

词表生态系统:构成要素及关联关系*

贾君枝

摘 要 词表通过为数据添加明确的语义,推动了结构化数据网络的形成与发展。从数据生态、开放数据生态系统到关联数据生态系统的演化,以词表、信息技术、资源、人等要素共同构成的词表生态系统随之产生,同时推动着词表功能的扩大,词表价值的提升,以及关联数据生态系统的高质量发展。本文综合分析国内外研究文献,结合生态系统的理论与方法,对词表生态系统的基本构成进行详细分析,从各构成要素的功能角度讨论其对词表所造成的影响,认为高质量词表应从语义、语法、可获得性、互操作层面入手进行建设。在此基础上,深入剖析词表责任人、词表与词表、词表与数据集之间所产生的各种关系,展示以词表为核心所形成的纵横交织网络的特征,以了解词表数据集网络呈现的关联性,总体把握词表生命周期链条中各责任人彼此关联、环环相扣的态势,共同推动词表质量及价值的提升。图2。表1。参考文献15。

关键词 信息组织 生态系统 词表

分类号 G254

Vocabulary Ecosystem: Components and Correlations

JIA Junzhi

ABSTRACT

Vocabulary promotes the formation and development of structured data networks by adding explicit semantics to data. From the data ecology, open data ecosystem to the evolution of the associated data ecosystem, the vocabulary ecosystem composed of vocabularies, information technology, resources and human comes into being, which promotes the expansion of vocabularies function, the enhancement of vocabularies value and the high-quality development of the linked data ecosystem. The construction of vocabulary is affected by various factors such as related technologies, users, resource objects. The produce, release, use, maintenance and update of the vocabulary depend on each other and interact with each other.

The paper focus on the vocabulary development from the perspective of the ecosystem, instead of being restricted to the vocabulary itself, organically considers the various components of ecosystem and their relationships. To study the problem of vocabulary construction from the perspective of dynamic development and its interactive relationship is of great significance for the vocabulary to maintain long-term vitality and high quality.

The paper uses a systematic analysis method to analyze the vocabulary, information technology, resource,

* 本文系国家自然科学基金项目“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)

and personnel of the vocabulary ecosystem. And it explains the possible impact of each element on the vocabulary ecosystem in terms of quality, function, and evolution characteristics. The literature research method and data analysis method are used to sort out the types of horizontal and vertical interweaving relationships generated in the vocabulary ecosystem.

The paper considers that high-quality vocabularies should be constructed from the aspects of semantics, grammar, availability and interoperability. From the view of the vocabulary life cycle, the relationship between producers, providers, managers, users and other responsible persons is analyzed longitudinally, which aims to overall grasp the interrelationship of all responsible persons in the life cycle chain. The vocabulary providers build a bridge between the vocabulary producers and users. And the vocabulary managers become an intermediary between the vocabulary producers and providers. The feedback information of the vocabulary users is helpful to form a well-ordered and value-grading knowledge transfer and utilization system. The paper also analyzes the horizontal relationship between two vocabularies, and the vocabulary and the data set, and subdivides relationships into reuse, mapping, restriction, metadata, and co-occurrence between vocabularies. It is considered that various relationships occur at different stages of the vocabulary life cycle. The reuse, mapping, and restriction relationships occur mostly during the vocabulary construction stage, the metadata relationship arises during the vocabulary release stage, and the co-occurrence relationship occurs during the use of the vocabulary. The relationships between the vocabulary and the data set include reuse and creation. By reusing existing vocabularies or creating new vocabularies, the data set can have a clear semantic structure, which can be used in the future. 2 figs. 1 tab. 15 refs.

KEY WORDS

Information organization. Ecosystem. Vocabularies.

