DOI:10.13530/j.cnki.jlis.180010

中文百科概念术语服务平台 SinoPedia 的构建研究*

陈 涛 刘 炜 朱庆华

摘 要 随着“数据的网络”的兴起,万维网的内容已不再是纯粹的文本,而是表达和模拟多种事物及事件之间相互关系的实体集合,其中实体名称、属性及取值词表的规范十分重要。国外已形成覆盖广泛的“关联开放数据(LOD)”服务。中文概念术语的缺乏已严重阻碍中文知识图谱和中文领域本体的标准化和推广应用。本文提出的 SinoPedia 平台采用 RDF 三元组对目前公共领域的百科概念术语赋予唯一的 URI 进行资源的持久化,并通过 SOOOPA 模块提供检索服务。同时,自建的资源词条已与 DBPedia、WikiData、上海图书馆人名规范档等多个开放资源做了实体关联。除检索服务外,SinoPedia 还提供了关联数据发布服务,可以充当关联数据发布中心(Hub)。通过扩展 LODVIEW 系统为不同关联数据站点(SPARQL Endpoint)提供统一的关联数据发布和内容协商服务。此外,SinoPedia 集成了 LODLIVE 系统,能够实现不同数据集之间关联数据的发现与融合。目前 SinoPedia 包括了 554 万条三元组数据,并提供 API 接口和 SPARQL Endpoint 两种数据调用方式,下一步将申请接入 LOD 云图。SinoPedia 将来可以作为数字人文领域的数据链接中心,推动数字人文研究的快速发展。图 7。表 3。参考文献 20。

关键词 SinoPedia 关联数据 知识图谱 数字人文 知识发现

分类号 G250.2 TP393

SinoPedia: An Unified Chinese Terminology Service Platform Based on Linked Data

CHEN Tao, LIU Wei & ZHU Qinghua

ABSTRACT

With the development of “Web of data”, the content of the World Wide Web is no longer purely text but a collection of entities that can express and simulate events and their interrelationships. It is very important to specify entity names, attributes, and vocabularies on the World Wide Web. Europe and the United States have formed extensive Linked Open Data (LOD) services. However, the lack of Chinese conceptual terms has severely hindered the standardization and promotion of ontology in Chinese Knowledge Maps and Chinese domains. The SinoPedia platform proposed in this paper uses RDF triples to assign unique URIs with respect to the current public domain encyclopedia terminology and persist resources. It follows the

* 本文系国家社会科学基金重大项目“面向大数据的数字图书馆移动视觉搜索机制及应用研究”(编号:15ZDB126)的研究成果之一。(This article is an outcome of the major project “Exploring the Mobile Visual Search of Digital Library in Big Data Era: Mechanism and Application” (No.15ZDB126) supported by National Social Science Foundation of China.)

通信作者:陈涛,Email: tchen@libnet.sh.cn, ORCID: 0000-0002-6609-4914(Correspondence should be addressed to CHEN Tao, Email: tchen@libnet.sh.cn, ORCID: 0000-0002-6609-4914)

Linked Data of W3C that will publish the resources by four publishing principles. Moreover, the SinoPedia, acts as a publishing center of resources and can provide Linked data-related services to access external Linked Data sets (SPARQL Endpoint). The SinoPedia is composed of SOOOPA retrieval module, LODVIEW publish module and LODLIVE discovery module. It has been associated with DBPedia, WikiData and the Shanghai Librarian Name Authority File using the SOOOPA module to provide search services, and self-built resource entries. SinoPedia can store RDF data using OpenLink Virtuoso database. The search module of SOOOPA can retrieve words, multi-words, simplified and traditional Chinese characters and resource URIs, which can make intelligent ranking of search results. The retrieval results also give a link to other open resources, and the relevant information of the entries can be seen in other data sources in these results.

In addition to these search services, SinoPedia also provides Linked Data publishing services that can act as Linked Data distribution centers (Hubs). The SinoPedia provides a unified RDF data publication and content negotiation service for different Linked Data sites accessed by SPARQL Endpoints. Our platform extends the system of LODVIEW to support SPARQL Endpoint configurations with multiple external data sources. Resources from different sources are re-assigned in SinoPedia to obtain a uniform resource URI address, and, these resources can be redirected to the origin resource. The raw data of this resource are published using the new URI address of SinoPedia platform.

The SinoPedia integrates the LODLIVE system to realize the discovery and integration of Linked Data between different resources. The unified publication of different data sets achieves the unity of data syntax layer (RDF structuring). The links of different data sets achieve the unity of the data semantic layer, that is, the integration of multi-source data is realized through association. LODLIVE's Discovery Module displays the Linked Data from different sources in the form of knowledge graph. This Discovery Module also implements semantic extension and knowledge discovery services for resources through correlation.

At present, SinoPedia currently contains 5.54 million triplet data that includes people, places and institutions, and 730 000 instances. SinoPedia also provides API interface and SPARQL Endpoint calls. Finally, SinoPedia endpoint will also be registered in the Linked Open Data (LOD) cloud to make up the deficiency of knowledge base of Chinese encyclopedia in the LOD. In the future, SinoPedia can be used as a data link center in the digital humanities field to get more resource information by connecting to SinoPedia, and promote the development of digital humanities research. 7 figs. 3 tabs. 20 refs.

KEY WORDS

SinoPedia. Linked Data. Knowledge graph. Digital Humanities. Knowledge discovery.

