汇"⑥。它当前包括 172 种语言。DBpedia 从维基词典数据中自动创建了 RDF,并借助众包的力量来修改本体词汇,提供了可供查询的 SPARQL 端点⑦,但本体中的词汇尚待成熟,而且当前仅满足不包括中文的 6 种语言的查询⑧。维基词典的内容还可以通过官方的 MediaWiki API⑨来获取。

由于 SPARQL 端点的访问限制、RDF 图中的信息不足,同时维基词典词条内部信息结构的统一程度不高,本项目没有采用 SPARQL 和 API 获取

① http://dbpedia.org/page/Liu。
② http://dbpedia.org/page/Ou_(surname)。
③ https://en.wikipedia.org/wiki/Wu_(surname)。
④ http://dbpedia.org/page/Wu_(surname)。
⑤ https://www.wiktionary.org/。
⑥ https://en.wikipedia.org/wiki/Wiktionary。
⑦ http://wiktionary.dbpedia.org/sparql。
⑧ http://wiki.dbpedia.org/wiktionary-rdf-extraction。
⑨ https://en.wiktionary.org/w/api.php。

信息的方式，而是直接调用维基词典页面的 URI。不同于维基百科，维基词典的词条名称非常有规律，可以直接构造出维基词典中查阅某一词条的 URI，比如汉字“刘”对应的维基词典链接为 https://en.wiktionary.org/wiki/%E5%88%98，其中%E5%88%98 是汉字“刘”对应的 UrlEncode 编码。通过构造 URI 的方式，可以实现单个汉字和维基词典页面完全的一一对应。此外，BabelNet[①] 也可以作为替代 Wiktionary 的语义资源。

3.3 对关联数据质量的验证

数据的验证是数据匹配中的重要步骤。由于数据集之间组织方式的不同和语义上的冲突，必须检查数据的匹配质量。本项目从数据内容的完备性和数据内容的冲突两个角度来验证数据。首先，以现代四百大姓氏为基础，可以发现上海图书馆的家谱数据对应上 377 个；英文维基百科有 295 个对应词条。以上海图书馆家谱数据中的姓氏为基础，英文维基百科中有 93 个姓氏无对应词条，这可以作为今后对不同数据源中数据的进行补充的依据。其次，通过将上海图书馆家谱数据和维基百科英文词条相关联，可以发现房、柏、区、强、危五种姓氏在英文名上的冲突，具体见表 2。

表 2 不同数据源间英文姓氏的表述冲突

姓氏	姓氏英文名(上海图书馆家谱)	英文词条名(维基百科)	维基百科连接	注释
房	Fang	pang	http://en.wikipedia.org/wiki/Pang_(surname)	fang\pang 均有
柏	Bai	bo	http://en.wikipedia.org/wiki/Bo_(Chinese_surname)	应为 bo
区	Qu	ou	http://en.wikipedia.org/wiki/Ou_(surname)	应为 ou
强	Qiang	jiang	http://en.wikipedia.org/wiki/Jiang_(surname)	应为 jiang
危	Wei	ngai	http://en.wikipedia.org/wiki/Ngai_(surname)	ngai 为广东话读音

综上可以看到，关联数据集的质量是对数据进行开发和利用的重要条件。结合本案例中的实践，关联数据集的质量主要体现在以下 4 个方面，前两个方面关乎数据的内容，后两个方面关乎数据的组织：(1)数据内容的完备性，即是否存在数据缺失；(2)数据内容的正确性，一定程度上反映为数据间的语义冲突；(3)本体结构的一致性，影响着对数据的批量查询，比如由于 DBpedia 之间各个页面本体属性的不统一，难以采用 SPARQL 对某一属性的值进行批量查询；(4)数据标识的一致性，也影响着对数据的批量查询，这里指 URI 名称中数据名的一致，由于 DBpedia 中各姓氏对应页面名称(即维基百科中的 URI 名)不一致，比如有的包含“surname”，有的则没有，难以准确地从姓氏构造出对应维基百科页面的 URI。此外，消费接口的多元化对于关联数据的消费极为重要。无论是官方接口，还是第三方接口，都有利于最大程度地发挥数据的价值。本案例中用到的四种关联数据资源，上海图书馆家谱数据、GeoNames、DBpedia 和 Wiktionary 都提供了多种获取数据的方式，为数据的开发和利用提供了多种可能。同时，也要根据数据的质量和使用方式来选择合适的消费接口，这一思路充分体现在本案例之中。

4 结语

关联数据的核心在于数据资源描述的统一和关联。随着单语言关联数据集的增多，语言的不同阻碍了这些资源的联系，因此关联数据的跨语言本体匹配的重要性突显出来。在总结了关联数据中跨语言本体匹配的相关概念和代表项目的基础上，本文以关联数据的移动应用 Learn Chinese Surnames 为案例，介绍了如何采用关联数据的消费技术，将中文家谱数据与三种多语言的关联数据集进行匹配，从而添加一层用英文描述的数据或链接。建立的跨语言应用完善了图书馆的家谱服务，有利于国际上与中文家谱相关的数字人文研究。

关于关联数据的跨语言匹配和案例项目的深化，未来的研究可以从以下几个方面进行。首先，对维基百科词条中文本的提取：提取英文维基百科中各姓氏的历史名人和人物关系，并关联到中文家谱本体中。其次，用中文家谱数据来补充在线百科的词条数据：将中文家谱数据嵌入到百度百科或维基百科之中。再次，利用 GeoNames 中的地理元数据，完善中文家谱数据的可视化。同时也可以进一步探

① https://datahub.io/dataset/babelnet。

索跨语言的家谱知识本体对家谱数字人文研究的影响。最后，对于其他任何多语言的数字图书馆项目，本体和数据的跨语言匹配仅仅是第一步。从跨语言的信息组织到完备的多语言数字图书馆的实现，仍有大量的工作亟待完成，比如多语言信息在关联数据中的表示和多语言信息的语义搜索。

参考文献

1 Schmachtenberg M, Bizer C, Paulheim H. Adoption of the linked data best practices in different topical domains. In: International Semantic Web Conference [C], Springer International Publishing, 2014: 245－260.