0 引言

词表定义了所聚焦领域的术语或者概念及关系,以实现描述对象的特征表述、类型划分及其条件限制。作为语义网的基本构成,词表既包含受控词表、名称规范档、语义网络、本体,还包括各种类型的元数据集。随着关联数据的飞速发展,词表为数据集描述提供了可共享的属性及属性值集合,通过为数据添加明确的语义,推动了结构化数据网络的形成与发展。早期,人们更多关注受控词表的建设。我国在 20 世纪 90 年代末出版了上百个叙词表,但仅有 13% 的词表在近十年有更新版本^[1],术语陈旧、开放度不高、机器可理解性差等问题与当前信息技术发展要求不相匹配,进一步阻碍了这些词表在网络环境下的拓展应用。可以看出,不

能单纯地将词表建设看成是一个独立问题,与词表相关联的技术、用户、资源对象所形成的生态系统在很大程度上影响着词表的发展。词表的生产、发布、使用、维护、更新各个环节彼此依赖、互相影响,共同推动着词表不断优化和完善,而一些词表却由于构建质量差、应用范围有限、开放程度低等原因被逐步淘汰。

从生态系统角度看待词表的发展,将不再囿于词表本身,有机地将生态系统的各种构成要素及其之间的联系进行综合考虑,从动态发展及其互动关联角度出发研究词表建设,对于词表保持长久的生命力及高质量具有一定的意义。本文综合分析国内外研究文献,结合生态系统的理论与方法,对词表生态系统的基本构成进行详细分析,从各构成要素的功能角度讨论其对词表所造成的影响,在此基础上深入剖析词表责任人、词表与词表、词表与数据集之间

所产生的各种关系,展示以词表为核心所形成的纵横交织的网络特征。

1 词表生态系统的构成

开放数据的发展,使任何人以任何目的都可以自由与免费使用、修改、再利用和重新发布数据。数据生态开始向开放数据生态系统演化。不同角色的数据处理主体参与到数据的产生、编辑、更新、消费使用流程中,不断推动新旧数据的产生与交替,这种动态循环、不断演化使得数据质量提升、数据价值激增。关联数据技术进一步确保了数据的可识别性且具有明确的语义特征,数据之间建立了丰富的链接而形成了数据网络,从而使以关联数据为核心的开放数据生态系统演化为关联数据生态系统,这为机器大规模处理数据及其数据的未来重用奠定了基础。词表作为知识组织工具,一直在资源描述与检索中发挥着作用。尤其在关联数据发布中,词表重用成为基本原则之一,使词表肩负着为数据添加语义的重要功能。开放数据集的增加带来了词表的第二个黄金增长期。新词表不断产生并得以迅速应用,愈来愈多的机构将已有词表发布为关联数据,许多词表注册平台出现,如 Bartoc、LOV、BioPortal、Heritage Data 等,旨在帮助用户准确地获取词表相关信息。2012 年 Baker 指出关联开放的词表集构成了词表生态系统,决定了关联数据生态系统的良好发展^[2]。词表生态系统不仅包含着词表本身,还包括与词表创建、维护、管理、应用整个生命周期发展息息相关的信息技术、资源环境与人,这些因素共同作用确保词表的功能得到充分发挥。

1.1 词表

词表既包含以分类表、叙词表、名称规范档为核心的术语集合,又包括以本体为核心的概念集合。受控词表作为值词表,用来提供属性值的取值来源;本体、元数据集常用来定义类与

属性,它们共同作用实现对资源对象的语义化表示,为数据增加语义,使机器能够准确地理解数据含义,实现跨部门、跨系统的数据重用及共享。通常一个高质量的词表应从语义、语法、可获得性、互操作等层面进行全面建设。

(1) 语义:词表的每一个概念都有清晰的定义,有面向人可读的标签、注释、概念类型(类或属性)等,定义概念类、属性之间的关系,明确各类型属性的定义域及值域,以便于词表使用者更好地辨识各个术语,更准确地参考引用。

(2) 语法:词表中的每个术语有统一资源定位符(URI)对其进行标识,对词表自己所定义的术语定义命名空间,便于其他词表或资源进行参考引用;词表采用 RDF、OWL 等形式化语言进行表述,以便于机器理解,同时提供 turtle、RDF/XML、RDFa 等多种格式供用户选择,以适应不同平台或系统需求。