0 引言

网络世界不缺中国用户,然而网络世界却极少有中文“资源”。这里的资源并不是指中文信息,而是指以互联网协会 W3C“资源描述框架(即 RDF)”形式表示和编码的数据和知识,建设一个“数据的网络”是提供知识服务乃至智慧服务的基础,并具有十分重要的意义。虽然我们有知乎、百度百科等知识社区,但它们的“知识”并不是计算机可理解的,也不是公共领域的,而以 Google 为代表的许多公司正在致力于在网络世界建立起以关联数据为基础的“知识图谱”,比较知名的如以维基百科的词条为基础的 DB-

Pedia①和 WikiData②等平台。这类通过众包建立的、免费的、通过标准接口提供开放服务的概念词表系统正在成为网络时代极其重要的知识服务基础设施,通常作为概念语义交换的知识中枢(Knowledge Hub)发挥作用,如在自动问答系统 Watson[1]中就用到了 DBPedia、WordNet、Yago 等数据。我国目前尽管也有类似的知识库存在,如国内中文百科知识图谱[1] zhishi.me③和复旦大学知识工场[2]的 CN-DBpedia④。它们虽然提供了资源的 RDF 格式,也可以通过数据下载和 API 方式进行调用,但都没有完全遵循 W3C 的关联数据发布四原则,不能作为知识流转的中枢。

1 研究现状

国外的关联数据集呈逐年增长趋势,从关联开放数据(LOD⑤)云图中可以看出,截至 2018 年,一千多个数据集(领域本体)中有近一万六千个跨领域的术语提供了相互链接,已有几百亿级的三元组数据可供使用,其中内容源自维基百科的 DBPedia 几乎在所有领域本体中都处于核心中枢的地位,除了生命科学、社交网络和政府信息三类。生命科学专业性强而且术语数量巨大,通用本体词表在其中用处不大;社交网络领域的本体大多不是知识性的术语;在政府信息类本体词表中 DBPedia 的术语虽然用得不少但种类不多,因此不处于中心位置。整个 LOD 云图中有近 1/3 的数据集链接到了 DBPedia,涉及的资源链接数高达 2 亿多,这些都是当前网络知识图谱的主要内容,可以看到 DBPedia百科术语词表起到了对网络知识图谱进行相互链接的中枢作用。

然而国内很少有可以媲美国外的本体词表服务。虽然可以认为语种只是概念的一个标签,但缺乏中文的标签等于是中文知识在网上的缺失,非常不利于中华文化的传播以及中文知识的交流。究其原因主要有发布技术和开放理念上的区别;就发布技术而言,本体词表只能用关联数据进行发布,但很多人并不清楚如何去发布关联数据,错误地认为将数据转为 RDF 结构后就是数据的发布。尤其是随着近年来政府开放数据的兴起,国内已有的政府机构在尝试以关联数据的方式来发布政府数据,但能够被正确发布且高效使用的数据集很少。在 2018 年的贵阳数博会中贵阳政府已以 RDF 格式来发布政府数据,虽然在发布技术和质量上有所欠缺,但已走出了改革创新的一大步。因此,我们认为构建一个关联数据统一发布平台能够在很大程度上解决这个问题,至少可以起到示范引领作用。

将数据转为 RDF 数据和关联数据的发布有本质的区别,关联数据发布的四原则[3-6]归纳为尽可能多地使资源能够在网络中以 RDF 数据形式流通。其中具体的形式就是要求发布的数据(资源实体)具有 URI 链接(原则一),并可以通过 HTTP 访问(原则二),同时尽可能提供资源的 RDF 数据(原则三和原则四)。RDF 数据的格式化是关联数据发布的前提,目前已经有越来越多的工具可以实现数据的 RDF 结构化,可以说 RDF 转换的门槛越来越低。OpenRefine⑥、上海图书馆转换工具 RDB2RDF⑦、usources.cn⑧、Lindas⑨等可以用来非常方便地将数据转为 RDF。

目前国内 RDF 数据的发布主要集中在使用 D2R⑩和借助 Drupal 的 RDF 模块。沈志宏[7]、游毅[8]、王忠义[9]、牛永骎[10]等使用了 D2R Server 进行关系型数据库的数据发布,使用 D2R Server 可以将关系型数据库中的字段通过 D2R Mapping 文件转为关联数据进行发布,而对于已经是 RDF 数据的数据集则不能进行发布操作。

① https://wiki.dbpedia.org
② https://www.wikidata.org
③ http://zhishi.me
④ http://kw.fudan.edu.cn/cndbpedia
⑤ http://lod-cloud.net
⑥ http://openrefine.org
⑦ http://data.library.sh.cn/tools/rdb2rdf
⑧ http://www.usources.cn
⑨ http://lindas-data.ch
⑩ http://d2rq.org

杨萌[11]、白林林[12]等人则借助 Drupal 的 RDF 模块进行数据的发布。使用 Drupal 可以进行数据的 RDF 结构化和提供 SPARQL 检索。SPARQL 检索可以将 RDF 数据查询出来并下载使用,但是这并不是好的机器可读方式。很多关联数据集都提供了 SPARQL Endpoint 访问方式[13],该方式以 RESTful 的接口方式进行数据的检索与获取,并不是直接发布和获取资源数据。上海图书馆的数字人文系列近几年逐渐推出多个关联数据集(家谱知识库[14]、古籍循证平台[15]、人名规范档、书目系统[16]等),这些数据集采用和扩展了 BibFrame 本体[17-18],并利用 OpenLink Virtuoso 提供相关的 SPARQL Endpoint。然而这些数据集的发布都依赖于各自的发布平台,查看和调用数据时需要去不同的平台进行访问。当需要将多个数据集集中发布时,一般的做法是将不同数据集的 RDF 数据导入到某一个统一的数据库中进行发布,这将带来数据更新的同步问题和维护成本的增加。

从这些发布方式可以看出,目前的很多关联数据集都是仅仅针对本地环境的发布,而不能作为发布中心(Hub)去发布多源(内部和外部)的 RDF 数据集。关联数据要求尽可能去关联更多的数据,而这些关联的数据如何去获取,以及如何和本地的资源进行信息的融合和可视化将是关联数据消费和传播的关键因素。SinoPedia 不仅可以作为中文开放百科知识库使用,还可充当发布中心提供外部数据集的相关关联数据服务,本文介绍了作者在构建 SinoPedia 时的实践和思考。

