

面向数据网络的信息组织演变发展^{*}

贾君枝

摘 要 数据与数据之间通过富含语义链接的方式形成了有价值的数据网络,在以语义网为核心的网络架构中,网络资源的语义化结构、细粒度描述、数据关联等特性对数据加工及利用提出了新的挑战,信息组织的处理对象及外延发生着变化,信息组织工具不断丰富且功能趋于统一。本文依据信息组织的不同功能和处理流程可以将其细分为资源层、元数据记录层、词表层,分别对各个层次的特点、变化及其未来发展进行分析,认为从文本中准确提取出反映语义含义的特征项是当前信息组织的关键问题,元数据集、知识组织系统成为词表的主要构成,知识组织系统呈现出描述对象概念化、语义关系多样化、描述语言形式化、数据开放关联性四大特点。图3。参考文献23。

关键词 信息组织 数据网络 知识组织系统 词表

分类号 G254

Development of Information Organization for the Data Web

JIA Junzhi

ABSTRACT

The network is regarded as the global data space in the semantic web architecture. The form of resource organization is transformed from file-centered to data-centered. The valuable data network is formed between data and data employing rich semantic links. The semantic structure, fine-grained description and data correlation of network resources pose new challenges and possibilities for data processing and utilization. With the development of Semantic web technology and the change of user demand for information, the objects and extension of information organization processing are constantly changing from document unit-centered to knowledge unit-centered. And information organization tools are constantly enriched and tend to be functionally unified and extend from controlled vocabularies such as classification schemes and thesaurus to knowledge organization systems such as semantic network and ontology to meet the needs of large-scale, diversified and semantic information processing. In the process of development and evolution, how to abstractly summarize the process flow of information organization, and whether the roles played by different levels have changed? What are the essential links between different types of tools? What are the characteristics of the development of information organization tools? This paper focuses on those core issues and has a study using the method of literature analysis and system theory. According to the different

^{*} 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)

functions and the process flow, information organization can be subdivided into resource layer, metadata record layer, and vocabulary layer. This paper makes a detailed and in-depth analysis of the characteristics, changes and future development of each layer. The resource layer is the resource objects processed and can be divided into unstructured, semi-structured or structured objects. The metadata record layer is formed after resource objects being described and indexed. Metadata set and knowledge organization system constitute the vocabulary layer, which is the schema of the data network, specifies the basic mode of information description, and clarifies the selected description elements and their corresponding value types and ranges. The analysis results show that the key problem of information organization is to extract the feature items which reflect the semantic meaning from the text accurately. Knowledge organization system aims at expanding the functions of information organization tools, expanding library resources to network resources, deepening the description and representation of resources from original document units to knowledge units, changing from human-oriented construction to machine understanding, and gradually approaching user needs with finer descriptions. The knowledge organization system has four characteristics: conceptualization of descriptive objects, diversification of semantic relations, formalization of descriptive languages and open relevance of data. This paper systematically describes the process flow of information organization and clarifies the characteristics and future development requirements of resource layer, metadata record layer and vocabulary layer in the current stage of development. The deficiency is that there is no discussion on non-information resource objects and information resource objects respectively. 3 figs. 23 refs.

KEY WORDS

Information organization. Data web. Knowledge organization system. Vocabularies.

0 引言

互联网经历了 Web1.0、Web2.0, 进入到当前 Web3.0, 即称之为数据网络时代, 在以语义网为核心的网络架构中, 网络被看作是全局数据空间, 资源组织形式从以文件为核心向以数据为中心转换, 数据与数据之间通过富含语义链接的方式形成了有价值的信息网络。网络资源的语义化结构、细粒度描述、数据关联等特性对数据加工及利用提出了新的挑战及可能。信息组织是信息有序化过程, 即通过一定的方法与工具对所处理的信息对象进行描述、标引、生成元数据的过程。元数据作为信息组织的最终结果表现形式, 成为数字化管理对象的重要构成。随着语义网技术发展及其用户信息需求的变化, 信息组织的处理对象及外延不断发生着变化, 相应地从以文献单元为核心的处理向以知识单元为核

心的处理转换, 信息组织工具不断丰富且功能趋于统一, 从以分类表、主题词表为核心的受控词表扩展到语义网络、本体等知识组织系统, 以适应大规模、多样化、语义化的信息处理需求。在这个发展演变过程中, 如何抽象地概括表示信息组织处理流程, 各层次所发挥的作用是否发生了变化? 其不同类型的工具之间的本质联系是什么? 信息组织工具发展的特点表现是什么? 这些都成为信息组织研究者及使用者关注的问题。

