

关联数据云图中出版类数据集特点分析

贾君枝 寇蕾蕾

摘要 出版类数据集作为关联开放数据(LOD)云图中的重要数据集合,成为继社交网络类、政府类数据集之后增速最快的一类,受到了广泛关注。论文旨在通过对 LOD 云图中出版类数据的深入分析,从已发布的数据集、关联应用、词表使用、元数据信息等最佳实践角度入手,探索国外关联开放数据的服务模式,进而为我国发展关联开放数据提供参考。图 4。表 7。参考文献 19。

关键词 出版类 数据集 关联开放数据

The Characteristic Analysis of Publication Datasets in Linked Open Data Cloud

Jia Junzhi Kou Leilei

Abstract: As an important dataset of Linked Open Data (LOD) cloud, publication datasets have become an fastest-growing category after social web and government, and received a widespread attention. This paper aims at exploring foreign linked open data service mode and providing reference for the development of China's linked open data through deeply analyzing the best practices of publishing resources, such as published datasets, interlinking, usage of vocabulary, and adoption of metadata. 4 figs. 7 tabs. 19 refs.

Keywords: Publication; Dataset; Linked Open Data

1 导言

万维网之父 Tim Berners-Lee 在 2006 年第一次提出关联数据(Linked Data)概念时,即采用 RDF 数据模型,利用 URI(统一资源标识符)命名数据实体,通过 HTTP 协议揭示并获取数据,同时强调数据的相互关联。其目的是构建一张计算机能理解的语义数据网,以便在此基础上构建更智能的应用。基于更好地应用关联数据的需求,Chris Bizer 等在 2007 年 5 月向 W3C SWEO(语义网教育和宣传小组)提交了 LOD(Linked Open Data)项目申请,该项目旨在号召人们将现有数据发布成关联数据,并将不同数据源互联起来,以可视化图形的方式将互联的关联数据集展现出来,通过链接现存、分散的数据来创造知识,开展数据整合服务,实现数据的增值。在过去的几年中,越来越多的数据提供者和网络应用开发者将各自的数据发布到网络上,并与其它数据源关联在一起,形成了一个巨大的数据网络。截至 2014

年 4 月,世界各机构已经基于 LOD 标准发布了数千个数据集,包含数千亿个 RDF 三元组。相比于 2011 年,数据集数量增长了将近一倍,其中描述不同实体的常用词表显著增加,但提供的起源和授权元数据较少。依据 datahub.io 类目标标准,LOD 数据分为八大类别,包括媒体类、政府类、出版类、生命科学类、地理类、社交网络类、跨领域及用户生成内容。媒体类包含提供电影、音乐、电视和广播节目及印刷媒体信息的数据集,如纽约时报、BBC 广播节目;政府类包含由联邦或地方政府发布的关联数据,其中包括许多统计性数据集;出版类包含馆藏资源、科学出版物及会议信息、大学读物列表、知识组织工具等相关数据集;地理类包含涉及地理实体、地缘边界、热点地区信息的数据集;生命科学类包括生物和生物化学信息、药物相关数据、以及有关物种及其栖息地信息的数据集;跨领域类包含基于语言资源、产品数据等基础知识的数据集;用户生成内容包含从由较大用户群体组成

的门户网站所收集数据的数据集。其中出版类数据集增长迅速,成为继社交网络类、政府类之后增速最快的一类,受到了广泛关注。出版类数据的来源机构包括出版单位、图书馆、博物馆、档案馆以及高校等,这些机构积极参与数据的开放互联运动,推动了数据的发布、获取、相互关联,提高了关联数据的质量。本文旨在通过对 LOD 中出版类数据的深入分析,探索国外关联开放数据的服务模式,进而为我国发展关联开放数据提供参考。

2 出版类数据集归类

LOD 云图中共有 1014 个数据集^[1],其中出版类共有 96 个,所占百分比为 9.47%。在出版类中,按照数据内容可将数据集划分为四类:馆藏资源、

科学出版物和会议信息、大学信息及读物列表、知识组织工具。对出版类数据集的详细分类,有助于用户清晰地了解资源分布情况,准确定位信息。

2.1 馆藏资源数据集

馆藏资源数据集是对文化机构中的馆藏资源进行描述,主要包含图书馆、博物馆、档案馆以及其它信息机构的数据集。通过对馆藏资源的语义描述和链接来实现资源内容的充分揭示及关联关系的规范表达,进而为文化机构中数据的深度聚合与知识发现提供服务。如表 1,馆藏资源占出版类资源的 24%,以德国国家图书馆关联数据(DNB)为例,目前已包含 192,556,756 个 RDF 三元组^[2],数据间可相互引用,且其在 LOD 中也被许多数据集链接。从形式来看,其应用格式包括 rdf 和 xml。

