

资源描述中的词表重用类型与实现方式*

贾君枝

摘 要 为提高资源描述的规范化及标准化,提升资源之间的互操作能力,各类型词表不断被创建及使用,词表重用已成为资源描述中的关键问题。本文从资源描述的基本结构出发,基于数据模型和资源标注两个阶段对词表层、模式层及数据层进行详细分析;探讨词表层、概念层两种重用类型,认为当前用户更关注于词表层重用,概念层重用将随着词表生态环境的完善得到进一步发展;以数据模型中的类与属性为划分对象,以 RDF 三元组形式入手对词表重用实现方式进行深入研究,有助于明确各种方式之间的差异性及适用环境,能够有效地指导用户运用已有词表对数据集的类、属性及其取值范围进行准确描述,从而提升数据集的质量。图 2。表 1。参考文献 16。

关键词 资源描述 词表重用 词表 数据集

分类号 G254

Type and Implementation of Vocabulary Reuse in Resource Description

JIA Junzhi

ABSTRACT

The purpose of the resource description is to add meaningful data to the described objects, so that machine and human users can effectively identify, find, and discover various types of resources. In order to improve the normalization and standardization of resource description, and enhance the interoperability between resources, various types of vocabulary are constantly created and used to describe resources and can express semantic content contained in resources as much as possible. Vocabulary reuse needs to select proper classes and attributes from the existing vocabularies to define the internal and external characteristics of the described objects, and use shared terms to define the common mode of information exchange. This will realize the accurate description and formal representation of datasets to improve the interoperability between datasets, and avoid ambiguity and conflict of expressions. In the process of data modeling, whether the dataset can be described accurately, whether the connection between data sets can be established clearly, and whether the human-computer readable and comprehensible environment can be provided largely depend on the reuse of vocabularies. The paper analyzes the basic structure of resource description and divides its processing into two phases: data model and resource annotation. According to the function, it is deeply

* 本文系国家自然科学基金一般项目“数据开放环境中的词表重用问题研究”(编号:19BTQ023)的研究成果之一。(This article is an outcome of the project “Research on Vocabularies Reusing in the Open Data” (No. 19BTQ023) supported by National Social Science Foundation of China.)

通信作者:贾君枝,Email:junzhij@163.com,ORCID:0000-0003-1486-673X (Correspondence should be addressed to JIA Junzhi,Email:junzhij@163.com,ORCID:0000-0003-1486-673X)

classified into vocabulary layer, schema layer, and the data layer. Then it discusses the types and implementation methods of vocabulary reuse by taking the bibliographic data of the British Library as an example and shows the differences and applicable environments of various ways, guiding the users of vocabulary reuse to accurately describe classes, attributes, and value ranges of datasets by using existing vocabulary, so as to improve the quality of datasets and offer the basis for their further application. There are two types of vocabulary reuse: vocabulary layer reuse and conceptual layer reuse occurred at the modeling stage according to different users' focus. The vocabulary layer reuse focuses on the selection of vocabularies, and the concept layer reuse focuses on the matching degree of class and attribute. It is evident that current users prefer the vocabulary layer to the conceptual layer. With the increasing openness of the vocabularies and sustainable improvement of its ecological environment, the reuse of the concept layer will be further widely used. Taking classes and attributes in the data model as the dividing objects, the realization of vocabulary reuse can be mainly divided into six types: classes or properties reuse, as a superclass/subclass or subproperty/subproperty, domain or range, classes operations, inverse properties, and value restrictions. The choice of vocabulary reuse ways depends on the matching degree of the target vocabulary and the class and attribute in the data model. Different ways of reuse of the same vocabulary influence each other and there are issues of class and attribute coordination and conflict resolution in multiple vocabularies reuse. 2 figs. 1 tab. 16 refs.

KEY WORDS

Resource description. Vocabulary reuse. Vocabulary. Datasets.