2 刘炜. 关联数据：概念，技术及应用展望[J]. 大学图书馆学报，2011 (2)：5－12.

3 Smith-Yoshimura K. Analysis of international linked data survey for implementers [J]. D-Lib Magazine. 2016, 22 (7/8): 1.

4 Trojahn C, Fu B, Zamazal O, Ritze D. State-of-the-art in multilingual and cross-lingual ontology matching. In: Towards the Multilingual Semantic Web [C]. Springer Berlin Heidelberg, 2014: 119－135.

5 Gracia J, Montiel-Ponsoda E, Cimiano P, Gómez-Pérez A, Buitelaar P, McCrae J. Challenges for the multilingual web of data [J]. Web Semantics: Science, Services and Agents on the World Wide Web, 2012, 11: 63－71.

6 夏翠娟，刘炜，陈涛，张磊. 家谱关联数据服务平台的开发实践[J]. 中国图书馆学报，2016，(3)：27－38.

7 夏翠娟，刘炜，张磊，朱雯晶. 基于书目框架（BIBFRAME）的家谱本体设计[J]. 图书馆论坛. 2014，34(11)：5－19.

8 Spohr D, Hollink L, Cimiano P. A machine learning approach to multilingual and cross-lingual ontology matching. The Semantic Web-ISWC 2011: 10th International Semantic Web Conference, Proceedings, Part I. [C]. Springer Berlin Heidelberg, 2011: 665－680.

9 Helou MA, Palmonari M, & Jarrar M. Effectiveness of automatic translations for cross-lingual ontology mapping [J]. J. Artif. Intell. Res.(JAIR), 2016, 55: 165－208.

10 Cimiano P, Montiel-Ponsoda E, Buitelaar P, Espinoza M, Gómez-Pérez A. A note on ontology localization [J]. Applied Ontology. 2010, 5(2): 127－137.

11 毕玉德，崔杞鲜，刘扬. 多语种词汇语义网建设中的几个问题. 见：全国第八届计算语言学联合学术会议论文集 [C]，南京. 2008：253－259.

12 Ngai G, Carpuat M, Fung P. Identifying concepts across languages: A first step towards a corpus-based approach to automatic ontology alignment. In: Proceedings of the 19th international conference on Computational linguistics-Volume 1 [C]. Association for Computational Linguistics, 2002: 1－7.

13 Wang S, Englebienne G, Schlobach S. Learning concept mappings from instance similarity. In: International semantic web conference [C] Springer Berlin Heidelberg: 2008: 339－355.

14 李娟子. 跨语言知识图谱构建 [EB/OL]. 第一届全国中文知识图谱研讨会. 苏州大学. 2013－10－12. [2017－11－16] http://bj.bcebos.com/cips-upload/kg/ljz.pdf.

15 Wang Z, Li J, Wang Z, Li S, Li M, Zhang D, Shi Y, Liu Y, Zhang P, Tang J. Xlore: A large-scale english-chinese bilingual knowledge graph. In: Proceedings of the 2013th International Conference on Posters & Demonstrations Track-Volume 1035 [C]. CEUR-WS. Org, 2013: 121－124.

16 Chen SJ, Chen HH. Mapping multilingual lexical semantics for knowledge organization systems [J]. The Electronic Library. 2012, 30(2): 278－294.

17 Matusiak KK, Meng L, Barczyk E, Shih CJ. Multilingual metadata for cultural heritage materials: The case of the Tse-Tsung Chow Collection of Chinese Scrolls and Fan Paintings [J]. The Electronic Library. 2015, 33(1): 136－151.

18 Isaac A. Case Study: Enriching and sharing cultural heritage data in Europeana [EB/OL]. 2012－06－13. [2017－11－16]. https://www.w3.org/2001/sw/sweo/public/UseCases/Europeana/

19 Stiller J, Petras V, Gäde M, Isaac A. Automatic enrichments with controlled vocabularies in Europeana: challenges and consequences. 5th International Conference, EuroMed 2014 [C]. Cham: Springer International Publishing; 2014: 238－247.

20 Damova M, Dannélls D, Enache R, Mateva M, Ranta A. Multilingual natural language interaction with semantic web knowledge bases and linked open data. In: Towards the Multilingual Semantic Web [C]. Springer Berlin Heidelberg, 2014: 211－226.

21 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.

22 袁义达，邱家儒. 中国四百大姓(上中下册)[M]. 南昌：江西人民出版社，2013：1－2.

作者单位：利物浦大学计算机科学系，英国；西交利物浦大学计算机科学和软件工程系，苏州，215123

收稿日期：2016 年 8 月 8 日

（转第 103 页）