信息组织的理论与方法随着信息技术的发展不断演变, 出现了各种类型的新术语, 需要在学科研究领域形成对这些专业术语的统一认识, 本文旨在梳理信息组织的发展变化过程, 抽象概括出信息组织的处理功能层次, 依此对每一层次的特点及发展变化规律进行分析, 寻求不同阶段对术语的准确解释, 旨在预测未来信息组织的发展方向, 以推动国内外信息组织理论与实践的发展。

1 信息组织处理功能层次

信息组织处理活动可以理解为对资源对象

按照一定的描述规则及语言进行描述标引的过程。依据不同功能、处理流程可以将其细分为三层:资源层、元数据记录层、词表层(见图1)。

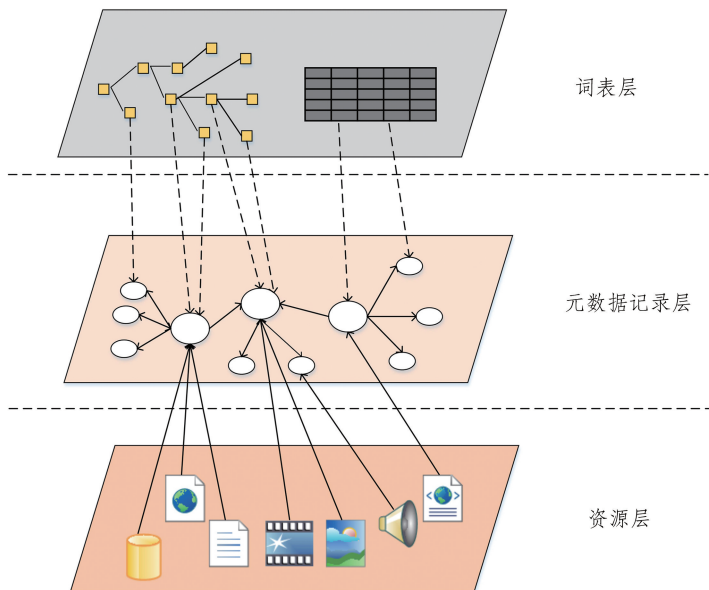


图1 信息组织处理功能层次

资源层实际为所处理的资源对象,可以分为非结构化、半结构化或结构化对象三种。非结构化资源是机器不可读的,以纸质文献或者多媒体资源为核心;半结构化资源拥有一定的格式,如 pdf、word、html 文档,但机器不能准确地理解到文本内容;结构化资源采用一定的数据模型表示高度结构化资源,如关系数据库、RDF 数据集等,结构化程度越高,机器越易于获取并表示资源的语义内容。

获取到资源对象后,需对资源对象进行信息描述和信息标引,则进入到元数据记录层。该层是信息组织的最终结果形式,也是信息检索效率提升的关键。元数据记录的生成需要依托于词表层,词表层属于最上层,是数据网络中的模式层^[1],规定了信息描述所使用的基本模式,明确了所选用的描述元素及其相应的取值类型及范围。万维网联盟将词表定义为语义网

环境下描述和表示关注领域的概念及关系,以实现描述对象的特征表述、类型划分及其条件限制^[2]。词表包括受控词表、本体、元数据集(DC、RDA、FOAF、SKOS)、语义网络(WordNet、FrameNet、HowNet)等,实际上可分为元数据集、知识组织系统两大部分,元数据集又称之为属性词表,知识组织系统称之为值词表。依据不同的应用需求及其所描述资源对象、标引人员偏好等差异,所使用的词表具有较大差异性,从而导致不同机构的元数据记录之间存在着差异,一定程度上增加了数据交互的困难。ISO25964-2 所制订的叙词表与其他词表映射的标准^[3],旨在解决不同词表之间的互操作问题。但随着语义网技术的引入,通过词表最大程度重用、构建丰富的元数据集间关系、描述语言 RDF 化等多种措施可以提升元数据记录及词表间的互操作性。

