

关联数据发布的若干规范及建议^{*}

陈 涛 张永娟 刘 炜 朱庆华

摘 要 伯纳斯-李提出的开放数据五星模型,多年来一直被业内视为关联数据发布的最高标准。五星模型从追求高质量数据的角度出发,对数据发布者提出了相关要求,然而遵循五星模型发布的诸多数据集并没有为使用者带来便利,而更像一个需要人们去探索的黑盒。芬兰阿尔托大学的学者由此提出了开放数据七星模型,从数据集的易用性方面对五星模型做了扩展。本研究以此七星模型为基础,分析了从四星到七星模型中需要注意的诸多问题,并从本体设计、资源存储、数据发布、状态监控等角度提出了相应的规范措施和建议。这些规范和建议来自于多个数字人文项目中的实践经验,具有技术的可复制性,旨在从数据集发布伊始对发布中的多项工作进行梳理和指导,共同推动关联数据在各行各业的发展,打造良好的关联数据生态圈。图5。表3。参考文献17。

关键词 关联数据 七星模型 技术规范 语义网 开放数据

分类号 G254

Several Specifications and Recommendations for the Publication of Linked Data

CHEN Tao, ZHANG Yongjuan, LIU Wei & ZHU Qinghua

ABSTRACT

The Five-Star model of open data proposed by Berners-Lee, has been regarded by the industry as the highest standard for the publication of linked data these years. Publishers should apply the Five-Star model requirements to pursue high-quality data. However, many datasets published following the Five-Star model do not bring convenience to consumers. These datasets are more like black boxes that users have to explore. Scholars in the Aalto University in Finland have proposed a Seven-Star model of linked data, which is an extension of the Five-Star model in terms of usability of the dataset. When the dataset is published, the sixth star requires that its corresponding ontology should also be released, which makes it easy for the user to understand the structure of the dataset. And the seventh star requires declaring the use of classes and attributes in the ontology of the dataset, which makes it easy to confirm the data status of the dataset for user.

This article is based on the Seven-Star model. As we all know, the one-star stage specifies data online, such as image and PDF file; two-star is machine readable, such as EXCEL file; and three-star is a

^{*} 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)

relatively simple non-proprietary format, such as CSV file. Not all of these three stars are the core of the linked data system, thus they will not be discussed in this article. We analyze the problems that need to be addressed from the four-star to seven-star models and put forward corresponding regulatory measures and recommendations in this paper. In the four-star model, which needs to publish dataset with W3C open standards, we propose common technical specifications on resource URI design, content negotiation, storage and publishing the RDF data based on the four principles of linked data publication. In terms of data storage, we also propose that the three data storage methods are universal according to the structure of the dataset. There are “single graph(SG) mode”, “multi-graph(MG) mode” and “data hub(DH) mode”. The publication of the dataset is the basis in the linked data application. Moreover, the linking between the different datasets is the purpose. Therefore, for the five-star model about the linked data, this paper proposes some norms and suggestions that are often overlooked from the perspective of ontology vocabulary design and reuse. The expanded six-star and seven-star models are not separated from the five-star model, and there put higher requirements on the application of the linked data in details. In the six-star model specification, this paper proposes three forms of ontology publishing: “point mode” (file mode), “line mode” (page mode), and “face mode” (visualization mode) according to the degree of readability of the ontology. The “point mode” and “line mode” modes are six-star standards, while the “face mode” mode is up to six and a half stars. In addition, metadata description of the dataset is often overlooked, and this has an impact on whether the dataset is easy to understand and reuse. The corresponding solutions are proposed for these issues. The standard for the seven-star model is relatively simple; therefore, it can be realized by the use of existing tools or statistics on simple classes and properties.

In summary, the norms and recommendations are presented based on the author's practical experience in the field of digital humanities and linked data during these years. These technical solutions are reproducible. The purpose is to sort out and to guide the publication of linked data. Then these measures can promote the development of linked data in various industries, and create a good ecosystem of the linked data. 5 figs. 3 tabs. 17 refs.

KEY WORDS

Linked data. Seven-star model. Technical specification. Semantic web. Open data.

0 引言

关联数据并非新的数据,而是数据的一种新的展示方式,其价值在于通过资源描述框架RDF数据模型,将网络中的非结构化数据和采用不同标准的结构化数据转换成遵循统一标准的结构化数据,以便机器理解,这种模型更适宜采用图形式计算而不是关系型数据计算模式^[1]。作为语义网的轻量级实现方法,关联数据使互联网迈出了向语义网进化的重要一步,

可以以语义网的标准在网络中促进结构化数据的发现和互操作^[2]。

任何信息资源,不管其规模有多大,一旦成为封闭的孤岛,就会失去其数据价值。从关联的数据开放云图(Linked Open Data Cloud-LOD)中可以看出,从2007年的12个数据集开始到2018年的1 200多个数据集,越来越多的机构逐渐将其数据发布为关联数据。美国国会图书馆、英国大英图书馆、法国国家图书馆、德国国家图书馆、瑞典国家图书馆等逐渐将其书目数据发布为关联数据^[3]。2017年3月15日,施普