2 SinoPedia 知识库系统架构

图 1 为 SinoPedia 知识库的平台框架,其中平台框架具体的分类从下而上为数据接入层、功能服务层、资源链入层三层。

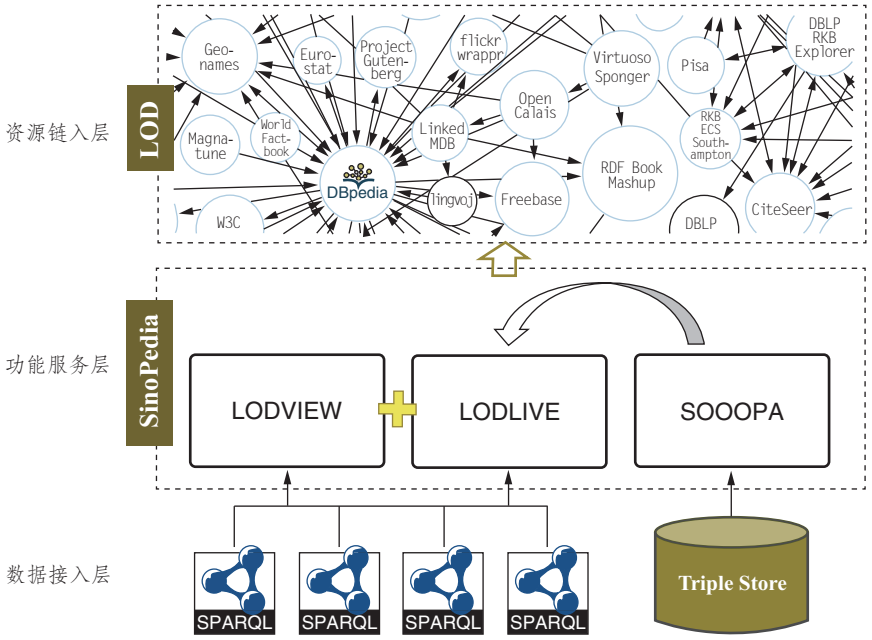

图 1 SinoPedia 服务平台框架

(1)数据接入层

该层指明了 SinoPedia 的主要数据源,其中数据来源主要分自建数据和接入数据两大类。①自建数据为平台收录的资源词条,这些词条

主要来自上海图书馆数字人文系统、DBPedia 和其他网络资源。其中上海图书馆数字人文系统中的数据和 DBPedia 数据为 RDF 数据,网络资源数据为 Excel、XML 等多种数据格式。这些异构数据目前采用线下转换的方式,将多种类型的数据转为 RDF 数据,并导入到三元组数据库。目前知识库所用三元组数据库为 OpenLink Virtuoso7.2.4.2。②接入数据为网络中其他在线的关联数据资源,这些资源通过 SPARQL Endpoint 提供访问,在 SinoPedia 中可以简单地通过相关配置文件进行资源接入。

(2)功能服务层

该部分给出了 SinoPedia 的核心功能模块,主要由 SOOOPA 检索模块、LODVIEW[①] 发布模块和 LODLIVE[②] 发现模块三大部分组成,其中各自模块的功能为:

①SOOOPA 检索模块直接连接 SinoPedia 的三元组数据库,来提供站点资源的检索服务和接入资源的转发服务。

②LODVIEW 和 LODLIVE 模块则需要连接不同的站点资源,这些站点资源为已发布的开放的关联数据集(SPARQL Endpoint)。LODVIEW 发布模块主要用来提供不同数据源中关联数据的发布和内容协商服务,连接的站点资源可以为开放的数据集,也可以是需要账户认证的数据集。

③LODLIVE 发现模块则以可视化的知识图谱方式提供不同数据源之间关联数据的浏览服务,要求接入的数据集为开放资源。

(3)资源链入层

SinoPedia 目前包含人、地、时、机构等资源类型,并做了和外部关联开放数据集的关联,如 DBPeida、GeoNames[③]、NobelPrize[④]、WikiData、上海图书馆人名规范档知识库[⑤]等。当外部资源链接到 SinoPedia 资源后,即可获取 SinoPedia 中该资源的 RDF 数据和其他相关联的外部数据。

3 SOOOPA 检索服务

SOOOPA 检索服务支持“单词”“多词 AND”“简繁体”“资源 URI”等多种检索方式,可以方便快速地找到所需资源的 RDF 数据。SOOOPA 检索采用了 Virtuoso 的 Free Text 检索机制[19],并对不同属性字段进行了差异性权重设置,检索的结果将根据相关度进行排序。如检索“李政道”,对应的 SPARQL 查询语句为:

```
SELECT DISTINCT ?uri ?name
WHERE {
  {
      ?uri dbpedia-owl:name ?name ; dbpedia-owl:name ?o .
      ?o bif:contains '李政道'
      OPTION (score ?ss) .
      bind (?ss * 10 as ?sc)
  } UNION {
      ?uri dbpedia-owl:name ?name; ?p ?o .
        FILTER (!contains(?name, '李政道'))
        ?o bif:contains '李政道'
        OPTION (score ?sc) .
  }
}
ORDER BY DESC (?sc)
OFFSET 0 LIMIT 10
```

① http://lodview.it
② http://en.lodlive.it
③ http://sws.geonames.org
④ http://data.nobelprize.org
⑤ http://names.library.sh.cn

该检索语句主要分为两部分,行 3—8 首先对标题进行检索,标题(属性为 dbpedia-owl:name)中命中的词条优先级将大于其他属性(如 rdfs:comment)命中的词条,这里设置了 10 倍的分值权重(?ss*10)。行 5 中使用了 bif:contains 对资源进行自由词检索(Free Text Search),该检索效率将大大优于简单使用 contains 或者 regex 进行的条件过滤效率,如:FILTER contains(?o, '李政道'或者 FILTER regex(?o, '李政道'))。第二部分检索为行 9—12,该部分用于检索除了标题外其他属性中命中的词条,这些词条的分值权重为 1。最后在行 15 中对检索出来的结构进行相关度倒排序。行 16 指定了返回的条数及偏移量(offset 用于分页)。检索结果见表 1,标题中含有"李政道"将优先显示。

表 1 检索结果得分排序

资源 URI	词条名称	得分
http://sinopedia.library.sh.cn/entity/person/cd9307f6b8804f05af73d52a27348423	"李政道"	40 000
http://sinopedia.library.sh.cn/entity/person/860173ab51064af5a0549498b282952c	"束星北"	1 440
http://sinopedia.library.sh.cn/entity/person/71f3fe27faa34cc48eb50fdf9a1d941c	"李中汉"	1 028
http://sinopedia.library.sh.cn/entity/person/a00fec69bdd4463082d47b84be32f2f3	"张朝阳"	960
http://sinopedia.library.sh.cn/entity/person/686c8fc2800849a68f842cf7c2723387	"杨振宁"	837
http://sinopedia.library.sh.cn/entity/person/e36338194be14098a94939fe343b05e5	"梅尔文·施瓦茨"	800
http://sinopedia.library.sh.cn/entity/person/e4e4dec552d94b6a977fa14188ca2250	"吴健雄"	571
http://sinopedia.library.sh.cn/entity/organization/983424fe86c740fca54361972f95f752	"芝加哥大学"	523
http://sinopedia.library.sh.cn/entity/person/b88672333c394340b80d9bf3de7a084f	"吴大猷"	458