2 资源层文本的构成

资源层是信息组织的处理对象,其通常记录在一定载体上,包含文本和多媒体信息。文本作为知识交流的主要方式,成为信息组织的重要研究对象。文本主要由内容、结构组成。内容表述信息的语义含义,是文本的核心部分,也是获取语义信息的重要来源。文本中语义选用一定语言体系中的词语及句法构成表达,词语作为概念的外在表达形式,概念具有清晰的语义含义,但词语的含义取决于上下文语境,与概念并不存在一一对应关系,从句法结构中有效地识别词语,并建立与概念之间的对应关系成为获取语义的关键途径。结构用以支持语义内容表述,体现为格式定义、文本编码等,提供外部呈现形式,如屏幕显示、打印输出。纸质文本与电子文本具有一定的差异性,纸质文本的结构和语义不可分离,互为支撑,以帮助人更好地理解文本的内容。电子文本随着结构化程度的增加,结构与内容逐步分离,相互影响度减小,相同的内容可以采用不同的编码方式表现,这样为机器准确获取到文本内容提供了便利。

如何从文本中准确提取出反映语义含义的特征项是当前信息组织的关键问题。分词及句法结构分析为文本理解提供了一定的可能,如何利用本体、机器学习技术等提高语义理解的准确性已成为当前研究的热点。随着资源数量的增加,用户不仅限于获得整个文档,可能更关注文本中的具体内容,如方法、观点、理论、数据等,这就对信息组织提出了更高要求,使信息处理对象从面向文献单元向知识单元(概念实体)转变,通过文本分析技术获取文本中所涉及的知识元,建立知识元间的关联,采用 URI(唯一资源定位符)方式标识这些非信息资源对象,确保用户能够获取到以主题、人物、事件等为中心的语义单元。

3 元数据记录及元数据集的发展

元数据记录是关于资源对象的结构化数据,定义了所描述对象的内外属性特征,用于描述、管理、知识产权声明等多种目的,既包含描述类元数据,如格式、文本长度、类型、出版信息,又包括表述内容的语义元数据,如主题词、分类号。随着文本数量的增加,元数据记录的价值在不断提升,其一定程度上能够作为文本的替代品,实现对文本的快速查询与管理。描述语言决定了元数据记录输出格式,其从 MARC 格式经历了 XML、RDF/OWL 的发展过程,进一步向标准化、开放性、多样化发展。

元数据记录生成离不开元数据标准,为提高信息处理标准化与数据交换效率,元数据标准制订成为必要,其旨在构建一个对数据语义内容达成共同理解的模式,从而在元数据标准下生成各种类型的元数据集,定义元数据间的关系及其语义、语法和可选值等^[4]。在数字化信息处理的不同阶段,随着数据描述、表示及存储技术的发展,信息处理所遵循的元数据标准从描述类元数据集开始向内容类元数据集转变。描述类元数据用于以文献单元为主的文本资源对象的处理,旨在将每一篇文献处理为一个数据对象,关注的是如何将非结构化数据转换为结构化数据,此过程使用的元数据集有 MARC、DC、EAD,以描述资源的外部属性及其格式为主,作者、题名、出版机构、刊名、卷期、出版时间、出版地点等外部属性显性度高,获取容易,易于操作,成为该阶段元数据记录描述的重点。

进入到内容处理阶段,以文献中的知识单元为主,考虑将每一篇文献解析为多个概念实体,注重概念之间的关系表示,关注的是知识网络的形成与表示,旨在将文献中的关键内容呈现给用户,并能帮助用户实现知识单元之间的无缝浏览。此阶段出现了内容型元数据集,重视资源的语义内容特征提取,如 RDA、DCMI 抽

象模型、DACS、CCO 等,这些元数据集多以概念类为核心,注重概念之间的关系表示,采用 RDF 语法表示其数据模型,形成了 RDF 词汇表,并注册到词表目录平台 LOV、JoinUp 供用户浏览下载。可以看出,这一阶段元数据集与知识组织系统功能趋同,两者之间界限模糊,一定意义上统称为词表。尤其在对资源对象进行描述时,

一个元数据集不足以完成全部描述内容,需要元数据集与知识组织系统等多个词表互相补充才能完成描述。关联开放数据(LOD)云图中,每个数据集作为描述对象,通常采用多个词表描述,图2显示了出版类数据集的词表数量,经统计每个数据集平均使用词表数量7个^[5]。