林格·自然集团(Springer Nature)推出了SciGraph(科研图谱)服务,再助科研人员一臂之力。这一新的关联数据平台整合了科研界的各种信息(如有关科研资助机构、科研项目及拨款、会议、科研单位和出版物等),同时会分阶段推出引用、专利、临床试验和使用数量等数据^[4]。目前,SciGraph已是LOD云图中最大的数据集之一(含有超过10亿的三元组),并链接到CrossRef和DBPedia等数据集。除了图书情报领域外,多国政府也逐渐试水关联数据,陆续将公共数据发布为关联数据。从全球参与开放数据运动的国家来看,既包括美国、英国、法国、奥地利、西班牙等发达国家,也包括印度、巴西、阿根廷、加纳等发展中国家^[5-6],中国政府数据也在尝试以开放标准进行发布。据2016年《联合国电子政务调查报告》显示,联合国已有128个成员国采用了机器可读方式开放数据集,128个国家中有92个国家已在用W3C的公开标准来开放数据。可见,关联数据已在众多领域快速发展^[7]。

2010年,伯纳斯-李提出开放数据开放程度评价五星模型,五星模型从数据使用的角度提出了数据开放的等级,从一星的数据上网,到二星的机器可读和三星的非专有格式,都只是在网络上(On the Web)实现数据的开放,数据集之间仍然存在数据隔阂,数据之间并没有消除信息孤岛。从四星的RDF到五星的关联数据,对数据发布者和数据消费者提出了更高的要求,这时数据才能成为网络中(In the Web)的一份子,才能将多源异构的数据集共构成巨大的网络实体数据库^[8],也只有三星及以上的数据才能称为高质量数据。然而,这些开放的数据集质量参差不齐,Hitzler^[9]、Hogan^[10]、Auer^[11]等人的研究指出,在发布的数据集当中,有一半左右都存在质量问题。同时这些开放的数据集使用门槛极高,比如一般关联数据发布时都将开放出相应的SPARQL端点(Endpoint),这些SPARQL端点很难理解,尤其是对于关联数据初学者和消费者而言。如何通过SPARQL端点快速准确了解数据集中的数据结构,如何能快速识别众多的

数据集,如何能构建高质量的关联数据集,这些都至关重要。芬兰阿尔托大学的学者则在五星模型的基础上扩展出了开放数据的第六星和第七星模型^[12]:①“六星模型”要求在发布关联数据集的同时,发布数据集中使用的本体,以增强数据集的可读性;②“七星模型”则需要申明本体在数据集中的使用情况,利于使用者判断数据集是否符合他们的使用要求,以增强数据集的可信性。基于七星模型,本文从实际应用出发,从技术规范的角度提出关联数据发布中的诸多行之有效的实践经验,以指导国内关联数据的发布。

1 开放数据四星技术规范

开放数据五星模型的第四星要求以W3C的开放标准(RDF和SPARQL)来识别事物的数据,并通过网络来共享数据。可以说从三星到四星的提升是数据质的转变。RDF的图谱模型使得从各种来源积累同一主题的信息变得简单,任何RDF系统都可以使用其他任何RDF系统的RDF数据,无需进行任何类型的协调,这使得大规模的数据集成变得更加容易。

这就要求以关联数据的四原则来发布数据:原则一,使用URI作为对象的标识名称;原则二,使用HTTP URI,使人们可以找到对象的标识名称;原则三,访问HTTP URI中的某个标识名称时,采用开放标准发现有意义的信息;原则四,提供相关的URI链接,以便查询者可以发现更多的对象。在发布关联数据集时,通常需要考虑如下技术规范。

规范1:设计资源URI时,该URI除作为对象标识名称外,还需考虑在数据发布后能够通过HTTP访问该资源。

采用HTTP形式进行资源的共享具有传统方式无法比拟的优势,这种访问方式远比将资源放在数据库或者文件中进行共享更为方便。正因为资源都采用HTTP形式流通,才有了互相融合的可能。很多系统在设置URI时,并未考虑将来系统的域名及数据的访问情况,仅仅将

资源 URI 作为系统内部 ID 使用。作为 ID 可以达到唯一性的效果,但在后期开放数据供机器可读和共享时,该 URI 将不能使用。因此,在设计资源 URI 时,需要事先做好将来的域名规划,以减少后期 URI 重构的代价。

对于事先未做好规划的资源,可在后期进行域名的跳转,如可以配置 Nginx。设计资源 URI 时,尽量采用有意义的唯一标识,如文献的 DOI、书籍的 ISBN 等,对于无明显标识的资源,往往使用随机号作为资源的 URI 标识。