整个排序的算法基于以下排序规则:

①单词的分数与文档中该单词的出现频率成正比。

②词组的分数既与文档中词组的出现频率相关,也与词组中单词的数据相关。

③"A NEAR B"的分数将取决于 A 和 B 之间相隔的距离,距离越远则得分越低。例如最高得分为 100,当 A 和 B 紧挨时,得分为 100;A 和 B 相隔 5 个单词时,得分为 80,以此类推。

④对于"A AND B AND C.... AND Z"的情况,将在 A、B、C...Z 最小得分基础上加上奖励分,越长的 AND,将会获得越高的奖励分。

⑤对于"A OR B OR C... OR Z"的情况,将在 A、B、C...Z 最高得分基础上加上奖励分,越长的 OR,将会获得越低的奖励分。

上述 SPARQL 对应的结果得分排序如表 1 所示,这里仅有 9 条相关记录,其中得分仅为相对分,而非绝对分。

每个词条后面都加上了内容协商服务,这些将依赖于 LODVIEW 发布模块,在词条的最后同样给出了已经关联的其他词条,如"李政道"关联了"上海图书馆"(人名规范档)、NobelPrize 和 DBPedia,如图 2 所示。当点击这些关联的词条,将会跳转到相关资源词条,当关联的词条所在站点资源已经接入 SinoPedia,相关资源词条将会在 SinoPedia 中打开该外部资源的 RDF 数据,如这里的 NobelPrize 资源;当关联的站点资源未接入 SinoPedia,点击相关词条时,将会跳转到资源所在网站访问,如 DBPedia 资源。

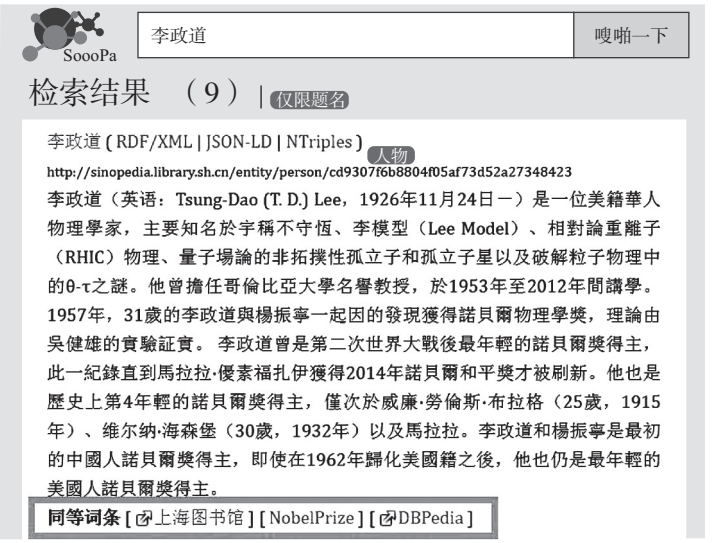

图 2 “李政道”关联词条

4 LODVIEW 发布服务

LODVIEW 系统是基于 Spring 和 Jena 的 Web 应用程序，可以与 SPARQL 端点一起以关联数据的发布标准来发布 RDF 数据。SinoPedia 在 LODVIEW 代码的基础上做了二次开发，以支持单个 LODVIEW 平台接入多个 SPARQL 端点，这样 SinoPedia 就可以充当外部资源的发布中心。不同站点的资源可以采用 SinoPedia 域名统一管理和发布，即在 SinoPedia 上显示不同站点资源，并针对这些接入站点资源提供关联数据的内容协商服务（主要有 RDF/XML、JSON、NT、Knowledge Graph）。

为了显示不同站点资源的单一显示，需要在 SinoPedia 中对这些站点资源进行重写。当外部 SPARQL 端点接入知识库后，知识库将会统一生成全局的资源 URI，知识库中采用的资源 URI 规则为：

http://sinopedia.library.sh.cn/${site}/${pattern}

其中变量 ${site} 为每个接入站点的标识，见表 2 所示。

表 2 不同接入站点的资源标识

数据源	SPARQL Endpoint	SinoPedia 资源标识
诺贝尔奖知识库	http://data.nobelprize.org/sparql	Nobel
第一次世界大战知识库	http://ldf.fi/ww1lod/sparql	ww1
Getty 艺术与建筑叙词表	http://vocab.getty.edu/sparql	Getty
地理信息知识库	http://geo.linkeddata.es/sparql	Geo

${pattern} 为资源在原始站点的命名规则，如诺贝尔奖知识库中“李政道”的资源地址为：http://data. nobelprize. org/resource/laureate/69，对应到 SinoPedia 中对应的地址为：http://sinopedia.library. sh. cn/nobel/resource/laureate/69，这里主要将原始域名替换为 SinoPedia 域名，并加入站点标识 nobel，这样实现了不同站点资源的统一管理。表 3 给出了目前接入站点的不同资源 URI 映射。

表 3 不同接入站点资源分配

资源名(来源)	资源原始 URI	SinoPedia 资源 URI
李政道(诺贝尔奖知识库)	http://data. nobelprize. org/resource/laureate/69	http://sinopedia. library. sh. cn/nobel/resource/laureate/69
1st Army(第一次世界大战知识库)	http://ldf.fi/ww1lod/635e1f7d	http://sinopedia. library. sh. cn/ww1/ww1lod/635e1f7d
Diamond Brook(Getty 艺术与建筑叙词表)	http://vocab.getty.edu/tgn/1135223	http://sinopedia. library. sh. cn/getty/tgn/1135223
Lago(地理信息知识库)	http://geo. linkeddata. es/resource/Lago/Sein%2C%20Lago%20del	http://sinopedia. library. sh. cn/geo/resource/Lago/Sein%2C%20Lago%20del