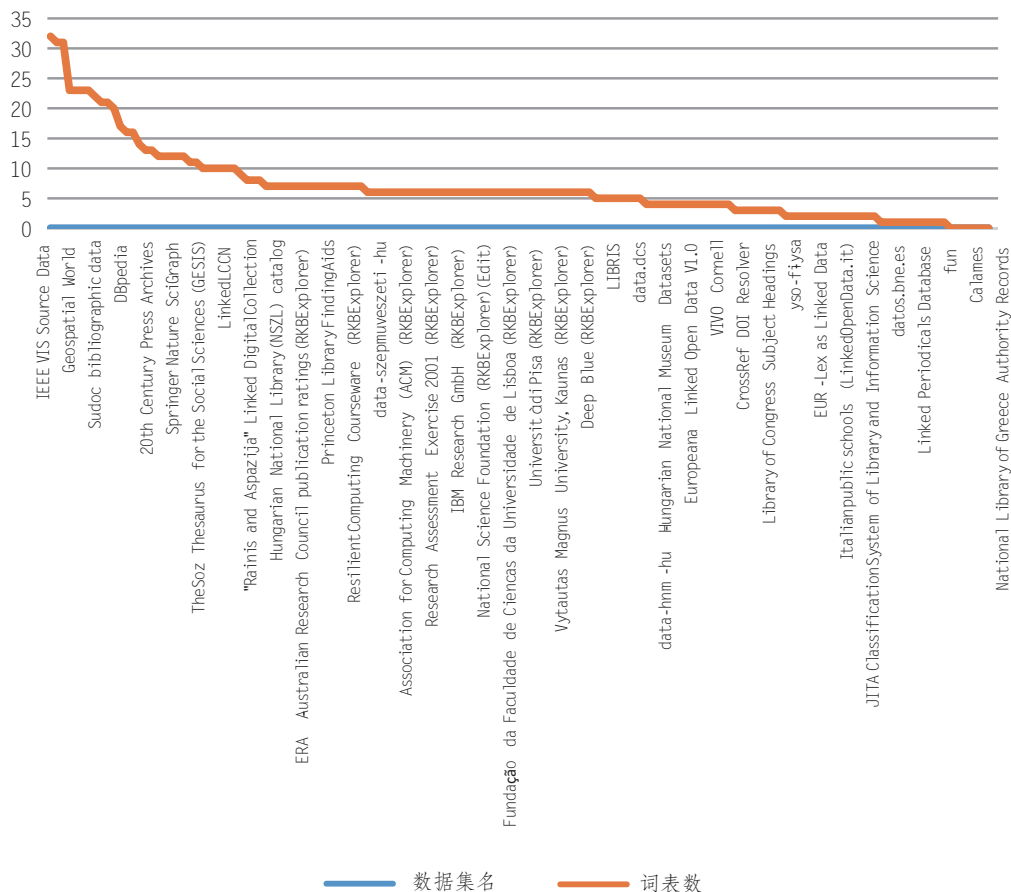


图2 出版类数据集的词表数量

4 知识组织系统的类型、功能与特点

在1998年召开的第三届ACM数字图书馆国际会议上,Hodge首次提出了知识组织系统的概念^[6],自此,从以分类表、叙词表为中心的传统信息组织工具发展成为包括术语表、地名表、

人名表、分类表、叙词表、本体等的知识组织系统。知识组织系统是用来组织信息的各种类型模式集合,其将领域内部的知识进行系统化表示,旨在将描述对象中的自然语言表达形式转换成知识组织系统所规定的表达形式,既实现对术语的统一化表述,又充分展示了术语之间的语义关系,从而支持对资源内容的检索与浏

览。知识组织系统是在数据网络发展背景下出现的,旨在拓展信息组织工具的功能,从以馆藏资源为中心向网络资源拓展,资源的描述及表示由原有的文献单元向知识单元深化,从面向人构建向注重机器理解转变,描述颗粒度变细且逐步向用户需求靠近。

4.1 知识组织系统的类型

学者对知识组织系统的划分依据不同,提出了多种类型,可以从不同的维度探究信息组织工具的差异。Hodge 根据知识组织系统的结构、复杂性、术语间关系及历史功能,将其分为三种类型:术语表、受控词表、关系表^[6]。术语表重视术语的定义及罗列,包括名称规范档、词典、字典、地名表;受控词表强调应用规范性词汇表示资源的主题内容,包括分类表、主题词表;关系表关注术语间的关系定义,包括叙词表、语义网络(如 WordNet)、本体。曾蕾教授根据受控程度及语义结构强度将知识组织系统划分为四种类型:术语表、元数据模型、分类表、关系模型^[7],将其 Hodge 提出的术语表作了进一步细分,将同义词表列入术语表,名称规范档、名录、地名表列入元数据模型。可以看出,各个知识组织工具之间既有联系,又有区别,是在不同时期服务于不同的应用目的,具有不同应用功能。术语表作为基础词表,可以为其他词表提供有用的数据来源,关系模型的形成需借助术语表、受控词表的丰富术语及其可以利用的关系类型。当前环境下这些工具功能逐步走向趋同,关系模型成为各个工具转换的目标。