规范 2:除了 SPARQL 端口形式访问数据外,需要提供资源的内容协商获取方式,以支持机器可读。

很多关联数据集提供了 SPARQL Endpoint 的调用方式,该方式只是通过 RESTful 形式提供数据集的检索服务,并不能直接进行资源的在线流通和信息获取,因此对资源而言,需要提供在线的元数据调用方式。该规范涉及关联数据的发布问题,对于关系型数据库常用 D2R 来发布;而对于三元组数据库中的数据,则可以通过 Pubby 或者 LODView 来进行发布。这里推荐 LODView,该系统已支持采用内容协商的方式来发布资源,支持的格式有 RDF/XML、NTriples、TTL、JSON-LD 等。

规范 3:使用三元组数据库存储 RDF 数据,并可与关系型数据库及企业级的索引库共存。

RDF 数据的存储主要有文件存储、关系型数据库存储、三元组数据库存储三种方式。对于归档数据,可以采用文件存储的方式,并提供下载链接;关系型数据库同样可以用来存储三元组,如 Jena SDB、Sesame 等;三元组数据库则使用最广,作为图数据库的一种,使用三元组存储方式将会使数据库的维护更为容易,如为数据库增加新的谓词(类似于关系型数据库中表的列)和书写更为复杂的查询式来进行推理和规则处理。关系型数据库在简单关联方面具有优势,但是三元组存储已经在复杂查询、规则处理和推理方面表现出更优的性能。

并不是说使用了三元组数据库就可以完全

抛弃关系型数据库,至少在将来的一段时间内两者会一起工作。这主要有两种原因:一方面,很多三元组数据库并没有成熟的权限控制,因此用户账号、权限等信息仍然可以用关系型数据库存储,资源数据可以用三元组数据库存储;另一方面,很多现有的系统都是用关系型数据库进行存储,有的数据量极其庞大,很难让这些耗费了大量人力物力的数据转为 RDF 数据。在这些情况下,三元组可以通过在数据库顶端添加智能化的元数据来促成灵巧的数据库间的集成。

三元组数据库中的数据在检索时一般会对整张图(Graph)进行检索,对于流水账式的实时产生的业务数据(如传感器信息、公交卡刷新信息),则需要使用 SPARQL_{stream} 进行检索^[13-14]。SPARQL 在使用中经常会考虑性能问题,尤其是涉及字符串的操作,效率极其低下,比如排序。因此很多时候会结合企业级的索引库,如 Redis、Solr 和 Elasticsearch。

规范 4:RDF 三元组数据,可选用适当的图(Graph)模式进行存储。

RDF 数据有多种存储模式(见表 1):①可以存放在一个图中,即“单图(Single-Graph)存储”,这种方式存储和调用都比较简单。②当数据量很大时,该 Graph 也会相当庞大,然而 Graph 可以很容易地进行分布式存放,可以根据数据量的大小切分成多个存储单元,即“多图(Multi-Graph)模式”。这些图之间并不像关系型数据中表与表之间的关联那么密切,可以轻易地进行拆分,如可以将文献、期刊、学者、基金等信息分别存放在不同的 Graph 中。这样的好处在于,当对不同类型的信息进行检索或交互时,不需要考虑其他类型的数据。但是,并不是将所有信息拆分得越细越好,因为当需要对不同类型的数据进行联合检索时,性能上会有所折损。③对于不同数据集的数据,在使用关联数据进行数据集集成时,没有必要将不同数据集的数据在物理层面进行整合,完全可以采用不同的数据节点进行存储,即“节点(Data-Hub)模式”。

表1 RDF数据存储的三种模式

存储模式	SG 单图模式	MG 多图模式	DH 节点模式
节点数	单 SPARQL Endpoint	单 SPARQL Endpoint	多 SPARQL Endpoint
图数量	单 Graph	多 Graph	单/多 Graph
数据集数量	单/多 Dataset	单/多 Dataset	单/多 Dataset
访问方式	远程调用	远程调用	远程调用
数据量分布	小数据量	大数据量 每个 Graph 可单独使用	分布式

传统方式进行数据集成如图 1(a) 所示,需要将不同的数据集汇聚到某一个存储空间中,并在此存储空间中进行数据的处理。而采用“节点模式”,不同端点之间的数据可以采用关联数据特有的“联邦查询(Federated Query)”功

能进行数据的集成与交互,见图 1(b)。采用联邦查询的好处是,每次获取的数据都是新鲜的,是从来源处获取的。如果来源数据更改,下一次运行联邦查询时,将会获得最新的数据。

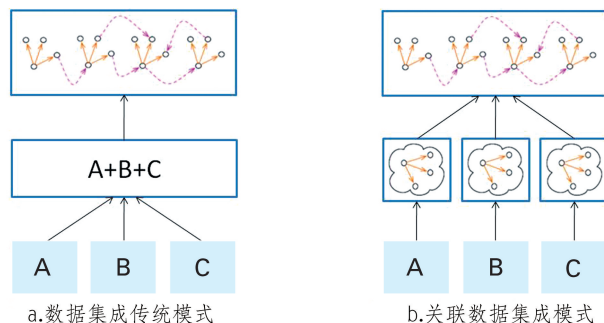


图1 多源数据集成模式

2 开放数据五星技术规范

开放数据四星规范的目的是实现多源异构数据的语法和用法统一,即使用 RDF 来描述资源,使用 SPARQL 来检索资源,使用内容协商来获取资源。这些做法的出发点还是针对单个数据集而言,相比四星而言,五星标准则要求通过各种技术手段和关联算法实现不同数据集之间的关联和融合。目前已有一些工具可以进行资源的发现与关联,如 SILK、LIMES 等,这方面可以和 NLP(自然语言处理)、机器学习的相关算法相结合,来提高实体识别和关