原始的 LODVIEW 在其配置文件中仅可配置一个资源地址,SinoPedia 中对此做了扩展,以支持多站点的接入方式,目前配置文件片段为:

```
<> a conf:Configuration;
    conf:endpoint <http://localhost:8890/sparql>;
    conf:bindEndpoint <http://lodview.it/conf#endpoint_nobel> ;
    conf:bindEndpoint <http://lodview.it/conf#endpoint_geo> .

<http://lodview.it/conf#endpoint_nobel> a conf:Endpoint ;
    conf:IRInamespace <http://data.nobelprize.org/>;
    conf:endpoint <http://data.nobelprize.org/sparql> ;
    conf:httpRedirectPrefix "/nobel/";
    conf:authUsername "";
    conf:authPassword "";
    rdfs:label '/nobel/'.

<http://lodview.it/conf#endpoint_geo> a conf:Endpoint ;
    conf:IRInamespace <http://geo.linkeddata.es/>;
    conf:endpoint <http://geo.linkeddata.es/sparql> ;
    conf:httpRedirectPrefix "/geo/";
    conf:authUsername "";
    conf:authPassword "";
    rdfs:label '/geo/'.
```

其中,在 conf:Configuration 类中定义了需要绑定的站点资源,这里给出了 nobel(endpoint_nobel)和 geo(endpoint_geo)这两个外部接入资源的配置;绑定的站点资源都是独立的 conf:Endpoint,互相结构关系见图 3。

conf:Configuration 类中保留了 LODVIEW 之前的系统配置信息,包括一些检索语句属性(defaultQueries、defaultRawDataQueries 等)以及相关资源属性(imageProperties、longitudeProperties、latitudeProperties、linkingProperties 等)。Configuration 类通过 bindEndpoint 属性和实际接入的站点进行关联,可以同时关联多个站点。conf:Endpoint 则为扩展的站点配置类,包括 label(站点标识)、authUsername(连接帐号)、authPassword(帐号密码)、IRInamespace(资源命名空间)、httpRedirectPrefix(资源跳转前缀)、endpoint(站点 url)等。

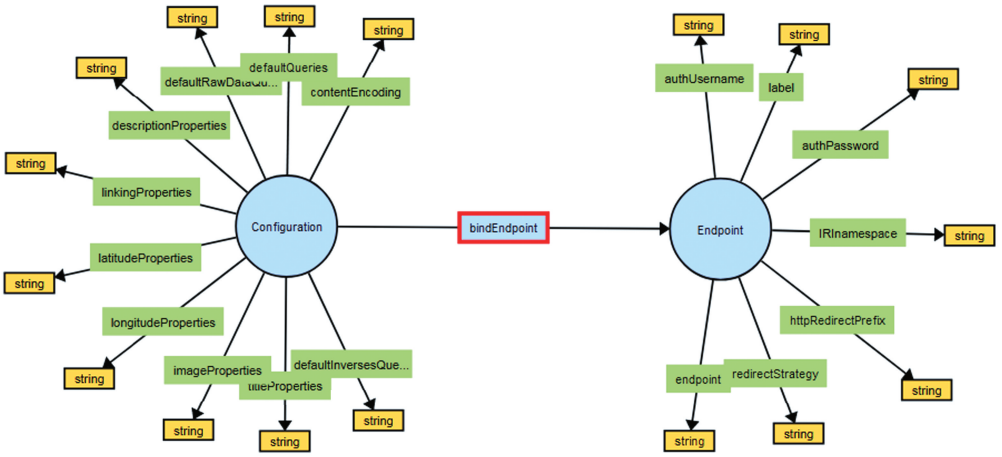

图 3 LODVIEW 配置文件扩展

以诺贝尔奖知识库为例,见配置文件中<http://lodview.it/conf#endpoint_nobel>部分。在 SinoPedia 中该站点标识(rdfs:label)为 nobel;站点 SPARQL 访问地址(conf:endpoint)为 http://data.nobelprize.org/sparql;资源命名空间(conf:IRInamespace)为 http://data.nobelprize.org/,表示当资源地址以该命名空间开头时,将通过 SinoPedia 平台进行数据的发布。如在 SinoPedia 中访问“http://data.nobelprize.org/resource/laureate/21”时,将会跳转到“http://sinopedia.library.sh.cn/nobel/resource/laureate/21”进行访问。

对于“李政道”词条对应的 HTML 页面如图 4 所示。

图 4 “李政道”资源页面

SinoPedia 针对接入站点提供了内容协商服务,当站点接入后,即可调用 RDF/XML、JSON、NT 等多种内容协商方式。“李政道”对应的 NTriples 数据获取地址为 http://sinopedia.library.

sh.cn/entity/person/cd9307f6b8804f05af73d52a2-7348423.ntriples,数据片段如下:

```
<http://sinopedia.library.sh.cn/entity/person/cd9307f6b8804f05af73d52a27348423>
<http://www.w3.org/2000/01/rdf-schema#comment> "美籍华裔物理学家.1926 年 11 月 25 日生于上海.抗战时期在国立浙江大学(当时在贵州省)和国立西南联合大学学习.1946 年赴美国芝加哥大学深造.1950 年获博士学位.1950—1951 年在加利福尼亚大学(伯克利分校)任教.1951—1953 年在普林斯顿高级研究院工作.1953—1960 年在哥伦比亚大学工作(1955 年任副教授,1956 年任教授).1960—1963 年任普林斯顿高级研究院理论物理学教授.1963 年起任哥伦比亚大学教授.是美国科学院院士.和杨振宁共同获得了 1957 年诺贝尔物理学奖."@zh .
<http://sinopedia.library.sh.cn/entity/person/cd9307f6b8804f05af73d52a27348423>
<http://xmlns.com/foaf/0.1/gender> "男"@zh .
<http://sinopedia.library.sh.cn/entity/person/cd9307f6b8804f05af73d52a27348423>
<http://dbpedia.org/ontology/name> "李政道".
<http://sinopedia.library.sh.cn/entity/person/cd9307f6b8804f05af73d52a27348423>
<http://www.w3.org/2002/07/owl#sameAs> <http://dbpedia.org/resource/Tsung-Dao_Lee>.
<http://sinopedia.library.sh.cn/entity/person/cd9307f6b8804f05af73d52a27348423>
<http://www.w3.org/2002/07/owl#sameAs> <http://data.nobelprize.org/resource/laureate/69> .
<http://sinopedia.library.sh.cn/entity/person/cd9307f6b8804f05af73d52a27348423>
<http://www.w3.org/2002/07/owl#sameAs> <http://data.library.sh.cn/entity/person/t5y9kqr3p5o8zl84>.
```