4.2 知识组织系统的主要功能

知识组织系统不再仅仅服务于信息检索过程,提高信息检索效率,还表现为增强知识管理能力。

(1)作为组织与管理资源文档的工具。网络环境下,资源数量增加导致资源获取难度增加,知识组织系统可以有效地实现对资源文档的表示及索引,通过构建特定领域的潜在语义

结构模型,支持文档的快速检索。

(2)形成特定领域或学科的语义知识地图。知识组织系统充分利用专家知识,描述了特定领域或学科的基本概念及概念间关系,形成了概念语义地图,有助于支持信息检索者(机器及用户)的知识获取及学习过程。

(3)作为知识交流及互操作的工具。知识组织系统所提供的概念框架,架起了用户信息需求与文档集、自然语言与受控词汇之间的桥梁,较好地推动了计算机与计算机、人与计算机的交流过程,通过建立知识组织系统之间的映射与连接,可以实现跨语种、跨库、跨系统的操作,满足用户多角度的检索需求。

4.3 知识组织系统的发展特点

(1)描述对象概念化

从术语表到语义网络、本体,知识组织系统的描述对象从原有的以词汇为中心向以概念为中心转变。概念是人类的思维单元,其由代表一定语言的词汇或符号来表示,可以使用自然语言或者人工语言表示。由于语言表述的丰富性及自由性,不同的语言词汇可指代同一个概念^[7],同一词汇在不同的语境中含义不同,从而导致词汇与概念不具有一一对应关系。指代物、概念、术语三者构成了对事物的基本认识。概念反映了一定的指示物,术语用以表述指代物的符号系统。术语表仅罗列现实存在的各种术语,通过释义、注释等方式对词汇含义进行区分,并未从概念角度对词汇进行分类,无法体现出词汇在概念层面的抽象。叙词表引入了规范词,将指代同一概念的多个词建立用代关系,并建立了概念之间的属分关系、交叉及并列关系,较大程度上体现了以概念为核心的特征。本体则以概念为出发点,所构建的共享概念模型强调知识表达的明确性、形式化,其所定义的函数、关系、公理充分将概念、属性及其概念之间的关系进行表示,从而为机器可理解及其信息共享提供基础。以概念为核心的知识组织系统描述相关资源,意味着不同资源共享共同的概

念体系,计算机非常易于抽取及集成相关信息,一定程度上提高了资源间数据交流及集成的效率。

(2) 语义关系多样化

知识组织系统是由代表知识领域的概念及其概念关系构成。概念之间彼此依赖、互相联接,所形成的关系称之为语义关系,这种关系通常由概念之间的外延所决定,而概念的含义可以通过语义关系进一步明确。知识组织系统中,语义关系从等同、等级、相关三种关系发展为针对不同领域的多种关系,如因果、来源、代理等。语义关系越丰富,定义越严格,所形成的概念语义网络才能为知识组织系统的进一步应用奠定坚实基础,更好地适应其他数据模型的需求。Hjørland 总结了概念及概念关系获取有五种方法:经验主义、理性主义、解释学、实用主义、批判论^[8]。在现有的知识组织系统构建中,不同的知识组织系统概念获取方式侧重点不一。分众分类法以批判论为核心,强调个体有使用词汇及概念的自由,通过众多用户对资源设置标签的方式,构建开放式的知识组织系统。受控词表(分类表、叙词表)的构建将经验主义、理性主义、实用主义三者结合,在明确词表构建目标前提下,既重视专家的知识,又注重概念在文档(包含检索集)中的使用频率。本体更强调理性主义,通过分析及逻辑方法,抽取领域中核心的概念及关系,基于共同认知框架形成共享化模型,以实现领域内部概念集合的抽象描述。

在知识组织系统中,从相容及不相容两种概念间关系可推演出三种基本的语义关系:等同关系、等级关系、相关关系。等同关系又可细分为同义、同源、反义。同义关系表示两个词汇都指代同一概念,如简称与全称、变异名称、不同语种形式等都属于此类型;同源关系表示随着时间变化而采用不同的术语形式,如机构不同时期的称谓,同义词表、名称规范档、语义网络充分体现了同义关系及同源关系;反义关系表示两个概念之间互相排斥,外延没有重合,如