联的准确度。

不同数据集之间的资源在进行关联时,经常需要面对如下问题:①不同数据集在描述同一资源时,采用不同的资源描述属性或词表(RDF Vocabularies)。例如很多时候在描述作品的创作者时,会采用 dc:creator、dcterms:contributor 等,甚至有些采用自己定义的本体属性。②在生成资源 RDF 的时候,基本上都是发布者采用自己定义的资源 URI 规则。这就会导致:同一个资源在不同数据集中有不同的资源地址,而这些资源地址对应到真实世界中的同一个资源实体。例如“李政道”在不同数据集中有如下 URI:

SinoPedia 中 <http://sinopedia.library.sh.cn/>

entity/person/cd9307f6b8804f05af73d52a27348423
NobelPrize 中 [http://data.nobelprize.org/re-
source/laureate/69](http://data.nobelprize.org/resource/laureate/69)

DBPedia 中 [http://dbpedia.org/resource/Ts-
ung-Dao_Lee](http://dbpedia.org/resource/Tsung-Dao_Lee)

上海图书馆人名规范档中 [http://data.
library.sh.cn/entity/person/t5y9kqr3p5o8zl84](http://data.library.sh.cn/entity/person/t5y9kqr3p5o8zl84)

考虑到不同资源关联时的实际问题,在实际应用中经常考虑如下三个规范。

规范 5: 描述资源时,尽可能使用已有的公共本体,便于资源的发现与关联。

在使用本体进行资源描述时,尽量使用网络上已公开的公共本体,如: DC、FOAF、BIB-

FRAME、SKOS、GEO 等,具体的可以查看 LOV 云图^[15],LOV 上列出了 600 个常用的本体词表。然而很多时候,在设计本体时,并没有复用已有的本体,此时需要将不同本体进行对齐 (Alignment) 处理。使用 SPARQL 的 CONSTRUCT 可以实现不同属性之间的对齐和知识推理,通过对齐可以很容易地将远程资源的属性映射为本地使用的自定义的属性,如下方 SPARQL 语句通过 SERVICE 实现远端资源 (<http://example.com/sparql>) 中作者信息 (?author) 的联邦检索,并将检索后的作者信息通过属性重构转换为本地使用的 dct:creator 属性描述。

```
PREFIX ex: <http://example.com/ns/>  
PREFIX dct: <http://purl.org/dc/terms/>
```

```
CONSTRUCT {
```

```
  ?uri dct:creator?author .
```

```
} WHERE {
```

```
  SERVICE <http://example.com/sparql> {
```

```
    ?uri ex:author?author.
```

```
  }
```

```
}
```

规范 6: 描述资源的词表尽量使用英文驼峰形式,避免使用特殊字符,并严格区分类和属性。

很多词表为了使用的方便,在定义时使用中文字符,这并不利于词表的传播、发现与关联。定义词表时,可以将中文名称作为多语种进行维护,并区分大小写,如: bf:Title 为 Title 类, bf:title 则表示资源具有 bf:title 这个对象属性,属性值为 bf:Title 类。驼峰形式可用来表示多个单词之间的组合形式,如: sg:hasBookChapter 用来表示书籍和其章节的关系。

规范 7: 复用和扩展本体属性时,需区分对象属性 (Object Property) 和数据属性 (Datatype Property)。

很多机构在定义本体属性或者在复用其他本体时,不太关注属性是对象属性还是数据属性。数据属性要求连接的宾语是非 URI 形式,

如字符串、整型、时间类型等;而对象属性则用来连接不同资源,宾语可以是具体的资源 URI,或者是空节点。举例来看,当用来描述某个资源的语种时,可以采用 language 属性, language 有 dc:language 和 dbpedia:language。

通过这两个属性的定义,可以看出 dc:language 为数据属性,要求后面的宾语为字符形式;而 dbpedia:language 为对象属性,要求后面的宾语为语言资源的 URI。