除了 NTriples 格式,LODVIEW 发布服务还支持多种其他格式的数据获取,如:

RDF/XML: http://sinopedia.library.sh.cn/entity/person/cd9307f6b8804f05af73d52a273484-23.rdf

JSONLD: http://sinopedia.library.sh.cn/entity/person/cd9307f6b8804f05af73d52a27348423.json

5 LODLIVE 发现模块

LODLIVE 插件旨在使用简单的图谱界面和使用关联数据标准来浏览相关的多源 RDF 资源[20]。以“李政道”为例,“李政道”在 SinoPedia 中已经关联到“上海图书馆(人名规范档)”“DBpedia”“nobelprize.org”等在线资源。通过 SinoPedia 的图谱展示功能可以查看不同资源数据的整合情况,利于信息的发现(见图 5)。

图 5 中 A 区为 SinoPedia 中的词条资源,该资源描述了李政道词条基本信息(姓名、性别、出生年月、介绍等);B 区为上海图书馆人名规范档资源,主要有李政道的国籍、籍贯、民族等信息;C 区为 NobelPrize 站点资源,在该资源中可以看到李政道在诺贝尔奖网站中的信息,主要有获得的奖项(Physics 1957,属性 http://data.nobelprize.org/terms/nobelPrize 连接)、所在的机构(Columbia University,属性 http://dbpedia.org/ontology/affiliation 连接)。同时,进一步看出 1957 年诺贝尔物理学奖由李政道(Physics 1957, Tsung-Dao Lee)和杨振宁(Physics 1957, Chen Ning Yang)两人共同获得。

使用 LODLIVE,通过 A 站点资源可以查看 B 站点和 C 站点资源,形成了不同数据源的信息整合,当不同的数据整合到一起后,甚至可以发现有意义的信息。如图 6 所示,同样在李政道词条中,我们在 NobelPrize 站点资源中打开李政道的出生地“中国”和“上海”,会发现出生在上海的诺贝尔奖得主共有三人(李政道、高锟、埃德蒙·费希尔);而出生在“中国”的非诺贝尔和平奖得主则有十人,除了熟悉的高行健、杨振宁、

莫言、屠呦呦、崔琦外，沃尔特·豪泽·布喇顿和根岸英一也出生于中国。因此通过不同数据集中关联数据的展示，可以发现更多的资源信息，以支持数字人文方面的研究。

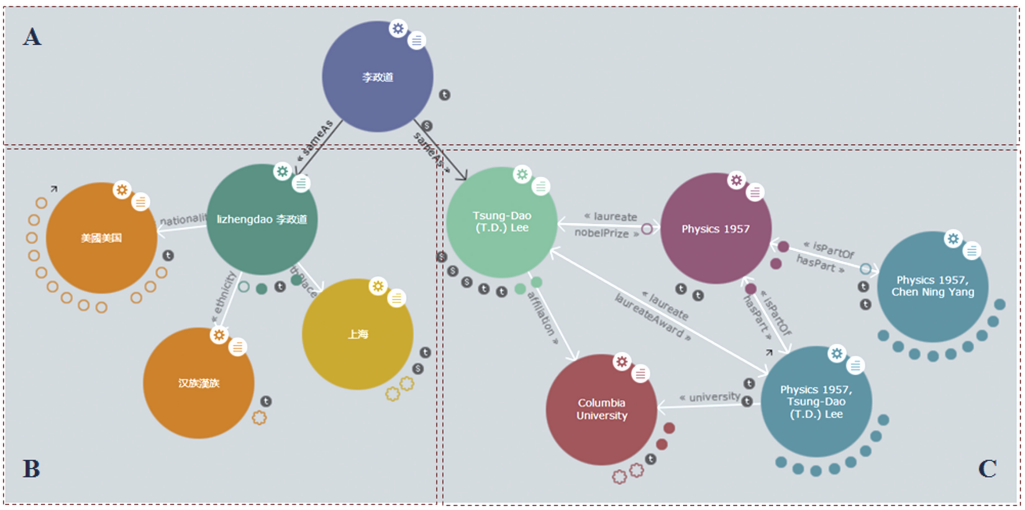

图 5 “李政道”多源数据关联图谱

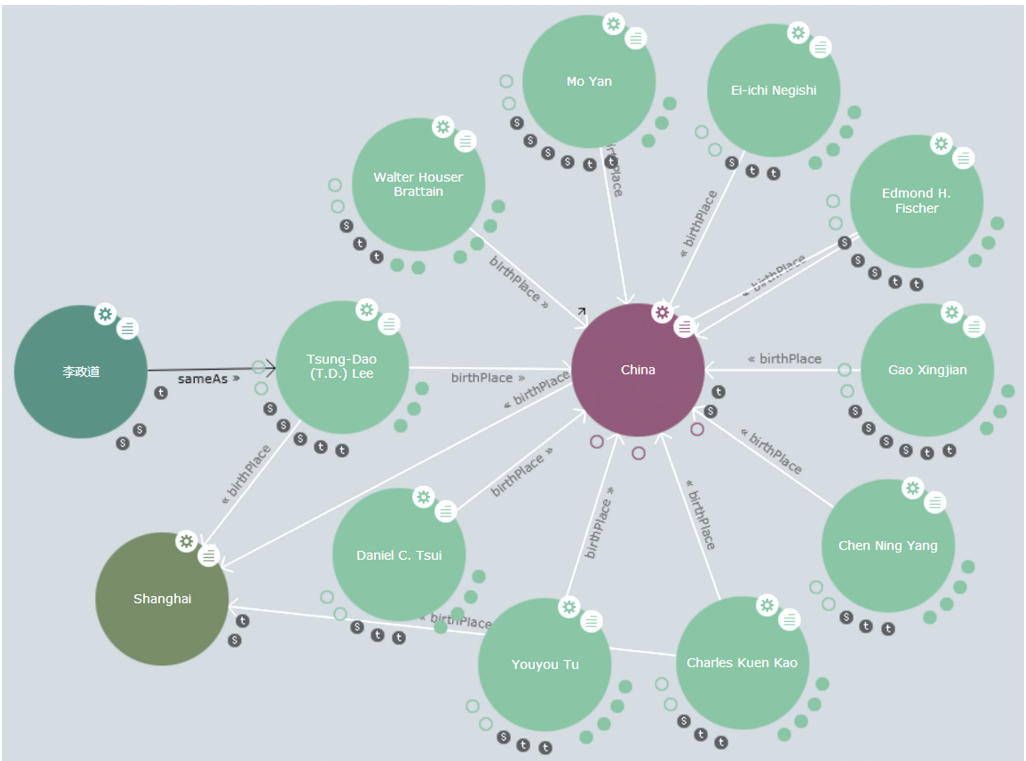

图 6 出生在“中国”和“上海”的诺贝尔奖得主数据关联图谱

6 SinoPedia 使用分析

SinoPedia 目前已收纳 73 万左右的实体,近554 万的 RDF 三元组,涉及人、机构、地理、朝代等多个类别,并与多个外部资源做了关联,如图 7 所示。