0 引言

关联数据、知识图谱等语义网络技术的快速发展使得数据迅速地产生、聚合及传播,从封闭走向开放、从孤立走向互联的数据生态环境逐步形成。图书馆、政府、企业、个人等众多数据拥有者参与到数据开放运动中,积极推动机构内外部数据集的整合、发布,并在此基础上进行应用开发,这对基于词表的资源描述与表示、数据之间的关联等数据处理活动提出了更高要求。词表作为重要的知识组织工具,一直在信息组织与检索领域发挥着重要作用,而且随着信息技术发展,其内涵及外延也在发生着变化。万维网联盟将词表定义为:语义网环境下描述和表示所关注领域的概念及关系,以实现对所描述对象的特征表述、类型划分及其条件限制^[1]。词表不仅包括传统受控词表(分类表、主题词表)、各种类型的元数据集,还包括语义网络、本

体等。资源描述旨在运用词表中的术语描述资源内在的特征及资源之间的关系,以实现资源的有效存取。由于自建词表成本高、效率低,万维网联盟一直强调词表重用原则,尤其要求政府机构所发布的数据集尽可能重用标准化词表,而不是自建词表。因此,资源描述过程中最大程度地选用已有词表已达成共识。词表重用需要从现有的词表中选择适合的类及属性来定义所描述对象的内外特征,使用共享词表的术语定义表述信息的数据模型,以实现数据集的准确描述及形式化表示,提高数据集之间的互操作能力,避免表述的歧义及冲突^[2]。

数据建模过程中数据集能否被准确清晰地描述,能否有机地建立数据集之间的联系,能否提供充分的人机可读、可理解环境,很大程度上取决于词表的重用方式。如何有效地从现有的词表中高效且准确地获取自己所需的术语,实现对资源的有效语义表达,成为数据发布者所面临的关键问题。本文旨在分析资源描述的基

本结构,从处理流程中剖析词表重用可能发生的环节,并对词表重用类型及实现方式进行深入探讨,以图书馆书目数据为例,形象地展示各种方式的差异性及其适用环境,以指导词表重用的用户能够运用已有词表对数据集的基本特征、属性类别及其取值范围进行准确的描述,从而提升数据集的质量,为进一步应用奠定基础。

1 研究综述

语义网的发展基础是词表共享,近年来围绕词表的研究得到了进一步的深化,不再限于词表的自动构建及标引问题,而是更加注重词表的描述表示及词表使用,正如 Thomas 等提出:词表的维护者不仅要重视术语的 URL 定义、元数据提供及其再利用问题,而且要关注词表使用现状、词表间关系、维护更新说明等^[3]。关联数据集中词表重用影响因素、重用行为及词表重用策略分析成为当前关注的重点。研究表明,词表及数据集本身的特点影响着重用,Schaible 调查分析了参与关联数据建模的人员,发现 87% 的人认为重用词表对于形成高质量的数据集有直接作用,词表重用的影响因素包括词表声誉度、词表的更新与维护频率、数据集的质量、数据集的消费及其互操作等^[4]。关联数据集是词表重用的重要体现,多个学者展开了对数据集的调研,旨在分析词表重用行为及策略。Nogales 等应用网络分析法对关联开放词表的使用情况进行分析,讨论了词表使用总频次、数据集的词表使用及词表的语言、专业归属度等问题^[5];李捷佳分析了关联开放数据云(LOD)出版类数据集词表的重用特点,发现建模人员更愿意使用领域依赖度高的专业词表作为重用对象^[6];Schaible 等通过调研关联数据集,将关联数据模型化所采用的重用策略划分为重用流行性词表、最小化词表数、聚集关键领域等^[4]。为提高词表重用效率,帮助用户更好地选择重用词表,研究者们探讨了词表推荐方法及工具。Schaible 等提出了词表重用方法

TermPicker,利用关联开放数据云的 RDF 类及属性信息为用户推荐未来使用词表^[7];Vandenburg 等基于关联开放词表(LOV)的特点来确定高质量词表以供用户重用^[8];Ellefi 等创建了词表推荐系统,旨在为元数据设计人员提供数据建模过程中重用的词表术语列表及元数据信息^[9]。

从现有研究成果看,研究者们更多地关注词表重用后用户行为分析及词表推荐工具设计,对词表重用过程的剖析存在不足,缺乏对词表重用类型及方式等内容的深入探讨,这些问题一定程度上会导致用户盲目地使用已有词表进行资源描述,影响数据集再利用的效率。为此,本文重点分析总结词表重用的类型,探讨词表重用的实现方式,为准确的资源描述奠定基础。