表 1 出版类中馆藏资源数据集

数据集名	描述	URI 地址
DNB	德国国家图书馆	http://www.dnb.de/EN/datendienste/linkedData
Bluk BNB	大英图书馆书目资源	http://www.bl.uk/bibliographic/datafree.html
data.bnf.fr	法国国家图书馆书目资源	http://data.bnf.fr
Princeton	普林斯顿大学图书馆	http://findingaids.princeton.edu/
oszk.hu	匈牙利国家图书馆	http://nektar.oszk.hu/wiki/Semantic_web
Sztaki LOD	匈牙利档案	http://lod.sztaki.hu/
libver.math	维罗纳公共图书馆	http://libver.math.auth.gr
worldcat	在线联合目录	http://www.worldcat.org/
Sudoc.fr	法国学术联合目录	http://punktokomo.abes.fr/2011/07/04/le-sudoc-sur-le-web-de-donnees/
theses.fr	法国学位论文搜索引擎	http://www.theses.fr
LIBRIS	瑞典联合目录的规范数据	http://libris.kb.se
ZDB	德国联合目录的连续出版物	http://www.zeitschriftendatenbank.de/services/schnittstellen/linked-data/
BritishMuseumCollection	大英博物馆	http://collection.britishmuseum.org
Amsterdam museum	阿姆斯特丹博物馆	http://semanticweb.cs.vu.nl/lod/am/
ArchivesHub	英国档案中心	http://data.archiveshub.ac.uk
muninn project	世界档案馆有关一战的记录	rdf.muninn-project.org
Gutenberg	古登堡电子书	http://www.gutenberg.org/
Open Library	加州图书馆图书	http://openlibrary.org/
Bib Base	谷歌书目信息目录	http://data.bibbase.org
L3S DBLP	计算机科学书目数据	http://dblp.l3s.de/d2r/
airforcehistoryindex	美国空军历史索引库	http://airforcehistoryindex.org
MisMuseos.net	艺术资源的语义索引库	http://mismuseos.net/

2.2 科学出版物和会议信息数据集

科学出版物和会议信息数据集主要提供了与科学出版物和会议信息相关的数据集(如表2)。该数据集占比18%,为科学研究的开展提供了丰富的来源信息。会议信息数据集中,以关联数据

会议(Colinda)为例,其提供了会议的地点、时间等基本信息,包含从2003年至2013年大约15,000个会议的信息,并与地理、维基百科及计算机科学等会议建立关联^[3]。同时以可视化图形的形式展现出来,便于用户获取各种类型会议的信息。

表2 出版类中科学出版物和会议资源数据集

数据集名	描述	URI 地址
colinda	关联数据会议	http://colinda.org
semanticweb.org	语义网会议	https://semanticweb.org/
TWC IEEE VIS	科技会议中论文数据集	http://ieeewis.tw.rpi.edu
PUB Bielefeld	比勒费尔德大学的出版物	http://pub.uni-bielefeld.de/
jewishencyclopedia	犹太百科全书	http://jewishencyclopedia.com
Deutsche Biographie	德意志传记	http://www.deutsche-biographie.de/
oreilly	电子书和科技会议	http://oreilly.com
crossref	学术出版商协会	http://crossref.org
techniques	自然出版集团信息,包括技术、主题、出版物、产品、引证、贡献者、目录。	http://ns.nature.com/techniques/
subjects		http://ns.nature.com/subjects/
publications		http://ns.nature.com/publications/
products		http://ns.nature.com/products/
citations		http://ns.nature.com/data-citations/
Datacitations		http://ns.nature.com/contributors/
contributors		http://ns.nature.com/citations/
catalogs		http://ns.nature.com/catalogs/