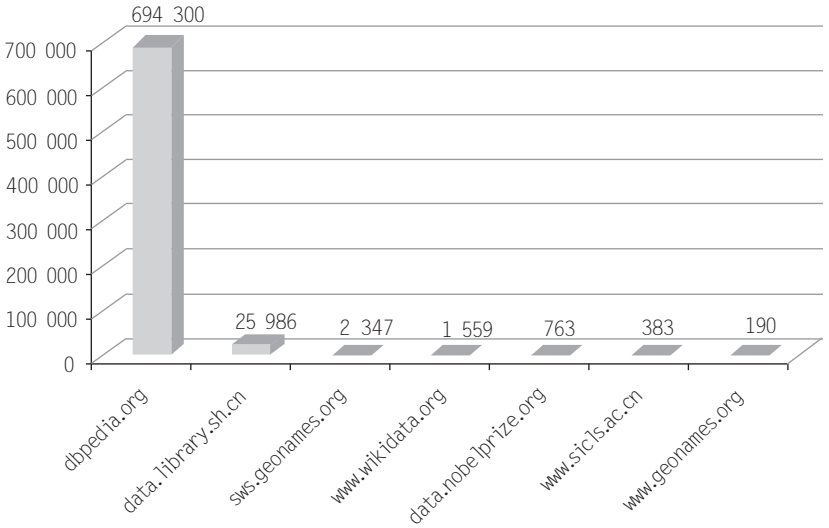

图 7 SinoPedia 关联资源统计

其中与 DBPedia 的链接占 95.7%,数目达 694 300条,主要因为平台中大部分词条来自 DBPedia。再者,从 LOD 云图中可以看出,DBPedia 已成为众多资源链接的核心。通过链接到 DBPedia 可以方便地与 WIKIDATA、YAGO①、FREEBASE②、VIAF③、DNB④等资源进行关联。

SinoPedia 有部分数据来自上海图书馆的数字人文系统,因此与这些资源的关联大概有 25 986 条,是总数据的 3.9%。SinoPedia 之所以没有像国内一些高校或企业那样采用网络爬虫技术抓取百度百科、知乎和类似网站的数据,并转换成 RDF 来提供一定范围的知识图谱服务,是因为这些爬取的数据不是公共领域的数据,即便支付一定的费用,也不可能享有完全开放使用的权利。国内即便是用户贡献内容的网站,内容大都属于网站公司,而不是创建者可以任意支配或声明的。这种现象已成为阻碍国内开放数据服务的主要瓶颈,SinoPedia 的研发也是为了打破这项制约,目前 SinoPedia 中的资源数据以 W3C 的标准完全开放。

SinoPedia 以内容协商方式提供多种机器可读格式,可作为数字人文关联数据搜索引擎使用。平台目前提供在线检索(http://sinopedia.library.sh.cn/sooopa)、SPARQL Endpoint(http://sinopedia.library.sh.cn/isparql/)、接口等多种调用方式。

其中,接口调用方式为:

http://sinopedia.library.sh.cn/data/search?${property}=${query}&p=${page}

①${property}为属性词,可以选择 name(仅检索标题名)和 q(全字段检索);

②${query}为检索词,可以为单个词或者

① http://yago-knowledge.org
② http://freebase.com
③ http://viaf.org
④ http://www.dnb.de

多词,多词之间用“+”号连接;

③ $ {page} 为页码。

例如:用 q 属性进行调用,接口为 http://sinopedia.library.sh.cn/data/search?q=鲁迅+学院&p=1,将返回题名或者描述中含有“鲁迅”和“学院”的词条,共有 227 条;当使用 name 属性进行调用时,接口为 http://sinopedia.library.sh.cn/data/search?name=鲁迅+学院 &p=1,仅返回题名中同时含有“鲁迅”和“学院”的词条,共计 3 条。

7 结论与展望

传统图书馆是传统社会的大脑中枢,通过提供各类基于馆藏的知识服务而起到整个社会的知识中枢的作用。数字图书馆要继续成为人类社会的知识中枢,仅仅向读者直接提供数字化的信息是远远不够的,不能满足当今社会日新月异的发展需要。要向网络上的各类“代理”提供数据和知识,并进一步将整个网络整合为一个统一的知识库,这就是 SinoPedia 这类概念术语服务的价值所在。

“数字学术”(Digital Scholar)领域对信息体的规范性和互操作性有较高的要求,尤其是近年来正在成为热点的“数字人文”。本文构建的 SinoPedia 平台大有所为,它不仅可以作为独立的知识库进行资源检索,还可以作为关联数据发布中心(Linked Data Hub)来发布多源的关联数据集,并提供相关资源的关联数据发布和内容协商服务。关联数据的发布可以看成是语法层的统一,而关联则可以实现不同资源的语义层的统一,SinoPedia 平台通过集成 LODLIVE 模块实现对关联的多源数据进行知识整合和知识图谱展示。目前,SinoPedia 已成功接入了多个关联数据源端点(SPARQL Endpoint),实现了在单一平台中展现不同数据源(本地和外部)的 RDF 数据,减少了不同数据集数据浏览时反复跳转的问题。

SinoPedia 平台已经上线,在提供服务的同时,将会在以下几个方面做进一步的研究与探索。

①SOOOPA 检索模块仅针对本平台的私有数据进行检索,并不包括接入站点的资源。接下来会对 SOOOPA 检索模块进行进一步的完善与升级,将会对接入站点的资源进行跨域的联邦检索。由于不同站点资源的检索存在效率上的缺陷,因此将考虑采用索引机制对检索资源的相关属性进行缓存索引,以提高检索的效率。当然,这将涉及资源信息的同步问题,这也是将来需要考虑的重点和难点。