2 资源描述中词表重用类型

资源描述旨在为所描述对象添加各类有意义的数据,以便于机器和人类用户能够有效地识别、查找、发现各类型资源。为提高资源描述的规范化及标准化,提升资源之间的互操作能力,各类型词表不断被创建及使用,使得资源描述有章可循,并尽可能地表达资源所包含的各类语义信息。

2.1 资源描述层次划分

随着统一资源定位符(URI)、资源描述框架(RDF)等语义网技术的发展,资源描述的结构可以看作是一个包含主体、谓词及客体的三元组陈述。主体指代被描述的对象,谓词用来表示资源对象的属性(包含资源对象间的关系),客体表示属性的取值。资源对象可以是万事万物,既包括信息资源,如期刊、图书、报纸、网页、图片等有一定载体形态的文献及多媒体资源,也包括非信息资源,如人、机构、地点、物体、事件、抽象概念等实体对象。一个资源对象通常有多个特征,因此会形成多个陈述,由若干个三

元组构成,这些陈述构成了以资源对象为核心的描述。不同资源对象之间有多种语义关系,从而形成了包含多个描述的 RDF 数据集。如何将资源对象所包含的语义特征及关系类型,运用三元组方式简洁而准确地表达出来,成为资源描述的关键。

通常情况下,资源描述要面对一系列具有共同特征的资源对象集合。因此需要基于用户需求及资源特点,抽象出反映这些资源的普遍、共有的属性特征,从当前词表中挑选出与这些特征相匹配的术语进行表示,最终形成包含各个资源对象的描述集。资源描述按照处理流程可划分为数据建模及资源标注两个环节,其中,数据建模提供了资源表示的基本结构,资源标注是描述集生成的过程。依据此流程,按照不同功能又可将资源描述划分为三层:词表层、模式层、数据层。词表层和模式层作用于数据建模环节,数据层作用于资源标注环节。词表层为模式层的形成提供了丰富的术语来源集合,模式层为数据层的形成提供了基本的描述结构。

数据建模运用实体关系模型,定义模型中的实体类、属性及类间关系,形成数据模式图,旨在形成对问题域中客观存在事物的准确描述。实体类即独立存在且可以唯一标识的事物,具体而言,通常表示相似对象的类、组或类别。属性是实体或实体类型的特征,没有独立含义,通常依赖于实体或实体类。数据模型的构建,需基于用户需求分析分离出用户感兴趣的各类实体,并厘清各实体间的语义关系。数据建模形成了模式层(又称为本体层),从语义层面抽象出资源背后的知识逻辑,将资源表述的各种事实置于知识体系下进行规范化表达,形成具有层次结构的知识库,以实现各类事实的有序化组织及推理。

模式层的形成需依赖于词表可共享的术语集合,以刻画各种资源对象的特征。词表基于

人类对客观世界知识达成的共识,揭示了领域层面可共享的知识体系,提供了数据建模的词汇术语集合,用来表示模型中的实体类及属性。词表在不同模式层是可共享的,因此可以将模式层所依赖的词表抽离出来,并置于模式层之上,以便于不同机构在资源描述中广泛地采用共同的词表进行表达,这有助于推动资源描述所形成的数据模式层面的交流与共享,并实现以词表术语为纽带的数据关联。

资源标注是数据模型实例化的过程,是基于所构建的数据模型对具体资源对象进行 RDF 描述,最终形成关于描述资源对象的多个陈述。资源标注形成了数据层,将不同来源的数据按照模式层所规定的形式进行知识抽取,置于对应的类及属性下,形成有序化描述事实的集合。数据层位于模式层之下,通过模式层为各类型数据添加语义内容,使标注形成的数据集更易于后期基于语义理解的各种应用,比如搜索、问答及翻译等。