“生—死”,叙词表中叙词与非叙词关系中包含反义关系。等级关系表示两个概念之间存在包含关系,是知识组织系统中最基本的类型,分为上下位关系、部分整体关系、实例关系。上下位关系中,下位类既具有与上位类同样的属性,又有自己特有的属性,两者属于梯阶层级关系;部分整体关系表示整体概念是由若干部分概念所组成,如元素—集合,部门—机构,材料—物体等;实例关系表示通过罗列其主要元素定义概念,各主要元素作为个体概念,被看作是实例。如摩托车—大阳摩托车。等级关系在受控词表、本体中是最突出的语义关系,分类表通过类目层层划分表示其类目的等级关系,叙词表等级关系对其上下位关系、部分整体关系、实例关系不做区分,但本体对这三种关系类型进行了明确区分,在继承、传递、对称性方面体现为不同的特性。相关关系指除等级、等同关系之外的其他关系,语义关系指代不明确,包括参照关系(see also)及其他。参照关系将存在一定联系的两个概念建立关系,如产品与厂家、作品与作者等;叙词表定义了相关关系,本体的严格性限制了对此类关系的引用,其根据不同的情境定义其他更为具体的语义关系,如疾病本体中会定义医患关系、师生关系、疾病与治疗等关系类型。

(3) 描述语言形式化

知识组织系统的形式化描述推动了语义网的发展,为语义网提供了丰富的知识基础设施。早期信息组织工具为纸质的分类表、主题词表,随之将其转换为计算机可读形式,便于数字化存储及管理,但已有格式并不能使机器很好地理解及标引资源,本体的形式化表述推动知识组织系统整体向此方向演进。形式化描述语言有简单知识组织系统(SKOS)、RDF/RDFS、OWL语言等,其对各个知识组织系统的类、属性及其关系进行明确界定,以增强其语义表述能力,拓展其在数据网络环境下的应用范围,为进一步实现各个系统之间的交互,为外部资源及应用软件集成提供便利。

SKOS 强调为叙词表、分类表、标题词表等受控词表的表达提供强有力且简单的机器理解方式^[9],其提供了表达概念模型的基本结构及内容,为已有受控词表导入数据网络提供了一条标准的、低成本迁移路线。SKOS 概念体系是多个 SKOS 概念及概念间语义关系的集合体。每个 SKOS 概念体系可以简单地对应为一个叙词表、分类表或其他知识组织系统。词表中每一词汇都可看作是 SKOS 概念类,词汇之间的关系采用 SKOS 属性进行定义,除此之外,SKOS 定义了词表之间的映射关系,依此将多个词表之间的词汇建立对应关系,为词表互操作的实现提供了形式化描述方法。当前许多受控词表纷纷转化为 SKOS 形式,SKOS 也成为数据集描述中频繁使用的词表之一,依据关联开放词表 (LOV) 平台统计,目前有 152 个数据集使用 SKOS 词表^[10]。2009 年美国国会图书馆将国会标题表 (LCSH) 的 MARC 数据转换成 SKOS^[11],联合国粮农组织将 AGROVOC 转换成 SKOS 格式并提供免费下载^[12],国际十进分类法 (UDC)、杜威十进分类法 (DDC) 分别实现了 SKOS 化^[13],我国学者也着手探讨《中国图书馆分类表》《汉语主题词表》《中国分类主题词表》的 SKOS 实现^[14-16]。

资源描述框架 (RDF) 旨在提供一种用于表达语义信息,并使其能在应用程序间交换而不丧失语义的通用框架。使用资源—属性—值的三元组形式描述资源的元数据信息,如题名、作者、修改日期、版权、许可信息、多个共享资源的模式表等^[17]。这种最小限制、灵活的描述方式,已被普遍应用于词表及数据集的描述中,成为实现数据整合、提升资源价值的重要描述语言。1999 年 DC 提出 RDF 模型,美国国会图书馆已提供 LCSH 的 RDF 形式下载^[10]。

RDF 表达能力终究有限,出现了扩展 RDF schema 及其 OWL 语言,添加了更多的用于描述属性和类的词汇,如增加了类之间的不相交性、基数限制、等价性、丰富的属性特征、枚举类,通过提供更多具有形式语义的词汇,使 Web 信息

拥有确切的含义,可被计算机理解并处理,因此 OWL 成为本体描述的重要语言。WordNet、FOAF 拓展其语义网应用,采用 RDF/OWL 表述^[18,19],美国国家癌症研究所发布了 OWL 版本的 NCI 叙词表 (National Cancer Institute Thesaurus)^[20]。