(3) 可获得性:词表由相应机构定期维护管理,对概念及概念间关系适时动态地更新,并生成不同版本,以不断纳入新术语及关系类型;设计供用户检索浏览的平台,提供概念、类型、主题等多种途径的检索入口,帮助用户准确地查找到所需术语,提供数据可下载的页面,或者 API、SPARQL 端口,方便用户获取到相关术语定义;提供词表的元数据信息,涉及词表创建者、创建日期、权限、维护者、版本、语种、更新时间、链接地址等管理信息,还包含所属领域、类型、主题、词表间关系等描述信息及规模、格式、获取方式等技术信息。

(4) 互操作:词表建立与同一领域内其他词表或者与最常用词表之间的多个映射,创建词表术语之间的对照表,应用于词表互操作。

1.2 信息技术环境

信息技术环境提供词表设计、管理、发布、集成等所需的一系列信息技术集合。从词表的计算机编制及管理到词表的语义描述、发布到应用,都离不开信息技术的支撑。技术环境侧重于从早期词表管理技术向当前词表应用的技

术转移。尤其以 URI、RDF 为核心的语义网技术推动了词表向开放获取、形式化表示发展。词表应用的核心技术有语义标引、集成技术。语义标引运用词表的概念结构表征资源的特征及资源之间的关系,将词表中的概念赋予资源中的实体对象,通过语义标识从而使歧义词具有了明确的语义,并且使实体对象之间具有了明确的链接类型。集成技术是将不同词表或者资源之间建立概念及其关系的对应关系,以实现资源或词表的统一存取技术。每个词表表述范围有限,描述颗粒度有差异,词表的集成将使各个词表之间建立对应关系,以实现多个词表的无缝浏览。如国际虚拟规范档(Virtual International Authority File, VIAF)将全球范围内 30 个国家的 40 多个机构名称规范档的数据实现聚类、链接,为各国不同用户提供便利的存取服务^[3]。不同数据集的集成旨在将具有不同格式、结构、数据模型的资源实体对象建立对应关系,通常借助于词表的对应关系实现资源之间的连接,以实现资源聚类与检索,如 Europeana 作为欧盟数字文化遗产平台,收藏了来自 3 700 个图书馆、档案馆、博物馆等机构的 53 000 000 个单件,为实现资源整合,将来自于其他机构的元数据转换为 EDM 模型^[4]。

1.3 资源环境

从词表的内涵及外延看,其基本功能等同于知识组织系统^[5],学者常春指出知识组织生态系统中,知识组织系统对应的环境是文献信息^[6]。实际上词表应用环境由各类型资源集合构成,文献信息仅仅是其中一部分。资源环境是指词表用以描述标引的各类资源对象,通过词表的有效表示,可以将这些资源对象转换成机器可读可理解的数据集,以在检索、问答、数据挖掘、知识发现中发挥重要作用。由于资源集是词表应用的环境,词表对资源的描述能力通常借助于资源检索系统的检全率和检准率等指标进行评估。了解资源特征,选择适宜的词表,准确地理解词表中的概念及关系,注重描述

标引的规范性及一致性,成为资源层面运用词表的关键。从有形的物质载体如图书、期刊到数字化载体如电子书、Web 网页、数据,从文本到多媒体资源,从单篇文档到文档中的抽象概念,资源对象类型呈现多样化。据 LODSTATS 统计,当前描述资源的实体类集中在人、机构、概念、日期、地址、参考文献、图书、文件、项目等。资源描述对象的变化对词表提出了新的要求,会不断出现新型词表以适应此变化,如 DC 旨在描述网络文档而出现,FRBR 词表改变了传统的整体文献单元描述方式,转换为对作品、代理、主题等实体对象的描述。