同样,在下方本体片段中,定义了 bibframe 模型中的 bf:title (对象属性) 和 bf:mainTitle (数据属性),通过取值属性 rdfs:range 可以看出, bf:title 属性主要是用于宾语是 Title 类的资源 (<http://id.loc.gov/ontologies/bibframe/Title>); 而 bf:mainTitle 属性则要求宾语为字符串 (<http://www.w3.org/2000/01/rdf-schema#Literal>)。

```
<owl:ObjectProperty rdf:about="http://id.loc.gov/ontologies/bibframe/title">
  <rdfs:range rdf:resource="http://id.loc.gov/ontologies/bibframe/Title"/>
  <skos:definition>Name given to a resource.</skos:definition>
  <rdfs:comment>Used with Work, Instance or Item</rdfs:comment>
  <rdfs:label>Title resource</rdfs:label>
  <dcterms:modified>2016-04-21 (New)</dcterms:modified>
</owl:ObjectProperty>

<owl:DatatypeProperty rdf:about="http://id.loc.gov/ontologies/bibframe/mainTitle">
  <rdfs:domain rdf:resource="http://id.loc.gov/ontologies/bibframe/Title"/>
  <rdfs:range rdf:resource="http://www.w3.org/2000/01/rdf-schema#Literal"/>
  <skos:definition>Title being addressed. Possible title component.</skos:definition>
  <rdfs:label>Main title</rdfs:label>
  <dcterms:modified>2016-04-21 (New)</dcterms:modified>
</owl:DatatypeProperty>
```

3 开放数据六星技术规范

开放数据的五星模型主要从实例数据的角度提出了评级标准,但是在使用这些开放数据集时,遇到了以下问题:①难以理解数据集的特征,②难以重用数据集中的数据,③难以评估数据集的质量。大多数关联数据集提供 SPARQL 端点或者数据下载的方式分享数据,然而如何去理解和重用数据,已成为关联数据共享的主要问题。因此,芬兰阿尔托大学的相关学者提出了开放数据的七星模型,七星模型中并没有要求六星的可读程度,因此我们将六星分为两个半星。

六星:发布数据集同时发布数据集中使用的本体。

六星半:以高可读形式发布数据集中使用的本体。

规范 8:发布数据集的同时,尽量以高可读性的形式发布数据集对应的本体。

很多关联数据集只是将实例数据进行了发布与关联,并没有发布对应的本体。除非数据集全部复用已有的本体,否则建议发布数据集的同时发布本体。本体词表中定义了该数据集

的所有类和属性之间的关系,通过理解这些关系,可以快速了解数据集的数据结构,因此在发布数据集的同时,应发布对应的本体。对于已经发布对应的本体的数据集,在七星模型中,可达六星。

发布本体时,建议采用内容协商的方式进行发布,即采用 HTML 与 RDF 共存方式。HTML 方式更便于人类进行阅读,RDF 格式则利于机器进行调用和关联。有些现成的工具可以快速方便地进行 RDFS/OWL 文件到 HTML 页面的转换,如 LODE (Live OWL Documentation Environment)、OntoSpec。以上海图书馆家谱本体为例,进行 RDFS/OWL 文件和 LODE 方式的对比,下方本体片段为家谱本体中 sh1:Person 类及相关属性的定义,即使是经验非常丰富的设计人员也很难理解类及属性之间的关联关系,以这种方式发布也仅比没有发布稍好。这种发布方式可以称为“点”模式发布,即每次只能看到本体词表的某一点(类或属性),不同点之间的关系需要一个去查看。图 2 为 LODE 自动生成的家谱本体 HTML 页面,在页面中可以很清晰地看到类与属性的关系,这种方式相比“点”模式更为直观,称为“线”模式发布,即可以根据某个类,看到该类下的所有属性。

```
<owl:Class rdf:about="http://www.library.sh.cn/ontology/Person">
  <rdfs:subClassOf rdf:resource="http://www.library.sh.cn/ontology/Agent"/>
  <rdfs:subClassOf rdf:resource="http://xmlns.com/foaf/0.1/Person"/>
  <rdfs:subClassOf>
    <owl:Class rdf:about="http://bibframe.org/vocab/Person"/>
  </rdfs:subClassOf>
  <rdfs:comment>个人。</rdfs:comment>
  <rdfs:label>个人</rdfs:label>
</owl:Class>
<owl:ObjectProperty rdf:about="http://purl.org/vocab/relationship/spouseOf">
  <rdfs:comment>配偶。</rdfs:comment>
  <rdfs:range rdf:resource="http://www.library.sh.cn/ontology/Person"/>
  <rdfs:domain rdf:resource="http://www.library.sh.cn/ontology/Person"/>
  <rdfs:label>配偶</rdfs:label>
</owl:ObjectProperty>
<owl:ObjectProperty rdf:about="http://www.library.sh.cn/ontology/firstMigratedAncestor">
  <rdfs:comment>谱籍所在地最初迁居之祖为始迁祖。</rdfs:comment>
  <rdfs:range rdf:resource="http://www.library.sh.cn/ontology/Person"/>
  <rdfs:domain rdf:resource="http://www.library.sh.cn/ontology/Family"/>
  <rdfs:label>始迁祖</rdfs:label>
</owl:ObjectProperty>
```



图2 LOD 生成的家谱本体页面(sh:Person) - “线”模式

除此之外,我们可以用可视化的方式展示本体词表,如图3所示。这种方式可从更为宏观的角度对本体结构进行全局的理解,即“面”模式发布。从中不仅可以看出 sh:Person 类(图中

“个人”)中的数据属性和对象属性,甚至可以看出与对象属性相关联的其他类,如“始祖”“始迁祖”“房祖”“本族名人”“支祖”等对象属性的定义域(domain)为“家族”类,作用域(range)为