词表层、模式层、数据层三层之间的关系如图 1 所示。以大英图书馆的书目数据为例,自 2011 年 7 月开始,大英图书馆将英国国家书目数据(The British National Bibliography, BNB)发布为关联数据,目前近 440 万条记录已转换成三元组^[10]。大英图书馆对书目记录进行描述时,首先定义了针对图书资源类的数据模型,包含了主题、出版、作者三大主要实体类 37 个子类及近百个属性的数据模型图,详见图 1 中的模式层。数据模型的搭建充分利用了各种词表的术语,其自定义词表 blt 所包含的类、属性数量少,大量重用了 DC、BIBO、FOAF、BIO、RDA、ORG、EVENT 及 DDC、LCSH 等多个词表,见图 1 中的词表层,共同完成了对资源的建模。数据层显示资源名为“Complete works of Voltaire”的 RDF 图,明确其不同属性的取值,生成了多个三元组,部分 RDF/XML 表示如下:

```
<rdf:Description rdf:about="http://bnb.data.bl.uk/id/resource/013767349">
```

```

<rdf:type rdf:resource="http://purl.org/dc/terms/BibliographicResource"/>
<dc:title>Harry Potter</dc:title>
<dc:creator rdf:resource="http://bnb.data.bl.uk/id/person/RowlingJK"/>
<dc:language rdf:resource="http://lexvo.org/id/iso639-3/eng"/>
<dc:subject rdf:resource="http://bnb.data.bl.uk/id/concept/lcsh/WizardsFiction"/>
<bibo:isbn10>0747594562</bibo:isbn10>
<blterms:publication
rdf:resource="http://bnb.data.bl.uk/id/resource/013767349/publicationevent/LondonBloomsbury2007"/>

```

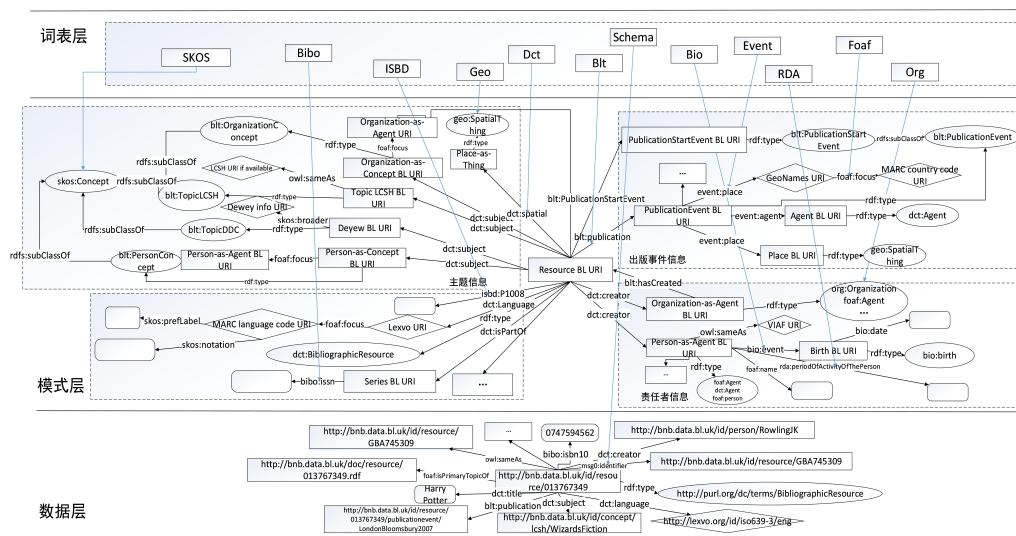


图1 资源描述层次

2.2 词表重用类型

词表作为为数据集添加语义的重要工具,提供了定义数据集及数据之间明确链接类型的术语集,成为数据网络形成的粘合剂。为适应语义网发展及资源描述的需求,词表作为互联网中的关键资源,已逐步走向形式化、开放化、共享化,以确保人或机器能够获取及理解资源的语义信息。为确保资源描述过程中能够更好地引用词表的术语,每个词表都设有 URI 命名,词表中的术语也采用 URI 方式进行定义,这种持久的 URI 形式使用户更易于获取到 URI 对应实体对象的描述信息,而不再重新定义,同时也有利于定义实体之间的关联以实现跨机构的

数据聚合。词表概念表述的清晰、稳定、可解引、可获取等特征在一定程度上影响着词表的重用。

词表重用主要发生在数据建模环节。为提高数据建模效率、提升数据集之间的互操作能力,借鉴已有词表并从已有词表中选取已定义好的类、属性是数据建模的基础。数据建模有自底向上和自顶向下两种方法,自顶向下方法直接设计数据模型,结合用户需求及专家经验构建,一定程度上取决于专家对领域知识及其各类型词表的熟悉度;自底向上方法则从丰富的数据源中提取出关键实体及关系,在此基础上定义数据模型,更强调从现实情境中进行归