2.3 大学信息及读物列表数据集

大学信息及读物列表数据集包含了各大学发布的关于人、部门、设施、课程、赠款和出版物等以关联数据形式出现的信息,能够帮助学生全面了解学校概况,发现不同课程、不同知识之间的相互关系,便于学生学习(如表3)。该数据集占比30%,居于出版类数据集的首位。以曼彻斯特大学阅读书目为例^[4],其提供了曼彻斯特大学图书馆的检索界面,用户可以查看资源列表、学科模块以及相关课程,并可通过定制方式快速查询到自己感兴趣的阅读书目。

2.4 知识组织工具数据集

这部分数据集包含了主题词表、本体、分类表、元数据等知识组织工具,可以被其它数据集引用(如表4)。该数据集占比29%。其中数据量最大的当属美国国会图书馆标题表(LCSH),它包含7,332,816个RDF三元组,自1898年以来一直被用于对国会图书馆的资源进行编目^[5]。LCSH关联数据服务内容包含国会图书馆标题表、主题和形式的细分信息、体裁/形式标题词表、儿童(AC)标题词表及创建规范记录所需的验证字符串^[6]。

表 3 出版类中大学信息及读物列表数据集

数据集名	描述	URI 地址
southampton-ac-uk	南安普顿大学门户网站	http://southampton. ac. uk
SouthamptonECSEPrints	南安普顿大学电子与计算机科学信息	http://eprints. ecs. soton. ac. uk/
VIVOFloridaUniversity	佛罗里达大学门户网站	http://vivo. ufl. edu
DWS-Group	曼海姆大学数据科学研究小组信息	http://dws. informatik. uni-mannheim. de/
bookmashup	书籍信息的集成平台	http://www4. wiwiss. fu-berlin. de/bizer/bookmashup/
DSC Sheffield	谢菲尔德大学计算机科学部门信息	http://data. dcs. shef. ac. uk/
VIVOIndianauniversity	印第安纳大学教职工信息	http://vivo. iu. edu
morelab	德乌斯托大学跨学科艺术工作室信息	http://www. morelab. deusto. es
Aspire Manchester	曼彻斯特大学阅读书目	http://www. readinglists. manchester. ac. uk/index. html
st-andrews	圣安德鲁斯大学阅读书目	http://resourcelists. st-andrews. ac. uk/
bourne-mouth	伯恩茅斯大学阅读书目	http://readinglists. bourne-mouth. ac. uk/index. html
Bolton	博尔顿大学阅读书目	http://lists. bolton. ac. uk/index. html
Hope	利物浦霍普大学阅读书目	http://lists. hope. ac. uk/index. html
Derby	德比大学阅读书目	http://liblists. derby. ac. uk/index. html
Brunel	伦敦布鲁内尔大学阅读书目	http://readinglists. brunel. ac. uk/index. html
Aspire Portsmouth	朴茨茅斯大学阅读书目	http://lists. lib. portsmouth. ac. uk/index. html
AspireQuml	伦敦大学玛丽学院阅读书目	http://lists. library. qmul. ac. uk/
AspireHarperAdams	哈珀亚当斯大学学院阅读书目	http://readinglists. harper-adams. ac. uk/
Aspire Keele	基尔大学阅读书目	http://lists. lib. keele. ac. uk/index. html
Aspire Brunel	布鲁内尔大学阅读书目	http://readinglists. brunel. ac. uk/index. html
Aspire NTU	诺丁汉特伦特大学阅读书目	http://resourcelists. ntu. ac. uk/index. html
Aspire UCL	伦敦大学学院阅读书目	http://readinglists. ucl. ac. uk/index. html
Aspire Roehampton	罗汉普顿大学阅读书目	http://resourcelists. roehampton. ac. uk/index. html
Aspire Sussex	苏塞克斯大学阅读书目	http://liblists. sussex. ac. uk/index. html
Aspire Plymouth	普利茅斯大学阅读书目	http://lists. lib. plymouth. ac. uk/index. html
Aspire Uclan	中央兰开夏大学阅读书目	http://readinglists. central-lancashire. ac. uk/index. html
Aspire MMU	曼彻斯特城市大学阅读书目	http://lists. lib. mmu. ac. uk/index. html