1.4 人员

信息资源管理专家霍顿最早提出信息是有生命的,由一系列逻辑上相关联的阶段构成,经历“成长期、发展期、衰退期”这样的自然规律^[7]。词表作为知识概念体系的载体,同样经历了生产、发布、使用、维护、更新一系列处理阶段,通过不断循环往复,最终形成一个稳定系统。每个阶段都有词表相关人员参与,并发挥着不同作用,各个阶段对应的人员类型主要分为词表生产者、提供者、管理者及使用者。在词表整体建设过程中,不同责任人承担着不同职能。

词表生产者负责词表的设计及构建,当前已从早期的以专家为核心的手工构建转换为专家与语料库结合的半自动构建。词表的质量水平取决于词表生产过程,词表设计人员需要确认所关注领域及其词表的应用范围,明确用户关注的实体,从各种语料库抽取相应的实体及关系并由专家做进一步筛选,清晰地定义类、类间关系及属性、实例,采用形式化语言对其进行表述。在该阶段概念模型定义中,为提高词表的重用效率,减少不必要的重复,设计人员应制定词表重用的原则,处理协调好单个或多个词表、词表类或属性的选用、不同词表类间的融合关系。为提高词表间的互操作,应尽可能将自己定义的类与属性建立与已有词表的对应关

系。词表生产者提供了词表的出处信息,其权威性一定程度上影响到词表的使用范围。

词表提供者负责提供词表的存取服务,构建词表服务平台,便于用户发现、获取到词表的概念及关系。关联数据环境下需通过发布环节将其转换为关联词表供用户下载。词表提供者可以是词表的生产者,或者是其他机构,也可以是专门致力于多个词表的统一存取服务的第三方服务平台(包括数据仓储和词表注册平台)^[8]。词表提供者需提供关于词表的各种详细信息,包括词表的元数据信息、内容信息(概念类名、类间关系、属性定义域、值域、映射关系等)、多种格式显示信息、用例、词表应用领域及使用统计信息(使用的机构及数据集基本信息)。服务平台还需提供数据浏览查询功能,以迅速定位到所关注词表及词表内的类、属性,提供数据下载端口,供人和机器获取。

词表管理者负责词表的日常维护及更新。词表借助概念体系反映领域中的知识体系,而知识的更替处于动态中,词表中概念、关系的增删改也需要定期维护。不仅需要专家及时参与修订词表,讨论新版本的出版,还需要管理者从信息系统角度对词表进行实时更新。具体内容包括:从各种语料中及时捕获出现的新概念、新关系提供给专家参考;提供词表逻辑检验,以防止新定义的概念及关系与既有内容冲突;记录词表的变动信息,实时追踪并显示各个概念的变化情况;建立与用户的良好互动,及时获取用户使用词表的反馈信息,为词表的修订提供参考;收集、记录词表被不同机构应用的情况。

词表使用者指将词表用以描述标引、资源集成及浏览检索的机构或个人、机器。随着词表应用范围扩大,词表使用者类型也在增加,有词表生产者、信息处理机构、知识库生产者或知识图谱构建者、一般用户。词表生产者在复用已有词表的基础上形成自己的词表,包含完全或部分复用单个词表或多个词表两种情况。图书馆、博物馆、档案馆、出版机构、数据库商等信

息处理机构,结合自身所描述资源的特点,明确词表选择原则,从众多的词表中选择适合的词表用以对资源进行描述标引;将词表提供给检索用户,帮助用户构造检索式;将词表用来修订、扩充用户提问;向用户提供相关检索词推荐;通过分类、主题进一步过滤检索结果。知识库生产者或者知识图谱构建者运用词表进行实体识别及概念对齐,利用词表所定义的明确概念及概念关系,采用一定的语义匹配算法从资源中甄别出具体的人名、地名、机构名,并将这些实体与概念类建立对应关系,从而使资源中各类信息置于知识体系中,以便较好地实现语义理解。一般用户使用词表的目标不很明确,仅仅是浏览、检索词表基本信息。