纳总结,并借助词表中的概念及概念关系,设计算法实现实体及关系的识别与聚类。无论哪一种方法,都涉及词表的选用及表示问题。

依据用户关注点不同,词表重用可以分为两种类型:词表层重用、概念层重用。通常情况下,两者结合在一起使用,共同实现对数据的描述与表示。词表层重用注重词表的选用,概念层重用更关注词表中概念及关系的重用,而不太关注词表的来源。Schaible 等人的研究表明,选取流行度高的词表已成为首选重用策略^[11],这在一定程度上表明当前用户更关注词表层重用。

(1) 词表层重用

词表层重用首先需确定描述对象的领域范围,查找此领域内部合适的标准化词表,从词表中直接选取与数据模型的类、属性相对应的术语进行表示,如果在该领域词表中找不到相关概念,则通过选择其他领域词表或者自定义词表并建立与所用词表的映射关系来完成构建。

词表层重用适用于所描述的资源对象类型较为单一、领域内标准化词表具有高流行度及使用人员对词表熟悉度高等情况。流行度高的词表重用使得同一领域内数据建模所使用的词表较为集中。如果不同机构的资源采用相同词表进行描述,这在一定程度上能够提升数据集

之间的互操作效率。比如关联开放数据(LOD)平台发布的 21 个书目数据集^[2],所使用的主要是 DC、RDA、SKOS、BIBO、FOAF、FRBR 六个词表,其中 RDA、BIBO、FRBR 是书目领域的词表,FOAF 词表用来描述书目所涉及的作者、出版机构等责任者,SKOS 词表用来表示书目的主题概念及类号,各个复用词表分布如图 2 所示。而美国国会图书馆^[12]、SHARE-VDE^[13]、LD4P 项目^[14]的书目数据都是基于 BIBFRAME 单个词表建模。同样在一些专业领域中,如医学、农业、军事、环境、统计等,复用词表较为集中。用户选用熟悉度高的词表能够体现出一定的用户偏好度,具有主观性、随意性的特点,同样的资源类型可能因数据发布者差异而选用不同词表。词表层重用侧重于数据描述中使用的词表数量较少,用户通常希望所选择的目标词表能够最大程度地满足建模要求。例如,从 VOID 词表的 SPARQL 站点获取到 1 881 个数据集,通过分析发现,每个数据集使用的词表数平均为 3.8 个,其中使用 4 个词表进行描述的占比为 33%^[15],因此用户首先关注的是找到能够满足建模需求的合适词表。在词表重用过程中,元数据信息(词表发布机构、主题、版本、更新频率、获取途径、使用频率等)的获取对于词表的识别与选取也具有一定的参考价值。

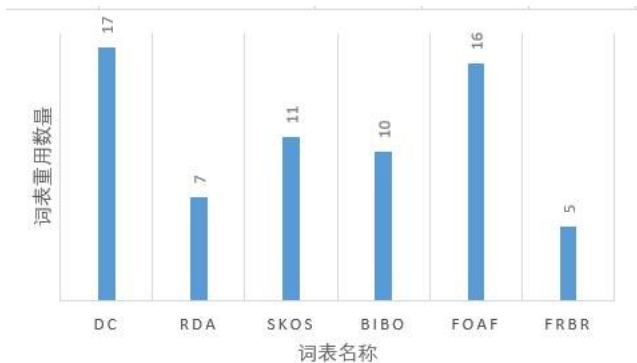


图 2 LOD 中国图书馆书目数据重用的词表

(2) 概念层重用

概念层重用首先通过分析描述对象得到重

要的概念集合及其关系,寻找能够包含这些概念及关系的词表,形成一系列可参考的词表集

合,再基于词表的发布机构、流行度、类及属性等特性确定所选用的词表。概念层重用更关注词表中出现的概念集合能否满足用户需求,这需要通过对词表的评估与排序来确定所选择的集合。概念层重用多适用于资源描述对象有一定复杂性、涉及多个领域或者跨领域、各领域内可选的标准化词表数量不多、用户对词表不熟悉等多种情况。概念层重用要求数据描述中使用多个词表,各个词表相互补充共同实现对资源对象的描述。词表重用过程中,词表的类及属性的定义清晰度、可获取、可参引、语种、用例等将会影响到选取过程。