②目前平台的自建数据来源和种类还很有限,为了支撑数字人文的研究,接下来会加入更多的资源种类和来源,如通过机器学习自动翻译技术,为更多的 DBPedia、WikiData、Freebase 中的概念术语添加中文标签,并引入更多的人物词条(科学家、知名学者等)、全球历史名胜古迹、历史事件、科学数据等。

③逐渐丰富关联关系。关联数据是链接不同数据源的最佳方案,接下来将会对平台中的历史人物和 CBDB①(中国历代人物传记库)以及 LOC②(美国国会图书馆规范数据)做进一步的资源关联。

此外,我们会尝试申请加入 LOD 云图,填补中文百科知识库在 LOD 家族中的空白,为中国数字人文的研究以及其他知识库的构建提供平台和科技创新的支持。

参考文献

[1] Fang Zhijia, Wang Haofen, Gracia J, et al. Zhishi.lemon: on publishing Zhishi.me as linguistic Linked Open Data[C]// ISWC 2016. Lecture Notes in Computer Science, vol 9982, 2016:47-55.

① https://projects.iq.harvard.edu/cbdb/home

② http://id.loc.gov

[2] Bo Xu, Yong Xu, Jiaqing Liang, et al. CN-DBpedia:a never-ending Chinese knowledge extraction system[C]// International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems , IEA/AIE 2017:428-438.

[3] Heath T, Bizer C. Linked Data: evolving the Web into a global data space[J]. Synthesis Lectures on the Semantic Web: Theory and Technology, 2011, 1(1): 1-136.

[4] Berners-Lee T. Linked Data[EB/OL]. (2009-06-18) [2018-06-20]. https://www.w3.org/DesignIssues/LinkedData.html.

[5] W3C Working Group. Best practices for publishing Linked Data[EB/OL].(2014-01-09) [2018-06-15]. https://www.w3.org/TR/ld-bp/.

[6] Marjit U, Sharma K, Biswas U. Discovering resume information using Linked Data[J]. International Journal of Web & Semantic Technology (IJWesT), 2012, 3(2):51-62.

[7] 沈志宏, 刘筱敏, 郭学兵, 等. 关联数据发布流程与关键问题研究[J]. 中国图书馆学报, 2013(3):53-62.(Shen Zhihong, Liu Xiaomin, Guo Xuebing, et al. A research on publishing workflow and key issues of Linked Data: experience with publishing scientific literature and scientific data as Linked Data[J]. Journal of Library Science in China, 2013(3):53-62.)

[8] 游毅. 面向馆藏数据库的关联数据发布研究[J]. 国家图书馆学刊, 2014(5):74-81.(You Yi. Research on collection database-oriented Linked Data publishing pattern[J]. Journal of the National Library of China, 2014(5):74-81.)

[9] 王忠义, 周杰, 黄京. 数字图书馆多粒度关联数据的创建与发布[J]. 情报学报, 2016, 35(8):885-896.(Wang Zhongyi, Zhou Jie, Huang Jing. The creating and publishing of multi-granularity Linked Data for the digital library resources[J]. Journal of the China Society for Scientific and Technical Information, 2016, 35(8):885-896.)

[10] 牛永骎, 常娥. 基于 D2R 发布学者关联数据集探究——以图书情报领域为例[J]. 图书情报工作, 2017, 61(19):13-21.(Niu Yongqin, Chang E. Research on publishing scholar repository Linked Data based on D2R[J]. Library and Information Service, 2017, 61(19):13-21.)

[11] 杨萌. 基于 Drupal 发布学者知识库关联数据的研究[J]. 图书馆研究, 2015(5):22-26.(Yang Meng. Research on publishing scholar repository Linked Data based on Drupal[J]. Library Research, 2015(5):22-26.)

[12] 白林林, 祝忠明. 基于 Drupal 的中文古籍书目关联数据发布研究[J]. 图书情报工作, 2017, 61(4):123-129.(Bai Linlin, Zhu Zhongming. Research on publishing Chinese ancient books bibliographic data to Linked Data based on Drupal[J]. Library and Information Service, 2017, 61(4):123-129.)

[13] 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.)

[14] 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(5):27-38.(Xia Cuijuan, Liu Wei, Chen Tao, et al. A genealogy data service platform implemented with Linked Data technology[J]. Journal of Library Science in China, 2016(5):27-38.)

[15] 夏翠娟,林海青,刘炜.面向循证实践的中文古籍数据模型研究与设计[J].中国图书馆学报,2017,43(6):16-34.(Xia Cuijuan, Lin Haiqing, Liu Wei. Designing a data model of Chinese ancient books for evidence-based practice[J]. Journal of Library Science in China, 2017, 43(6):16-34.)

[16] 夏翠娟,许磊. 中文关联书目数据发布方案研究[J]. 数字图书馆论坛, 2018,(1):8-16.(Xia Cuijuan, Xu Lei. Research and implementation of Chinese linked bibliographic data[J]. Digital Library Forum, 2018(1):8-16.)

[17] 娄秀明,危红. 书目格式的过去与未来[J]. 图书馆杂志, 2015(5):25-31,111.(Lou Xiuming, Wei Hong. The past and the future of bibliographic format: from MARC to BIBFRAME[J]. Library Journal, 2015(5):25-31,111.)

[18] 胡小菁. BIBFRAME 核心类演变分析[J]. 中国图书馆学报, 2016(5):20-26.(Hu Xiaojing. Evolution of BIBFRAME core classes[J]. Journal of Library Science in China, 2016(5):20-26.)

[19] Using full text search in SPARQL[EB/OL]. [2018-06-12]. http://docs.openlinksw.com/virtuoso/rdfsparql-rulefulltext.

[20] Camarda D V, Mazzini S, Antonuccio A. LodLive, exploring the Web of data[C]//Proceedings of the 8th International Conference on Semantic Systems, 2012:197-200.

陈 涛 上海图书馆/上海科学技术情报研究所,南京大学信息管理学院,博士后。上海 200031。

刘 炜 上海图书馆/上海科学技术情报研究所研究员。上海 200031。

朱庆华 南京大学信息管理学院教授,博士生导师。江苏 南京 210023。

(收稿日期:2018-06-26)