2 词表生态系统中的关联关系

词表生态系统的各个构成要素之间彼此作用,相互关联,交织成纵横交错的网络,使词表的价值得到提升,不断适应外部技术环境及用户的需要。词表与词表、数据集之间产生横向联系,反映了词表外部环境给予词表的影响,体现了所形成的词表数据集网络的关联性。从词表生命周期角度纵向对不同责任人员之间的关系进行分析,以便整体把握生命周期链条中各责任人彼此关联、环环相扣的态势,共同致力于词表的建设。

2.1 词表责任人之间的关系

词表相关的责任人如生产者、提供者、管理者、使用者,共同推动着词表使用及发展,从而使词表生态系统成为一个动态系统,通过充分的信息共享、及时的信息反馈,形成闭环式信息管理,推动词表的质量提升及应用范围不断扩大。

词表生产者将知识概念体系化,并以人机可读可理解的形式展现,为后期不同责任者的使用提供了词表的基本信息,但词表质量、应用状况信息需要来自提供者、管理者及使用者的

各个反馈渠道。

词表提供者提供了词表信息可以共享的方式,实现了词表的可获取性,扩大了用户对词表的知晓度,并依赖于词表使用者的反馈信息不断提升其信息传递能力。词表提供者搭建了词表生产者与使用者之间的桥梁,可视化浏览、多途径检索、即时下载及其用户实时交互等技术发展使平台逐步成为专业化服务平台,用户能更准确及时地把握所关注领域的词表类型、词表内的概念及概念关系,并随时发表自己的意见。随着服务平台用户数的不断积累,丰富的用户评价、评注信息成为了词表评估、更新的重要依据。平台用户积聚效应提高其知名度,进一步吸引了更多的词表生产者主动将词表提供给专业化信息提供机构,平台的词表数量呈现出动态上升趋势,为词表实现更大范围的共享提供了基础。例如,Vandenbussche 发现 LOV 关联数据词表平台中词表集持续动态演化,词表数量从 2011 年的不到 100 个增加到 2015 年的 527 个,且各个词表修订日期不断更新^[9]。

词表管理者负责词表日常管理及完善更新,是确保词表可持续发展的重要机构,是连接词表生产者与词表提供者之间的中介。词表管理者不仅要保证词表描述的稳定性、一致性,还要监测统计词表在不同情境下的使用状况,及时收集来自用户方的使用反馈意见,从外部各种数据源及时捕获新术语,并将此信息反馈给词表生产者,为修订词表提供参考,将来自词表生产者的更新数据纳入词表管理系统,同步提供给词表提供者供其发布。

词表使用者是词表价值直接或间接实现的关键人员,词表使用者的基本信息及词表应用状况信息一定程度上反映了使用者行为特点及偏好,从各渠道收集其反馈信息,提供给词表的生产者、管理者,以作为词表修订、更新、维护及评估的依据。构建通畅的信息反馈渠道,引导各用户积极参与到词表生产与利用整个生命周期中,有助于词表生态系统信息传递方式从单向到双向转换,形成良好有序、价值逐级递增的

知识传递与利用系统。

2.2 词表间的关系

随着资源数量、类型的增加,各个机构为适应本地需求,纷纷制订适合自身发展的词表,使得词表的数量及种类呈现出快速增长势头。但各个词表之间并不是孤立存在的,它们彼此联系、相互制约,共同形成了词表网络,成为词表生态系统中重要的关联网络。词表间关系可细分为重用、映射、限制、元数据、共现关系。各种关系发生在词表生命周期的不同阶段,重用、映射、限制关系多发生在词表构建阶段,元数据关系产生于词表的发布阶段,共现关系产生于词表的使用阶段。