与词表层重用相比,概念层重用的工作量大,更适于机器自动获取及自下而上的数据建模环境,并更依赖于良好的词表生态环境。易于获得的可选词表集合、一站式的词表查询平台、词表术语的使用频率及丰富用例信息等都会帮助用户正确地选用词表。当前有一些词表

注册平台如 LOV(关联开放词表)、元数据注册平台(Open Metadata Registry)等,可以帮助用户定位包含词表的术语信息。随着词表开放度的增加及词表生态环境的逐步完善,概念层重用将得到进一步推动。

3 词表重用的实现方式

依据资源描述的特点,词表重用发生在模式层,用于帮助构建描述对象的概念模型,主要由类、属性及一系列限制规范构成。词表重用一旦建立完成,就可用于资源对象标注。三元组是最基本的描述单元,词表重用以三元组形式得以体现。将重用词表定义为目标词表,以数据模型中的类与属性为划分对象,词表重用的实现方式可分为六种类型,各类型的具体含义见表1所示,表中所选事例主要来源于大英图书馆针对图书的数据模型,其他数据集作为补充。

表1 词表重用方式

词表重用方式	举 例	说明	功能
类 属性	dct:creator(属性) bibo:Book(类)	重用目标词表 DC、BIBO 中的术语分别表示类或属性	直接重用,简单方便
父/子类 父/子属性	blt:PublicationEvent rdfs:subClassOf event:Event blt:bnbsubPropertyOf dc:identifier	重用目标词表 E-VENT、DC 中的术语分别表示自定义类或属性的父类或父属性	直接重用,用以建立层级关系
定义域与值域	intervals: hasXsdDurationDescriptionrdfs: domain owl:unionOf(time:DurationDescription time:Interval)	重用目标词表 TIME 定义自定义属性的定义域	间接重用,用以限制属性
类运算	holding:Itemowl:unionOf (bf:HeldItem frbr:Item rdac:Item).	重用目标词表 BF、FRBR、RDA 定义新类	间接重用,用以创建类
逆属性	dct:creatorowl:inverseOf blt:hasCreated	定义两个属性之间互逆,BLT 为自定义词表,DC 为重用词表	间接重用,用以限制属性
属性值限制	<owl:Class rdfs:about = "&gn;GeonamesFeature"> <owl:onProperty rdfs:resource = "&gn;featureClass"/> <owl:someValuesFrom rdfs:resource = " http://www. w3. org/2004/02/skos/core #ConceptScheme"/>	重用目标词表 SKOS 概念限制属性取值	间接重用,用以限制属性值

3.1 类或属性

目标词表的类或属性直接用来定义数据模型中的类或属性,以表达三元组中的谓词或客体。表1中 bibo:Book 表示图书类,dct:creator 表示作者属性,直接用来表示大英图书馆图书数据模型中的类及属性。该类型使用方便,已成为目前最常用的重用方式,复用词表的术语能够有机地架构起数据模型的结构,但要求目标词表术语能够充分表达数据的含义。

3.2 父/子类或父/子属性

目标词表的类或属性作为数据模型中的父/子类或父/子属性,旨在帮助数据模型构建语义层级体系。概念层级体系为数据集提供了丰富的上下文环境,为进一步实现推理奠定了基础。选择目标词表的类或属性作为父类或父属性,其实是概括化过程;选择目标词表的类或属性作为子类或子属性,其实是具体化过程。将数据模型的类或属性与目标词表建立联系,既能够丰富模型的概念体系,又能够借助词表提升数据集与外部数据集之间的交互。此重用方式使用较普遍,表1中将自定义的 blt:PublicationEvent 类定义为 event:Event 子类,将 blt:bnb 属性定义为 dc:identifier 子属性。

3.3 属性定义域、值域

自定义属性的定义域或者值域来源于目标词表。定义了属性后,需要定义其所属的定义域及值域。属性定义域声明给定属性所属资源是定义域类的实例,属性值域声明属性的取值是值域类的实例,而这些类则来源于重用词表。实质上,词表提供了对属性进行限制的类集合,要求用户从复用词表中选用相应的术语进行实例化,以实现描述的规范化及一致性。此重用方式在相同语义下可以转换为类的重用,即将属性的定义域及值域明确指定为某一词表的类。如 blt:Publication (Resource BL URI, PublicationEvent BL URI) 为自定义的属性,将其定义域“Resource BL URI”直接定义为 bibo:Book 类,