表 4 出版类中知识组织工具数据集

数据集名	描述	URI 地址
LCSH	美国国会图书馆标题表	http://id.loc.gov/authorities/
Agrovocoskos	联合国粮农组织受控词表	http://aims.fao.org/standards/agrovoc/about
ListaEncabezamientoMateria	欧洲主题词表	http://id.sgcb.mcu.es
STWThesaurusEconomics	经济领域叙词表	http://zbw.eu/stw
FAOgeopoliticalontology	联合国粮农组织地理政治本体	http://www.fao.org/countryprofiles/geoinfo.asp? lang=en
Bible Ontology	圣经本体	http://bibleontology.com/
MSC	数学主题分类表	http://msc2010.org/mscwork/
iconclass	多语言主题分类表	http://www.iconclass.org/help/lod
DeweyDecimaClassification	杜威关联数据服务	http://dewey.info
PlanetData Project Wiki	欧盟数据管理项目	http://planet-data.eu
LOD2 Project Wiki	欧洲委员会关联数据项目	http://lod2.eu
dev8d	英国科研机构的数字技术	http://dev8d.org
Project Gutenberg In RDF	古登堡项目元数据	http://www4.wiwiss.fu-berlin.de/gutendata/
BibSonomy	社会书签共享系统	http://www.bibsonomy.org/
skos.um	科技领域国际标准命名术语	http://skos.um.es/unesco6/
ciard ring	全球农业数据集分类目录	http://ring.ciard.net/
ASN;US	美国教育机构的成就标准	http://asn.jesandco.org
Nbnresolving	URN-解析器的示范	http://nbn-resolving.org
DOI	学术出版物的数字对象识别符	http://dx.doi.org
Semantic Web Grundlagen	语义网的数学基础	http://semantic-web-grundlagen.de
semanticweb.cs.vu.nl	语义网团体的数据集和词表	http://semanticweb.cs.vu.nl
semantic.ckan.net	语义网通用服务器	http://semantic.ckan.net/about/data/
datahub	数据中心的数据集信息	http://datahub.io
datacite	数据引证中心数据	http://datacite.org
JITA	图书情报学的分类表	http://www.destin-informatique.com/ASKOSI/Wiki.jsp?page=JITA%20Maintenance
idref.fr	法国学术联合目录应用程序	http://punktokomo.abes.fr/2011/07/05/idref-des-pages-html-et-rdf-plus-riches/

3 出版类中的关联应用实践分析

关联是关联数据的核心^[7]。关联的最佳实践是鼓励数据发布机构在数据集间设置 RDF 链接,借助 RDF 链接数据提供者可以将自己的数据集与整个数据网络建立连接,通过遵循 RDF 链路能够发现额外有用的数据,从而起到导航的作用。总体而言,LOD 中 56.11%的数据集至少和

一个数据集建立了 RDF 链接。为便于分析数据集间的关联情况,如果不同数据集的资源之间至少存在一个 RDF 链接,我们就认为这两个数据集间建立了连接^[8]。

3.1 出版类中数据集的出入度

“度数”一词来源于数学用语,就一个节点而言,靠近相邻节点的头部端点的数量称为该节点

的入度,靠近相邻节点的尾部端点的数量称为该节点的出度。针对关联数据环境中的某一个数据集,入度是指 LOD 中指向该数据集的 RDF 链接数,出度是指指向 LOD 中其他数据集的 RDF 链接数。度数能够很好地反映整体连接情况,入度值反映了该数据集被其他数据集利用的重要程度,出度值反映了该数据集对其他数据集的需求程度^[9]。数据集的出入度值越高,表明链接越紧密;反之,数据集的出入度值越低,表明链接越稀疏。在 LOD 入度排名前十的数据集类别中,出版类位于社交网络类及跨领域类之后,排名第三。如图 1 所示,出版类 96 个数据集中,入度值大于 10(包含 10)的数据集只有 6 个,占比 6.25%;其余数据集的入度值分布在 0—10 之间,占比 93.75%。而在社交网络类的 520 个数据集中,入度值大于 10(包含 10)的数据集有 63 个,占比 12.12%;入度值在 0—10 之间有 457 个,占比 87.88%。与社交网络类相比,出版类入度值在 10 以上的数据集所占比例较低,入度值在 0—10 之间的数据集所占比例较高。