(1)重用关系。词表建设并不是一蹴而就的,而是一个个词表互为借鉴、相互利用的过程。这一过程中,词表通过共享术语、定义交流信息的共同模式而实现了词表重用,也构建了词表间的重用关系。词表重用的最大作用在于降低词表构建成本,提高词表之间的互操作性,避免表述的歧义及冲突,减少冗余。根据两个词表之间概念体系的重叠程度,可分为完全复用和不完全复用。生成不同语言版本的词表,属于完全重用,如 Agrovoc 是一个多语言的农业叙词表,包含中、英、法等 29 种语言。部分重用已有词表的概念及关系,属于不完全重用,如都柏林核心元数据 (DC) 已成为复用频率最高的词表,在其基础上形成了 PROV (起源表)、VOID (互联数据库词表)、DCT (数据目录词表) 等重要词表。通过 LOD (关联数据云图) 平台搜索发现,使用 dcterms:title 元素所描述的词表数达 132 个。Duval 指出没有哪一个元数据标准可以符合多个应用需求,从一个或多个元数据模式中选择元素组装形成复杂模式以适应需求^[10]。根据词表重用数量可分为单个词表重用和多个词表重用。以一个词表为核心构建属于单个词表重用,如元数据对象描述模式 (MODS) 是基于 MARC 标准而形成的。综合重用多个词表旨在利用多个词表的优势创建一个适合本地应用的

新词表,遵循元数据模块化原则,将词表中元素作为模块通过组装而形成,采用其重用词表的名称空间定义其元素。

(2)映射关系。在词表构建过程中,如果要定义自己所需的术语,则尽可能建立该术语与常用词表概念之间的对应关系,通过形成的概念网络以扩充该术语额外信息,以丰富其上下文环境,并确保术语可以被及时访问到。词表作为关联数据发布时,也需建立与其他词表的外部链接,以便形成有效的词表网络,供用

户实现词表间浏览访问。因此,映射关系实质是通过构建两个词表之间概念的对应关系,实现词表层面的数据集成,为跨系统的资源集成提供基础。如统一医学语言系统(UMLS)作为超级叙词表,确立了与213个来源词表的映射,提升了生物医疗信息系统和服务的互操作性。常用的映射类型有等同关系、等级关系和相关关系,不同映射关系采用相应的表示方法(见表1),当前采用SKOS术语形式化表示的方法较常见。

表1 映射关系类型及表示

映射关系	含 义	表示方法
等同关系	词表间两个术语含义相同或近似	skos: exactMatch、skos: closeMatch、owl: equivalentClass、owl: sameAs、owl: equivalentProperty
等级关系	词表间两个术语之间具有包含关系	skos: broadMatch、skos: narrowMatch、rdfs: subClassof、rdfs: subProperty
相关关系	词表间两个术语之间有一定的关联性,包含相互排斥	skos: relatedMatch、rdfs: seeAlso、owl: differentFrom

(3)限制关系。在词表构建中,仅罗列术语的名称及定义,不明确术语的取值范围,有可能导致词表使用的不一致,通过增加对术语的限制说明可确保一定的规范性。一个词表用来作为另一词表术语的限制说明,规定其一定的取值来源及数据格式等,则这两个词表之间形成了限制关系。通过进一步明确词表中的术语定义,有助于实现各个应用程序间的互操作。限制关系的确立能使两个或多个词表的联系紧密度增加,它们共同作用完成对资源的描述。通常元数据标准采用此方式来定义其元素的限制符,而以受控词表为主的编码模式经常作为起限制作用的词表。如都柏林核心元数据计划(DCMI)限制符用来限制元素及元素的取值^[11],其中dcterms:subject元素的取值主要来自某一特定分类表、叙词表等,可以将来自于同一类号或叙词的资源进行集中。

(4)元数据关系。词表构建完成后,将词表的基本信息展示给用户供其选择查看,需通

过词表的元数据实现。一个词表的术语用来表示另一个词表的元数据,包括管理、内容及技术等信息,则这两个词表之间建立了元数据关系。同样Vandenbussche认为元数据关系是一个词表的术语被重用以提供另一个词表的元数据^[9],元数据旨在帮助用户查找、获取、定位到词表,元数据记录生成时,通常需选择一个或多个特定词表的元素对词表的信息进行描述。常用来做元数据词表的有通用类元数据集,如DC、DCT、PROV;表示特定数据集的元数据集,如VOID、DCAT、SCHEMA;表示本体的元数据集,有OMV等^[12]。经统计,LOV(关联数据词表)中2014年词表的元数据关系链接数达2637个,随着词表的增加现已达到3574个,对其元数据关系的出入度进行计算(见图1和图2),平均每个词表的元数据关系链接数达到5个,其中各词表使用的元数据词表排名在前的有DC、知识共享CC协议、FOAF、SKOS等词表。