通过类重用实现对定义域的限制说明。表1中自定义属性的定义域来自于目标词表 TIME 两个类的并运算。该复用方式通常结合类运算实现。

3.4 类运算

自定义类是在重用目标类运算的基础上获取。在对应词表中没有找到合适的类,可以通过词表术语之间的布尔组合运算、不相交类定义等方式明确数据模型中的类,通过交集、并集或补集运算以构造新类来表示描述对象。此类属于非直接重用,对象词表中的类指代的实例集合与数据模型不一致时,可以通过类运算进一步明确。表1中的 holding:Item 类是基于 bf:HeldItem、frbr:Item、rdac:Item 三个重用词表的类的并运算得到。

3.5 逆属性

将对象词表的属性定义为数据模型中属性的逆属性,建立二者之间互逆关系。即对象词表的属性不直接用来表示属性,需在数据模型中定义新属性,并与词表的属性建立逆属性关系。逆属性的定义应该建立在类重用的基础上,数据模型中属性的定义域及值域所指定的类应该来源于目标词表属性的值域和定义域。如表1中自定义的属性 blt:hasCreated 建立了与 dct:creator 的互逆关系,其定义域实际上是 dct:creator 的值域 dct:Agent。

3.6 属性值限制

使用重用词表对属性的取值进行全称或存在限制。全称限制 (owl:allValuesFrom) 用来规定指定属性的所有值都只能从重用词表的类中选取,存在限制 (owl:someValuesFrom) 要求指定属性至少有一个值从重用词表的类中选取。表1中 gn:GeonamesFeature 类指定 gn:featureClass 属性至少有一个值来自于 SKOS 词表中的概念。

词表重用方式的选择,依赖于目标词表与数据模型中类与属性的匹配程度。如果二者之

间完全匹配,则选择类或属性的直接重用。如果在目标词表中找不到完全对应的类,则考虑在目标词表基础上通过类组合运算构造新类,该方式属于间接重用。目标词表的类直接或间接方式都不能用来表示数据模型的类,则需自定义新类,并进一步考虑与目标词表的类建立子类或父类关系,以便于将新定义的类与词表建立关系。如果在目标词表中找不到完全对应的属性,则需建立新属性,定义其定义域及值域。取值可能是来自于词表中重用的类,也可能是数据模型中自定义类。与目标词表的属性建立子属性或父属性关系,有助于将新属性与词表建立关系。对于一些特定属性的取值,还需进一步考虑使用重用词表中的类成员进行值限制。如果定义域、值域都来自于重用类,则考虑能否将此属性定义为目标词表的逆属性,逆属性的复用在一定程度上与属性重用具有相同效果。通常情况下,类与属性、父(子)类与父(子)属性、属性值限制等重用方式的应用较多。

同一词表的类与属性的重用之间相互影响。属性的重用意味着其定义域与值域来自于词表的类,属性重用增加了同一词表内部类重用的可能性,比如 `det:creator` 属性的重用导致 `det:Agent` 类重用。同样,同一词表类重用的数量越多,类与类之间的语义关系重用的可能性越大,导致属性重用的数量增多。数据集重用的词表数量不一,可以分为词表最大化和最小化重用策略^[16]。词表最大化策略旨在重用尽可能多的词表来覆盖建模领域的所有术语,词表最小化策略旨在重用少量词表以满足建模需求。词表重用数量越多,重用的难度就会增加,也就会涉及多个词表的类、属性之间的冲突与

协调问题,因而用户选取重用方式时需要慎重考虑此类问题。

4 总结与展望

资源描述是数据集得以深层次开发利用的前提,当前信息技术环境使资源描述对象及其描述方式、语言发生了深刻变化,资源描述对象范围不断扩大,从馆藏集合、单件发展到知识单元,资源描述方式从人工发展到自动,资源描述语言从 HTML、XML 发展到 RDF,使得形成的描述集更益于用户使用。数据建模与资源标注是资源描述的两个基本过程,其有机地将词表、模式、数据三个功能层集成在一起,实现了数据的语义化表示。词表重用作为数据建模中的重要环节,旨在搭建数据层之上的共享知识层,为数据集提供良好的语义框架。依据用户关注点的差异,词表重用可以划分为两种类型:词表层重用注重词表的选取,概念层重用关注类及属性的匹配度,二者共同作用来实现对数据的描述与表示。词表重用方式的选择,依赖于目标词表与数据模型中类与属性的匹配程度,用户依据实际情况选择类与属性、父(子)类/属性、定义域与值域、类运算、逆属性、属性值限制等重用方式。