“个人”类。三种发布方式对比如表 2 所示,其中“点”模式和“线”模式可以达到六星标准;而

采用“面”模式可以达到六星半。

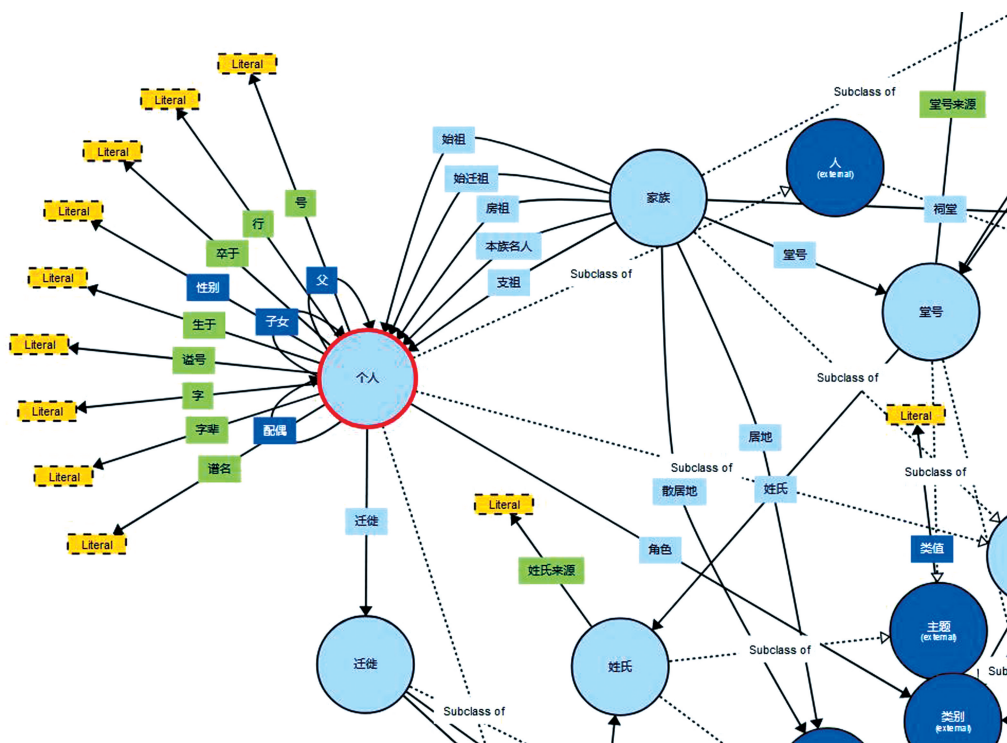


图 3 可视化展示家谱本体 (shl:Person) - “面”模式

表 2 本体词表发布三种模式

发布模式	“点”模式	“线”模式	“面”模式
发布方式	本体文件	HTML 网页	可视化图谱
发布特点	每次只能查看一个类或属性	可以查看类下所有属性	可以查看类下所有属性以及多个类之间的关系
发布难度	简单	中等	较难
星级评估	六星	六星	六星半

规范 9: 本体词表发布时, 需要加注本体 Metadata 元数据信息^[16]。

发布本体时,很多机构只关注词表本身,而忽略了本体自身的 Metadata 元数据描述。本体元数据中常用的属性/元素见表 3。这些元素可以分为以下几类:

识别属性: voaf: Vocabulary | vann: preferred-

NamespacePrefix | vann:preferredNamespaceUri

标题属性:dcterms:title|dcterms:description

版本属性: dterms:issued | dterms:modified |

owl:versionInfo

权限属性:dcterms:rights|cc:licence|dcterms:

```
creator|dcterms:contributor| dcterms:publisher
```

表3 本体元数据常用属性及描述

属性	描述	属性	描述
voaf:Vocabulary	指明类别	cc:licence	共用许可
vann:preferredNamespacePrefix	命名空间前缀	dcterms:creator	创建者
vann:preferredNamespaceUri	命名空间 URI	dcterms:contributor	贡献者
dcterms:title	本体标题	dcterms:publisher	发布方
dcterms:description	本体描述	rdfs:label	标签
dcterms:issued	最初创立时间	rdfs:comment	备注
dcterms:modified	版本修改时间	rdfs:isDefinedBy	定义方
owl:versionInfo	版本信息	vs:term_status	稳定、测试、不稳定、摒弃
dcterms:rights	版权所有		