(4) 数据开放关联性

数据开放关联理念在大数据的应用环境下得到充分体现,数据网络的形成基于 RDF 数据模型实现不同数据源实体之间的链接。数据集彼此关联为数据深层次开发与利用提供了保障。随着越来越多的机构参与到数据开放中,结构化数据之间的关联而形成数据网络规模不断扩大,不仅使用户能够无缝浏览到各种级别的数据,而且为创建跨领域、跨部门的数据集成应用提供了可能。知识组织系统中概念体系开放意味着可以被越来越多的资源所应用,未来所产生的价值将更大。万维网联盟图书馆关联数据小组 (W3C Library Linked Data Incubator Group) 旨在推动图书馆数据的全球性互操作,以提高图书馆数据在互联网环境下的重用性^[21]。许多受控词表、名称规范档走向开放关联,2009 年 DDC 分类法首次转换为 DDC 关联数据,将 DDC22 版足本中最高三级类目以 11 种语言发布为关联数据。UDC 简表发布了 SKOS 版关联数据,2009 年国际虚拟规范档 (VIAF) 发布关联数据,并建立与维基百科的连接,网站访问量明显增加。德国国家图书馆在 2010 年 4 月宣布将其名称和主题规范数据发布为关联数据,匈牙利国家图书馆也宣布将其名称/主题规范数据发布为关联数据。艺术与建筑叙词表 (AAT) 发布了 RDF 版,不仅提供概念体系的浏览,还提供 SPARQL 端点供用户下载数据,以较好地嵌入到外部应用^[22]。LOD 出版类型中的知识组织工具实现的关联发布有 30 个,占出版类数据集的 20%^[5]。

开放数据集大量涌现,词表重用即使用共享词表的术语定义交流信息的共同模式,实现数据集的 RDF 描述,成为关联数据发布的重要

步骤,也成为衡量开放数据质量的重要指标。词表重用能在一定程度上提高数据集之间的互操作能力,避免表述的歧义及冲突,提升数据集的质量,这样对知识组织系统提出了再利用要求,为其进一步应用提供了更广泛的空间。LOV

云图中^[9],高频使用前20的词表见图3。大卫·伍德等对关联数据集统计得出重用的高频率词表有:FOAF、DC、Geo、Bio、Bibo、SIOC、Vcard、CC等^[23],可以看出,两者之间重合的词表较多。

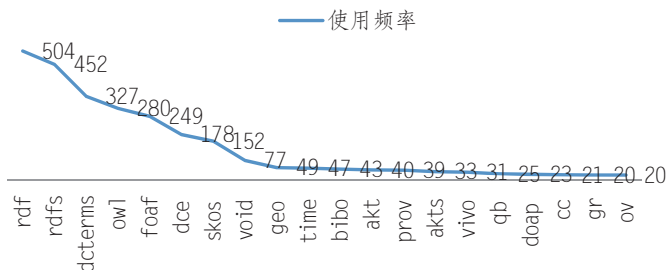


图3 LOD 高频使用的词表

5 结论与展望

数据网络的形成和发展与信息组织的理论和方法息息相关,一方面信息组织是揭示数据网络内容特征的最有效手段,另一方面信息组织工具及使用其描述的数据集是数据网络的重要构成。当前形势下,信息组织的发展水平影响着数据网络的质量及其后期的应用价值,反之,数据网络也会进一步推动信息组织的新理论及方法不断出现。本文通过剖析信息组织处

理功能层次,以理清资源层、元数据记录层及词表层之间的关系,明确各个构件在当前发展阶段所体现的特点及其未来的发展要求。可以看出,面对用户需求精细化及其以机器为中心的资源处理要求,信息组织功能作用在不断延伸拓展,从基本的标引、检索功能发展为互操作、语义交流、语义知识地图等,架构起人与机器、机器与机器之间进行语义交换的桥梁。知识组织系统和元数据集将成为开放数据集不可或缺的语义基础设施,服务于未来的资源深层开发与利用。

参考文献

- [1] Alam M, Buzmakov A, Napoli A. Exploratory knowledge discovery over web of data[J]. Discrete Applied Mathematics, 2018, 249(20): 2-17.
- [2] Vocabularies[EB/OL].[2019-04-23].<https://www.w3.org/standards/semanticweb/ontology>.
- [3] Information and documentation-Thesauri and interoperability with other vocabularies-Part 2: Interoperability with other vocabularies;ISO 25964-2:2013[S/OL].[2019-04-24].<https://www.iso.org/standard/53658.html>.
- [4] Information and documentation-Records management processes- Metadata for records - Part 1: Principles; ISO 23081-1:2017[S/OL].[2019-04-24].<https://www.iso.org/standard/73172.html?browse=tc>.
- [5] 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.)
- [6] Hodge G. Systems of knowledge organization for digital libraries: beyond traditional authority files[EB/OL].[2019-04-24].<http://www.clir.org/wp-content/uploads/sites/6/pub91.pdf>.