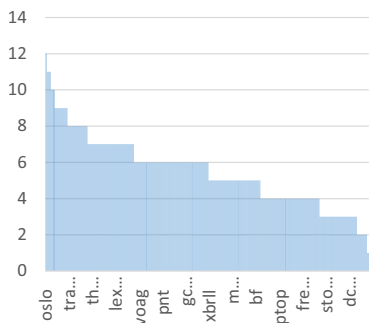


图1 LOV 中元数据出度统计

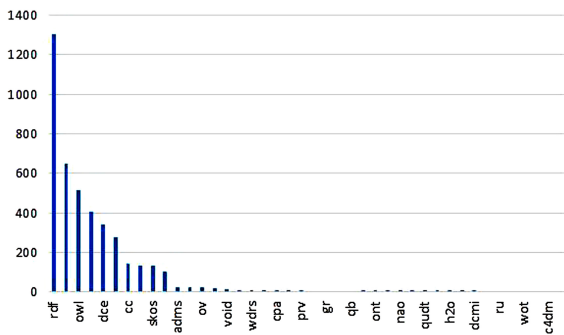


图2 LOV 中元数据入度统计

(5) 共现关系。在词表应用阶段,构建词表或描述数据集时,不同机构根据资源特征、词表选用偏好等重用不同的词表。共现关系指共同描述同一或多个数据集或词表时的词表间关系,即词表以描述对象数据集或词表为中介而建立起的词表间关系,是由词表使用者行为而形成的词表关系。与其他关系相比,共现关系是一种被动关系,并不由词表生产者、提供者决定。同一数据集的词表共现体现了词表间的互补性,不同数据集的词表共现体现了词表所覆盖的应用范围,分析词表共现情况一定程度上可以反映词表间的关联强度。词表间共现频率越高,则这些词表关联度越高,体现了词表间具有较好的相互补充功能。

2.3 词表与数据集关系

数据集是单个程序发布、维护或聚合的一组数据,通常以 RDF 三元组形式表示,因此数据集可以被视为某些资源的一组描述,这些资源通常共享一个公共 URI 前缀。其描述了一组明确定义的资源,使其成为具有特定语义含义的结构化数据,包含词表层面的数据集和资源层面的数据集。词表层数据集将词表表示成具有清晰语义和语法结构的数据,以适应语义应用。资源层数据集是词表应用的主要环境,发挥词表的解释、标注、扩展数据集语义的功能,通过抽取数据集中的实体对象,将其与词表中的概念类建立对应,依据词表定义的属性对实体对

象属性进行赋值,将实体对象之间的关系映射到词表的概念间关系上。创建及发布数据集时,有两种使用词表的方式:重用已有词表,创建新词表。为节省资源、提高资源互操作能力,确保数据集被广泛使用^[13],尽可能采用重用方式。将所使用的词汇表中的术语置于数据集三元组的谓词位置或者 rdf:type 的对象位置^[14],以表示资源的各种属性特征,这样将形成资源与词表相链接的数据集网络。随着大量资源数据集的不断发布,所采纳的词表数量也在增加,词表层数据集也在不断增加,导致该网络处于动态扩展中,其关联强度不断上升。据统计,2014 年 LOD(关联数据云图)收集了 1 014 个数据集,使用词表数达 649 个^[15],2019 年数据集有 1 239 个,所使用的词表数也在增加。通常一些具有高入度的词表成为核心词表,它们具有高流行度,在数据集中发挥着连接中心的作用。