词表元素: rdfs:label | rdfs:comment | rdfs:is-
DefinedBy | vs:term_status

下面给出 dbpedia 2014 版本体的元数据描

述,从中可以看到 dbpedia 本体始创于 2008 年,
当前的版本为 2014-07-15,许可协议为署名-
相同方式共享 3.0(CC BY-SA 3.0)。

```
<owl:Ontology rdf:about="http://dbpedia.org/ontology/">
  <rdf:type rdf:resource="http://purl.org/vocommons/voaf#Vocabulary"/>
  <vann:preferredNamespacePrefix>dbpont</vann:preferredNamespacePrefix>
  <vann:preferredNamespaceUri>http://dbpedia.org/ontology/</vann:preferredNamespaceUri>
  <dcterms:title xml:lang="en">The DBpedia Ontology</dcterms:title>
  <dcterms:description xml:lang="en">The DBpedia ontology provides the classes and properties used in the DB-
pedia data set.</dcterms:description>
  <foaf:homepage rdf:resource="http://wiki.dbpedia.org/Ontology"/>
  <dcterms:source rdf:resource="http://mappings.dbpedia.org"/>
  <dcterms:publisher>DBpedia Maintainers</dcterms:publisher>
  <dcterms:creator>DBpedia Maintainers and Contributors</dcterms:creator>
  <dcterms:issued>2008-11-17T12:00Z</dcterms:issued>
  <dcterms:modified>2014-07-15T10:00Z</dcterms:modified>
  <owl:versionInfo xml:lang="en">2014</owl:versionInfo>
  <rdfs:comment xml:lang="en">This ontology is generated from the manually created specifications in the
DBpedia Mappings Wiki. Each release of this ontology corresponds to a new release of the DBpedia data set which con-
tains instance data extracted from the different language versions of Wikipedia. For information regarding changes in this
ontology, please refer to the DBpedia Mappings Wiki.</rdfs:comment>
  <cc:license rdf:resource="http://www.gnu.org/copyleft/fdl.html"/>
  <cc:license rdf:resource="http://creativecommons.org/licenses/by-sa/3.0/">
</owl:Ontology>
```

4 开放数据七星技术规范

有了六星模型发布的本体结构,便于使用者对数据集的数据结构有所了解,但是这仍然不够。使用者在查看本体结构时,并不知道这些本体在数据集中的使用情况,即不清楚数据集的内容特征。很多时候,数据集在发布的时候,并没有完全参照之前定义的本体,这将给使用者使用这些数据带来困惑。当数据集提供 SPARQL 端点时,可以通过 SPARQL 语句去查询数据集中各类数据的使用情况,然而并非所有使用者都熟悉 SPARQL 语法,这加大了数据的使用难度;甚至很多时候,数据集并没有提供相应的 SPARQL 端点,只是提供数据的批量下载,此时需要通过自行搭建 SPARQL 端点来对数据集中数据的使用情况进行描述。

规范 10:提供数据集的数据状态,可按照本

体类和属性统计相关的实例数据。

这一规范在关联数据集的发布过程中往往容易被遗漏。数据集发布时,可以在对数据集进行描述的地方加上数据的使用状态,就可以很容易地提高数据集的星级。Miika Alonen 等提出了 Vocab.at 的解决框架和方案^[17],用来自动获取数据集中数据的使用状态,不过该网站目前已停止服务。在国内,usources.cn 平台中也有对 RDF 数据集使用状态的描述,如图 4 所示,当然这里只有针对类的统计,并没有统计相关属性的使用状态。此外,也可以借助 LD-OWL 在线工具,来查看数据集中数据的使用情况,如图 5 为目前 SinoPedia 知识库(<http://sinopedia.library.sh.cn/sparql>)的数据状态,从中可以看出该知识库中有人(Person)、地(Place)、机构(Organization)、高校(School)、医院(Hospital)等类,其中 Person 类的数量最多,共有 684 175 条。

上海市公安局开放数据[预览图谱] ☆☆☆ ✓ 已对接, 2016-09-08 22:40:46 最新更新

加入收藏

上海市公安局是上海市人民政府下设主管全市公安工作的职能部门,公安局受市政府、公安部的双重领导。上海警方以邓小平理论和“三个代表”重要思想和科学发展观为指导,紧紧围绕构建上海现代警务机制这一主线,牢牢把握全国公安机关开展抓基层、打基础、苦练基本功建设的契机,充分依托上海平安建设载体,着力提高维护国家安全的能力、驾驭社会治安局势的能力、处置突发事件的能力、为经济社会发展服务的能力,提高队伍执法水平和正规化建设水平,有效维护了上海社会政治和社会治安的稳定。

社保卡受理网点

(实例:267个/三元组:1602条,好评度:37.5%)

出入境受理点

(实例:21个/三元组:146条,好评度:无)

收起↑

看守所

(实例:23个/三元组:138条,好评度:无)

派出所

(实例:360个/三元组:1800条,好评度:无)

网吧

(实例:1656个/三元组:7110条,好评度:无)

居住证受理网站

(实例:269个/三元组:1345条,好评度:无)

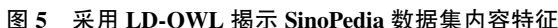
图 4 usources.cn 平台中的数据状态描述

5 结论及展望

本文在开放数据的五星标准基础上,结合芬兰阿尔托大学等专家扩展出的六星和七星模型,提出了关联数据开放过程中的一些常用规范和建议。这些规范和建议来自于诸多关联数据系统及平台实施过程中的思考,可以对关联