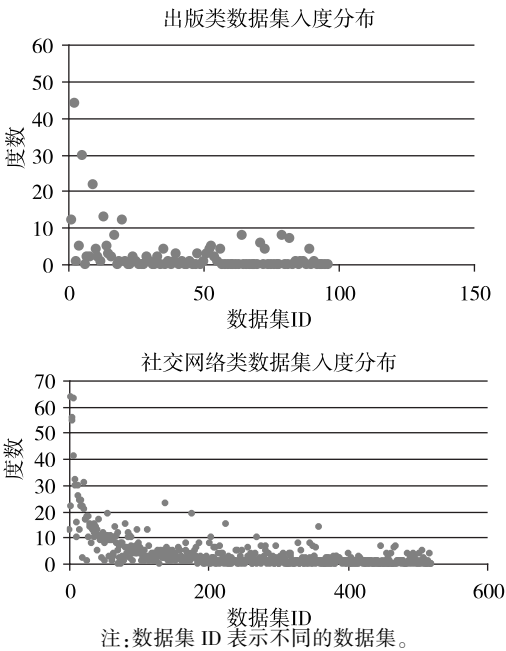


图 1 出版类、社交网络类数据集的入度分布情况

在 LOD 出度排名前十的数据集类别中,出版类仅次于社交网络类,排名第二。如图 2 所示,在出版类的 96 个数据集中,出度值在 20 以上的数据集只有 2 个,占比 2.08%;其余数据集出度值处在 0—20 之间,占比 97.92%。而在社交网络类的 520 个数据集中,出度值大于 20(包含 20)的数据集有 26 个,占比 5%;出度值在 0—20 之间的数据集有 494 个,占比 95%。与社交网络类相比,出版类出度值大于 20 的数据集比例较低,出度值在 0—20 之间的数据集比例较高。总体而言,除了社交网络类,与 LOD 中其他类别相比,出版类出入度值较高,这表明出版类在整个 LOD 中占有重要的地位。但出版类中只有少量的数据集被高度链接,而大部分数据集只是稀疏链接,这也与 LOD 的整体连接情况相符。

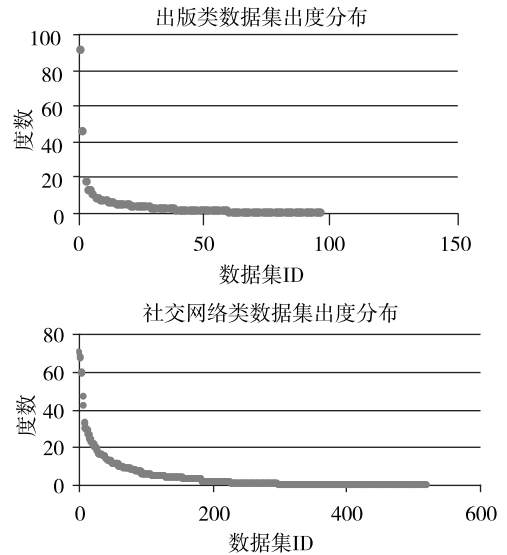


图 2 出版类、社交网络类数据集的出度分布情况

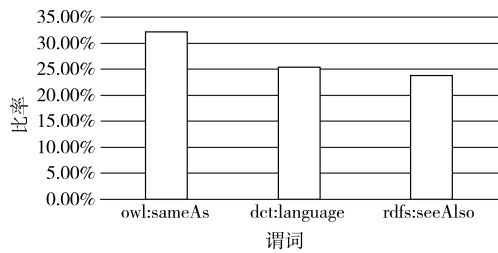
出版类数据集中出入度值较高的数据集如表 5 所示,可以看出,出度值最高的数据集是社会书签共享系统(BibSonomy)^[10],允许用户添加标签来提高数据访问能力;语义网会议(data.semanticweb.org)提供语义网相关会议的数据(如论文、报告、人),其出入度值都高,表明语义网已受到许多机构的关注。

表 5 出版类中出入度值排名前十的数据集

数据集	出度值	数据集	入度值
BibSonomy	91	data. semanticweb. org	44
data. semanticweb. org	45	L3S DBLP	30
PlanetData Project Wiki	19	LCSH	22
Agrovoc skos	12	DeutscheNationalbibliografie(DNB)	13
L3S DBLP	12	BibSonomy	12
Archives Hub Linked Data	8	FAO geopolitical ontology	12
data. bnf. fr	8	LOD2 Project Wiki	8
worldcat	7	DDC	8
morelab	7	Open Library	8
3LCSH	7	DSC Sheffield	7