3 总结与展望

随着词表的开放及应用范围扩大,愈来愈受人员、技术、资源等各种因素的影响。从词表生态系统角度看待词表建设问题,应注重从生产、发布、使用、维护、更新一系列处理阶段的词表责任人出发,揭示词表生产者、提供者、管理者、使用者不同的职能及其彼此之间的相互制约关系。从词表本身及资源、技术环境出发,明确词表发展的特点及外部环境因素所产生的影

响,尤其从词表间、词表与数据集所形成的关联网络可以看出,词表已成为重要的语义黏合剂,发挥着语义描述与标注功能。通过不断从外界获得反馈信息,词表能够持续地对其内容及结构进行优化,通过有效的信息发布环节,使用户

及时地发现并获取,从而扩大其应用范围,实现价值提升。如何进一步提升词表的开放度,促使不同词表责任人积极参与到词表建设中并实现互动,如何构建有效的信息反馈机制,是未来词表建设需要考虑的问题。

参考文献

- [1] 吴雯娜. 我国叙词表的编制历史与发展模式[J]. 情报理论与实践, 2018, 41(6): 39-44. (Wen Wenna. Construction history and development modes of Chinese thesauri[J]. Information Studies: Theory & Application, 2018, 41(6): 39-44.)
- [2] Baker T, Vandenbussche P Y, Vatan B. Requirements for vocabulary preservation and governance[J]. Library Hi Tech, 2013, 31(4): 657-668.
- [3] VIAF[EB/OL]. [2009-12-10]. <https://www.oclc.org/en/viaf.html>.
- [4] Sharing the beautiful thing - annual report and accounts 2015 [EB/OL]. [2020-01-01]. http://pro.europeana.eu/files/Europeana_Professional/Publications/europeana-annual-report-accounts-2015.pdf.
- [5] 贾君枝. 面向数据网络的信息组织演变发展[J]. 中国图书馆学报, 2019, 45(5): 51-60. (Jia Junzhi. Development of information organization for the data web[J]. Journal of Library Science in China, 2019, 45(5): 51-60.)
- [6] 常春. 面向叙词表构建的知识组织生态系统研究[J]. 图书情报工作, 2016, 60(15): 101-107. (Chang Chun. Research on knowledge organization ecosystem for the construction of thesauri[J]. Library and Information Service, 2016, 60(15): 101-107.)
- [7] Horton F W. Information resources management (IRM): where did it come from and where is it going[C]//Proceedings of the ACM'81 Conference. 1981: 277-278.
- [8] Zeng M L, Mayr P. Knowledge organization systems (KOS) in the semantic web: a multi-dimensional review[J]. International Journal on Digital Libraries, 2019, 20(3): 209-230.
- [9] Vandenbussche P, Atemezing G A, Poveda-Villalón M, et al. Linked open vocabularies (LOV): a gateway to reusable semantic vocabularies on the web[J]. Semantic Web, 2016, 8(3): 437-452.
- [10] Duval E, Hodgins W, Sutton S, et al. Metadata principles and practicalities[J]. D-Lib Magazine, 2002, 8(4): 1-16.
- [11] DCMI qualifiers[EB/OL]. [2020-01-01]. <https://www.dublincore.org/specifications/dublin-core/dcmes-qualifiers/>.
- [12] Dutta B, Toulet A, Emonet V, et al. New generation metadata vocabulary for ontology description and publication [C]//Metadata and Semantics Research: 11th International Conference (MTSR 2017). Tallinn, Estonia, 2017: 173-185.
- [13] 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: 11th International Conference. Anissaras, Crete, Greece, 2014: 457-472.
- [14] Binding C, Tudhope D. Improving interoperability using vocabulary linked data[J]. International Journal on Digital Libraries, 2016, 17(1): 5-21.
- [15] State of the LOD cloud 2014[EB/OL]. [2019-12-18]. <http://linkeddatacatalog.dws.informatik.uni-mannheim.de/state/>.

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: 2020-02-28; 修回日期: 2020-03-26)