数据的发布及应用提供参考,起到事半功倍的效果。

六星标准从数据集的可读性上做了阐述,而七星标准则从数据集的可用性角度提出了建议,扩展的这两星并没有从本质上去改变人们熟知的五星标准,只是从关联数据传播和消费的角度去规范数据集。虽然只是改进了一小步,但却能在一定程度上提升关联数据的传播能



虽然有了开放数据的七星模型,但并不能囊括关联数据的方方面面,也不能解决关联数据中的所有问题,比如数据的准确性、数据集的可靠性以及安全性等问题,这些仍然需要更多的同行和学者共同研究。“开放、共享”会是永恒的话题,将助推关联数据应用的蓬勃发展,而我们作为关联数据的消费者,也应成为关联数据的提供者。

[1] 刘炜, 夏翠娟, 张春景. 大数据与关联数据: 正在到来的数据技术革命[J]. 现代图书情报技术, 2013(4): 2-9. (Liu Wei, Xia Cuijuan, Zhang Chunjing. Big data and linked data: the emerging data technology for the future of librarianship[J]. New Technology of Library and Information Service, 2013(4): 2-9.)

[2] Heath T, Bizer C. Linked data: evolving the web into a global data space (1st ed)[M]. Morgan & Claypool,

- Synthesis Lectures on the Semantic Web: Theory and Technology, 2011:1-136.
- [3] 邹美辰, 胡瀛. 欧美国家图书馆书目数据关联化案例研究[J]. 图书馆理论与实践, 2016(11):61-66, 70. (Zou Meichen, Hu Ying. Research on bibliographic data association of national libraries in Europe and America [J]. Library Theory and Practice, 2016(11):61-66, 70.)
- [4] 白林林, 祝忠明. Springer Nature SciGraph 关联开放数据分析[J]. 知识管理论坛, 2018, 3(1):2-11. (Bai Linlin, Zhu Zhongming. Analysis of Springer Nature SciGraph linked open data[J]. Knowledge Management Forum, 2018, 3(1):2-11.)
- [5] 吴钢, 曾丽莹. 国内外政府开放数据平台建设比较研究[J]. 情报资料工作, 2016(6):75-79. (Wu Gang, Zeng Liying. A comparative study on the construction of open data platform between Chinese and foreign governments[J]. Information and Documentation Services, 2016(6):75-79.)
- [6] 张涵, 王忠. 国外政府开放数据的比较研究[J]. 情报杂志, 2015, 34(8):142-146. (Zhang Han, Wang Zhong. Comparison research on open data of international government[J]. Journal of Intelligence, 2015, 34(8):142-146.)
- [7] 联合国经济和社会事务部. 2016联合国电子政务调查报告[R]. [2018-07-06]. (United Nations Departments of Economic and Social Affairs. United Nations e-government survey 2016[R]. [2018-07-06].)
- [8] 5 star open data[EB/OL]. [2018-06-28]. <https://5stardata.info/en/>.
- [9] Hitzler P. What's wrong with linked data?[EB/OL]. [2018-06-12]. <https://semantic-web.com/2012/08/09/whats-wrong-with-linked-data/>.
- [10] Hogan A, Umbrich J, Harth A, et al. An empirical survey of linked data conformance[J]. Web Semantics: Science, Services and Agents on the World Wide Web, 2014, 14:14-44.
- [11] Auer S, Demter J, Martin M, et al. LODStats-an extensible framework for high-performance dataset analytics [C]//Proceedings of the 18th International Conference on Knowledge Engineering and Knowledge Management, Galway City, Ireland, October, 2012:353-362.
- [12] Hyvonen E, Tuominen J, Alonen M, et al. Linked data Finland: a 7-star model and platform for publishing and re-using linked datasets [C]//Proceedings of ESWC 2014 Demo and Poster Papers, Springer-Verlag, May, 2014:226-230.
- [13] Danh L P, Nguyen M Q, Chan L V, et al. Elastic and scalable processing of linked stream data in the cloud [C]//ISWC 2013, Part I, LNCS 8218, 2013:280-297.
- [14] Bolles A, Grawunder M, Jacobi J. Streaming SPARQL-extending sparql to process data streams [C]//ESWC 2008, LNCS 5021, 2008:448-462.
- [15] Vandenbussche P Y, Atemezeng G A, Marfàc P V, et al. Linked Open Vocabularies (LOV): a gateway to reusable semantic vocabularies on the web[J]. Semantic Web, 2017, 8(3):437-452.
- [16] Vandenbussche P Y, Vatant B. Metadata recommendations for linked open data vocabularies [EB/OL]. [2018-07-11]. http://lov.okfn.org/Recommendations_Vocabulary_Design.pdf.
- [17] Alonen M, Kauppinen T, Hyvönen E. Vocab.at-automatic linked data documentation and vocabulary usage analysis[EB/OL]. [2018-06-20]. <http://www.seco.tkk.fi/publications/submitted/alonen-et-al-vocab.pdf>.

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。

张永娟 中国科学院上海生命科学研究院副研究馆员。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, 副院长。江苏 南京 210093。

(收稿日期:2018-07-11;修回日期:2018-08-03)