3.2 出版类中使用的连接谓词

连接谓词能够将类和类、属性和属性关联起来。基于对象之间的关联关系,计算机可以进行有效地推理,进而实现不同数据集间的语义关联^[11]。图 3 列出了出版类资源使用 RDF 链接时



注:百分比是指每类中设置向外链接的数据集的数量同每类中数据集总的链接数之比。

图 3 出版类中使用频率最高的三个连接谓词

用到频率最高的 3 个连接谓词,这与 LOD 整体连接谓词使用情况相符。owl:sameAs 是最常用的连接谓词,表明“两个 URI 引用实际上指向同一事物”^[12]。由于在出版类中存在许多等价的 URIs,因此使用 owl:sameAs 属性能有效地实现数据的集成,有效地聚合指向同一事物对象的所有数据。dct:language 定义了语言属性,通过该属性将不同语言的数据集建立关联,实现了跨语言的数据操作。rdfs:seeAlso 表示对主题资源提供额外的信息,将围绕某一对象的所有相关属性建立链接,实现数据的集中展示^[13]。

3.3 出版类中使用的词汇表

关联数据利用词汇表的词汇,须通过定义属性及属性值来体现其语义特征。如果词表中的词汇出现在数据集中三元组的谓语位置,或者出现在 rdf:type 三元组的宾语位置,则认为该词表被数据集所用^[5]。

3.3.1 常用词表

关联数据中,一些被广泛使用的词表有助于建立不同数据集间的联系,实现数据的互操作。在 LOD 的 1014 个数据集中,超过 5%的数据集都会用到常用词表。其中 foaf、rdfs、dcterms、owl 等是许多主题领域的数据集最常用到的词表。此外,存在这样一种趋势:越来越多的数据集开始使用常用词表^[14]。从表 6 可以看出,出版类经常用到的词表既有 rdfs、owl 等描述语言,也有 dcterms、foaf、bibo 等元数据词表。其中,83%的数据集使用了都柏林核心词表 dcterms;用于描述人物、活动及其关系的 foaf 词表也被 76%的数据集使用;41.67%的数据集使用了 bibo 书目本体,它提供了描述引文和书目参考文献的主要概念和属性。skos 词表及资源清单(resourceList)用于创建大学读物列表。

3.3.2 专有词表的使用

由于常用词表并不能提供在网上发布数据集完整内容所需要的所有术语,因此还需要使用一些专有词表。专有词表是指仅被一个数据集使用的词表^[14]。需要注意的是,如果数据发布机构

表 6 出版类中常用词表使用情况

前缀	出现次数	比率	前缀	出现次数	比率
rdfs	43	44. 79%	owl	57	59. 38%
foaf	73	76%	sioc	30	31. 25%
skos	43	44. 79%	void	14	14. 58%
dcterms	80	83%	resourcelist	19	19. 79%
bibo	40	41. 67%	provenance	16	16. 67%

注:出现次数与比率是衡量词表使用频率的指标,出现次数指在出版类所有词表(包含重复)中的出现次数;比率是指出现次数与总数据集数量(出版类共 96 个数据集)之商。

使用专有词表,那么这种词表应在 RDF 模式或 OWL 定义下是可参引的。专有词表术语定义了除常用词表中术语之外的其它术语,应包含指向常用词表的 RDF 链接,以便更容易地对其进行解释。参引度是指词表中可参引术语的数量占词表中所有术语的数量比例^[15]。其值分布在 0—1 之间,参引值为 0 代表不参引,值在 0—1 之间代表部分参引,值为 1 代表完全参引。其中,部分参引的原因可能是意外使用词表中未定义术语或

对词表中已经弃用的术语没有做出恰当的标记。从图 4 中可以看出,出版类共使用了 54 个专有词表,其中有 12 个专有词表的术语为完全参引,5 个是部分参引,其余的均没有参引能力。由此可见,在出版类使用的专有词表中,它们的参引能力并不是很强,而根据上述描述可知,专有词表术语在 RDF 模式或 OWL 定义下的可参引,有助于更准确地解释词表中的术语,便于知识聚合和发现。