- [7] Zeng M L. Knowledge Organization Systems (KOS) [J]. Knowledge Organization , 2008 , 35 (2 / 3) : 160 - 182 .
- [8] Hjørland B. Concept theory [J]. Journal of the American Society for Information Science & Technology , 2019 , 60 (8) : 1519 - 1536 .
- [9] SKOS core vocabulary specification [EB / OL] . [2019 - 03 - 24] . <https://www.w3.org/TR/swbp-skos-core-spec> .
- [10] Linked open vocabularies [EB / OL] . [2019 - 03 - 24] . <https://lov.linkeddata.es/dataset/lov/vocabs/skos> .
- [11] Summers E , Isaac A , Redding C , et al. LCSH , SKOS and linked data [C] // International Conference on Dublin Core and Metadata applications . 2008 : 25 - 33 .
- [12] Caracciolo C , Stellato A , Morshed A , et al. The AGROVOC linked dataset [J]. Semantic Web , 2013 (4) : 341 - 348 .
- [13] 贾君枝 , 赵洁. DDC 关联数据实现研究 [J]. 中国图书馆学报 , 2014 , 40 (4) : 76 - 82 . (Jia Junzhi , Zhao Jie. The realization of DDC as linked data [J]. Journal of Library Science in China , 2014 , 40 (4) : 76 - 82 .)
- [14] 贾君枝. 简单知识组织系统与汉语主题词表 [J]. 中国图书馆学报 , 2008 , 34 (1) : 75 - 78 , 84 . (Jia Junzhi. Simple knowledge organization system and Chinese thesaurus [J]. Journal of Library Science in China , 2008 , 34 (1) : 75 - 78 , 84 .)
- [15] 曾新红. 中文知识组织系统形式化语义描述标准体系研究 (一) —— 扩展 SKOS 实现传统受控词表全描述 [J]. 中国图书馆学报 , 2012 , 38 (3) : 57 - 68 . (Zeng Xinhong. Formal semantic description standard family for Chinese KOS (I) : realizing full description for Chinese controlled vocabularies by expanding SKOS [J]. Journal of Library Science in China , 2012 , 38 (3) : 57 - 68 .)
- [16] 吴雯娜 , 鲍秀林. 国家叙词库的体系结构与数据模型 [J]. 中国图书馆学报 , 2016 , 42 (2) : 81 - 96 . (Wu Wenna , Bao Xiulin. The architecture and data model of national thesauri warehouse [J]. Journal of Library Science in China , 2016 , 42 (2) : 81 - 96 .)
- [17] RDF primer [EB / OL] . [2019 - 03 - 10] . <http://www.w3.org/TR/rdf-primer> .
- [18] RDF / OWL representation of wordNet [EB / OL] . [2019 - 03 - 10] . <https://www.w3.org/TR/wordnet-rdf> .
- [19] FOAF vocabulary specification 0.99 [EB / OL] . [2019 - 03 - 10] . <http://xmlns.com/foaf/spec> .
- [20] National cancer institute thesaurus [EB / OL] . [2019 - 03 - 10] . <http://biportal.bioontology.org/ontologies/NCIT> .
- [21] Library linked data incubator group final report [EB / OL] . [2019 - 03 - 10] . <http://www.w3.org/2005/Incubator/llld/XGR-llld-20111025> .
- [22] 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 .
- [23] 大卫 · 伍德 , 玛莎 · 扎伊德曼 , 卢克 · 鲁思 , 等. 关联数据 : 万维网上的结构化数据 [M]. 北京 : 人民邮电出版社 , 2018 : 36 - 37 . (Wood D , Zaidman M , Ruth L , et al. Linked data : structured data on the web [M]. Beijing : Posts and Telecom Press , 2018 : 36 - 37 .)

贾君枝 中国人民大学信息资源管理学院教授。北京 100872。

(收稿日期 : 2019 - 06 - 28)