同一词表不同的重用方式之间彼此互相影响,多个词表重用还存在着类与属性的协调与冲突问题,但本文没有深入探讨不同重用方式、不同词表之间的相互影响,未来研究将着重从这两个方面入手,以帮助用户更好地选择重用方式,有效地平衡词表之间的关系。

参考文献

- [1] Vocabularies[EB/OL]. [2019-04-23]. <https://www.w3.org/standards/semanticweb/ontology>.
- [2] 贾君枝. 词表生态系统:构成要素及关联关系[J]. 中国图书馆学报,2020,46(4):60-68. (Jia Junzhi. Vocabulary ecosystem:components and correlations[J]. Journal of Library Science in China,2020,46(4):60-68.)
- [3] Thomas B,Pierreyves V,Bernard V,et al. Requirements for vocabulary preservation and governance[J]. Library

- Hi Tech,2013,31(4):657-668.
- [4] Schaible J,Gottron T,Scherp A. Survey on common strategies of vocabulary reuse in linked open data modeling [C]//The semantic web:trends and challenges. ESWC 2014. Anissaras, Crete, Greece: Springer International Publishing,2014:457-472.
- [5] Nogales A,Siciliaurban M A,Garciabariocanal E,et al. Measuring vocabulary use in the Linked Data Cloud[J]. Online Information Review,2017,41(2):252-271.
- [6] 李捷佳. 关联数据的词表重用策略研究[D]. 太原:山西大学,2019. (Li Jiejia. Research on vocabularies reusing strategy of linked data[D]. Taiyuan:Shanxi University,2019.
- [7] Schaible J,Gottron T,Scherp A. TermPicker:enabling the reuse of vocabulary terms by exploiting data from the linked open data cloud:an extended technical report[J]. International Journal of Database Management Systems, 2015,4(5):21-37.
- [8] Vandenbussche P,Atemezing G A,Povedavillalon M,et al. Linked Open Vocabularies (LOV):a gateway to reusable semantic vocabularies on the web[J]. Sprachwissenschaft,2016,8(3):437-452.
- [9] Ellefi M B,Bellahsene Z,Todorov K. Datavore:a vocabulary recommender tool assisting Linked Data modeling [EB/OL]. [2020-01-20]. http://www.ceur-ws.org/Vol-1486/paper_32.pdf.
- [10] The British National Bibliography[EB/OL]. [2020-04-23]. <https://bnb.data.bl.uk>.
- [11] Schaible J,Gottron T,Scherp A. Extended description of the survey on common strategies of vocabulary reuse in linked open data modeling[EB/OL]. [2020-04-23]. https://kola.opus.hbz-nrw.de/opus45-kola/frontdoor/deliver/index/docId/820/file/2014_01_Arbeitsberichte.pdf.
- [12] Bibframe[EB/OL]. [2020-04-23]. <https://www.loc.gov/bibframe>.
- [13] Tiziana P. From MARC to Bibframe in the SHARE-VDE project [EB/OL]. [2020-03-03]. <https://www.loc.gov/bibframe/news/pdf/share-vde-alaal2018.pdf>.
- [14] Schreur P E. Linked data for production (LD4P):a multi-institutional approach to technical services transformation[C/OL]. [2020-06-02]. <https://www.researchgate.net/publication/324635180>.
- [15] 贾君枝,马文雯. 基于 VOID 词表的数据集及链接语义关系分析[J]. 情报科学,2020,38(10):97-103. (Jia Junzhi, Ma Wenwen. Structure and link semantic relationship analysis of vocabulary of interlinked datasets[J]. Information Science,2020,38(10):97-103.
- [16] Bosch T,Zapilko B,Wackerow J,et al. Towards the discovery of person-level data reuse of vocabularies and related use cases[EB/OL]. [2020-05-03]. <https://www.researchgate.net/publication/303342945>.

贾君枝 中国人民大学信息资源管理学院教授,博士生导师,北京 100872

(收稿日期:2020-12-07;修回日期:2021-01-07)