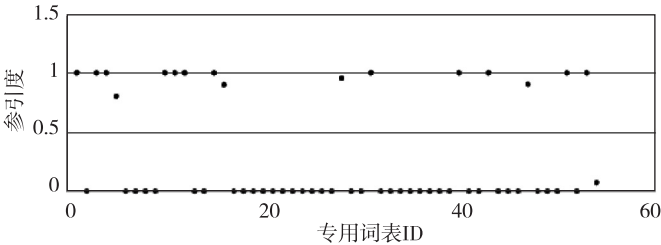


图 4 出版类专有词表术语的参引能力

4 采用的元数据信息

元数据是描述信息资源的特征和属性的结构化数据^[16],可以深入地揭示资源,便于资源发现。关联数据通过提供授权信息等元数据,能够确保数据自由共享,规范关联数据陈述。互联数据集词表(VoID),是一个表达 RDF 数据集元数据信息的词表,它提供了对整个数据集的描述,可以作为沟通数据发布机构和用户之间的桥梁^[17]。VoID 涵盖的元数据内容包括通用元数据、元数据存取、结构化元数据、数据集间的连接等信息。

4.1 数据集的 VoID 词表获取

每产生一个数据集,相应地就会产生一个 VoID 词表。用户通过其提供的元数据信息,可以

在短时间内定位到自己所需信息,实现高效检索。因此,数据集的 VoID 词表获取至关重要。由于数据集是一个包含多个 RDF 文档的集合,因此可以通过给定文档的 URI 来获得数据集的 VoID 描述^[18],具体方式包括:通过使用 thevoid:inDataset 属性将 RDF 文档反向链接(back-link)到 VoID 词表;通过在数据集的 URI 后添加/. well-known/void 属性来获取数据集的 VoID 描述。出版类中有 17 个数据集通过 VoID 文件来提供数据集的元数据信息,其中,6 个通过 back-link、3 个通过添加 well-known(知名信息)的方式来获取 VoID 词表,见表 7。

4.2 VoID 文件的内容

4.2.1 通用元数据

通用元数据是指从各个数据集中抽取的各类元数据的共性要素,具备通用可扩展的特征,如包括数据集的标题及描述、授权、主题等信息,可以帮助数据集的潜在用户决定是否使用该数据集来满足其检索需求,其通常遵循都柏林核心元数据标准。通常情况下,数据集的标题、描述等信息较为完备,但授权信息提供较少。出版类中,通过搜索三元组谓词部分包含“license”或“right”的字符串,发现仅有4个数据集提供授权信息,见表7。数据发布机构提供明确的授权信息,可以使用户明确使用条件,同时提供人类和机器可读的许可协议,允许数据的复制、传播、修改和再创作,减少版权问题,使数据可以更自由地共享^[19]。总体来

说,出版类中提供授权信息的数据集相对较少,应积极鼓励数据发布机构提供授权数据,为语义网的发展提供一个良好的知识共享平台。

4.2.2 元数据存取方式

VoID 词表定义了获取数据集 RDF 三元组的访问方式,包括 SPARQL 端点、RDF 数据转储等。RDF 数据转储是指当数据集的内容过大或需要很长时间压缩时,通过创建一个仅包含数据集元数据的转储文件对数据集进行备份^[17]。SPARQL 端点通过使用 void:sparqlEndpoint 属性来访问元数据。在转储方法中,通过使用 void:dataDump 属性将 RDF 转储文件与数据集建立关联。出版类有3个数据集通过 SPARQL 端点、1个数据集通过 RDF 数据转储的方式访问元数据,见表7。

表7 出版类中使用 VoID 词表提供的元数据信息

数据集	授权	反向链接	知名信息	SPARQL 端点	转储
http://semantic.ckan.net/about/data/					
http://aims.fao.org/aos/jita/					
http://lod.hebis.de/resource/					
http://worldcat.org		√			
http://wifo5-04.informatik.uni-mannheim.de/dws-group		√			
http://ieeevis.tw.rpi.edu/		√			
http://archiveshub.ac.uk					
http://bl.uk		√			
http://wifo5-04.informatik.uni-mannheim.de/gutendata		√			
http://aims.fao.org/aos/agrovoc/		√			
http://southampton.ac.uk	√				
http://rdf.muninn-project.org/ww1			√	√	√
wifo5-04.informatik.uni-mannheimde/dws-group			√	√	
wifo5-04.informatik.uni-mannheimde/gutendata			√	√	
http://l3s.de	√				
http://jesandco.org	√				
http://datahub.io	√				

5 讨论

LOD 在全球范围内实现语义网方面发挥着重要的作用,它促使人们发布用于连接和发现重要信息的数据集,并且将网络精简为一个单一的互联化的数据空间,最大程度地开放数据资产,

促进数据关联应用,挖掘数据的价值。由于目前我国对关联数据云图 LOD 的研究相对较少,本文通过从数据集归类、关联、词表使用、元数据信息等方面对 LOD 中出版类数据集的基本情况进行了细致的解释说明,以期为我国出版类数据的关

联数据化提供参考。然而,我们发现,LOD 中出版类数据集还并不完善,如数据集间关联度不高、许多数据集的内容无法开放获取等,这也是 LOD 云图中其他类数据集共同存在的问题。鉴于此,研究者还需要围绕此方面的问题做进一步讨论。

参考文献

- 1 Schmachtenberg M, et al. State of the LOD Cloud 2014[EB/OL]. [2015-06-18]. <http://linkeddatacatalog.dws.informatik.uni-mannheim.de/state/#toc0/>.
 - 2 German National Library. CATALOGUE OF THE GERMAN NATIONAL LIBRARY [EB/OL]. [2015-05-19]. <http://www.dnb.de/SharedDocs/Downloads/EN/DNB/service/link-edDataModellierungTiteldaten.pdf>.
 - 3 Selver Softic. COLINDA—Conference Link Data [EB/OL]. [2015-03-10]. <http://datahub.io/dataset/colinda>.
 - 4 Manchester Metropolitan University. Course reading lists [EB/OL]. [2015-04-16]. <http://lists.lib.mmu.ac.uk/index.html>.
 - 5 Library of Congress. Library of Congress Subject Headings [EB/OL]. [2015-05-11]. <http://datahub.io/dataset/lcsh>.
 - 6 Library of Congress. Library of Congress Online Catalog [EB/OL]. [2015-05-22]. <http://catalog.loc.gov/>.
 - 7 Mika P, et al. The Semantic Web-ISWC 2014 [J]. Lecture Notes in Computer Science, 2014, 8796:66-81.
 - 8 Bizer C, et al. Linked Data—The Story So Far [J]. International Journal on Semantic Web & Information Systems, 2009, 5(3):1-22.
 - 9 Rodriguez M A. A Graph Analysis of the Linked Data Cloud [J]. Corr, 2009(4):2-5.
 - 10 BibSonomy Developer Team. A blue social bookmark and publication sharing system [EB/OL]. [2015-05-30]. <http://www.bibsonomy.org/>.
 - 11 Gottron T, et al. Analysis of schema structures in the Linked Open Data graph based on unique subject URIs, pay-level domains, and vocabulary usage [J]. Distributed & Parallel Databases, 2015(4):515-520.
 - 12 Bizer C, et al. How to publish Linked Data on the Web [EB/OL]. [2015-04-08]. <http://www4.wiwi.fu-berlin.de/bizer/pub/Linked-DataTutorial/>.
 - 13 Ricci F, et al. Linking Search Results, Bibliographical Ontologies and Linked Open Data Resources [J]. Communication in Computer and Information Science, 2013, 390:62-65.
 - 14 肖琬,赵亮.中文元数据概论与实例[M].北京:北京图书馆出版社,2007:32.
 - 15 沈志宏,等. OpenCSDB:关联数据在科学数据中的应用研究 [J]. 中国图书馆学报, 2012(5).
 - 16 欧石燕. 面向关联数据的语义数字图书馆资源描述与组织框架设计与实现 [J]. 中国图书馆学报, 2012(6).
 - 17 W3C. Describing Linked Datasets with the VoID Vocabulary W3C Interest Group Note 03 March 2011 [EB/OL]. [2015-05-27]. <http://www.w3.org/TR/void/>.
 - 18 W3C. Cool URIs for the Semantic Web [EB/OL]. [2015-06-12]. <http://www.w3.org/TR/2008/NOTE-cooluris-20081203/>.
 - 19 W3C. Resource Description Framework (RDF): Concepts and Abstract Syntax [EB/OL]. [2015-05-29]. <http://www.w3.org/TR/2004/REC-rdf-concepts-20040210/>.
- (贾君枝 教授 山西大学经济与管理学院,寇蕾蕾 山西大学经济与管理学院情报学专业2014级硕士研究生)

收稿日期:2015-07-28