

词表语义类型框架与学科差异分析^{*}

——以图书情报领域为例

叶均玲 徐雷 刘瑜

(武汉大学信息管理学院, 武汉 430072)

摘要: 2019年科学出版社出版《图书馆·情报与文献学名词》, 这是图书情报领域的科技名词集合, 对该词表内容的语义类型分析可进一步为理解该领域的学科特征及知识体系提供新的视角。以《图书馆·情报与文献学名词》为分析对象, 从词汇基本特征、词表语义类型框架构建和词表分类框架的学科差异方面进行分析, 总结词表的词频特征和词汇组织形态的不足之处, 并从语义角度对词表的分类体系进行重新认识和组织, 通过与全学科词表分类体系的比较分析总结出优化图情学科词表的相关建议, 如进一步体现词汇间的语义关联、补充完善具有学科交叉性质的词汇、全面推动宣传普及工作。

关键词: 图书情报学; 词表分析; 顶层本体; 语义框架

中图分类号: G251 DOI: 10.3772/j.issn.1673-2286.2023.09.005

引文格式: 叶均玲, 徐雷, 刘瑜. 词表语义类型框架与学科差异分析: 以图书情报领域为例[J]. 数字图书馆论坛, 2023(9): 38-46.

学科领域词表是科技名词统一与规范工作的一部分内容, 也是一项重要的科技成果, 具有重要的科学研究与应用价值。词表的编制一方面凝结了某一学科领域的智慧, 揭示了学科的基本研究对象、研究方法和知识体系等内容; 另一方面, 词表一般遵循一定的信息组织原则与方法进行编制, 对学科领域的知识进行系统性组织, 有利于学科领域的发展以及科技名词的使用、宣传和普及工作。经全国科学技术名词审定委员会发布的词表依据学科知识体系, 收录具有学科学术特点、构成学科基础、常用且重要的名词。目前已陆续公布共计159个词表, 内容覆盖基础科学、工程与技术科学、农业科学、医学、人文社会科学、军事科学等领域。2019年科学出版社出版了《图书馆·情报与文献学名词》^[1]一书, 共收集编撰了3 436个该领域(不包含档案学)的词汇, 解决了图情学科中存在的用词随意、翻译不标准、

一词多义、一义多词等问题。《图书馆·情报与文献学名词》分为图书馆学情报学基础、图书情报工作管理、信息资源建设、信息组织、信息检索、情报分析与研究、信息服务与用户研究、文献学8个一级类目, 具体类别的划分主要依据词汇主题之间的相关性, 例如在图书情报工作管理类别下又划分有一般概念、组织机构、图书馆建筑与设施等7个二级类目, 二级类目下则是具体的词汇。这种以相关性为主的组织方式易于操作, 但在对词汇的语义内涵揭示方面存在不足, 具体体现在词汇之间缺乏丰富的语义关系、未对词汇的本质类属进行划分, 以及不同类别下词汇内涵存在交叉重叠。正如《图书馆·情报与文献学名词》前言部分提到的“该学科是包含广泛知识的跨学科领域, 不仅学科内部彼此交叉甚多、很难截然分开, 学科外部与其他知识也有着千丝万缕的联系”^[1], 这为从不同维度对词表进行重新

收稿日期: 2023-07-28

^{*}本研究得到教育部规划基金项目“科学出版物语义组织模式及其实现路径研究”(编号: 20A10486066)、中央高校基本科研业务费专项资金项目“图情档领域科学本体构建与应用研究”(编号: 1206413000233)资助。

分析与组织提供了可能。本文以《图书馆·情报与文献学名词》为分析对象,借助顶层本体对词表的语义类型进行再组织,同时和其他学科的术语分类体系进行对比,为对学科词表的认识理解以及编撰提供了新的视角和启示。

1 相关研究进展

词表编制有助于推进领域知识体系的建设和标准化工作,对词表的分析和应用研究则有助于审视词表编制工作中的不足,推进词表工作的进一步发展。其中,对领域词表的分析可从词表的外部特征和内部特征两个维度展开:外部视角分析主要对词表的基本情况和发展历程进行梳理,包括词表编制规则、基本结构^[2]、词表版本和编制机构^[3]等;内部视角分析则深入词汇内容,开展词汇基本特征分析和词表语义化、再组织等工作。词汇基本特征分析^[4]是对词表内容构成的初步揭示,包含词语分词、词汇聚类、词频分析等;而词表语义化、再组织工作基于信息组织方法与工具,丰富或重构词表编制方法和知识体系,例如基于简单知识组织系统(Simple Knowledge Organization System, SKOS)等构建术语词表间的概念关系^[5]、将叙词表转换为本体、构建词表顶层语义类型框架^[6]等。此外,在全国科学技术名词审定委员会公布不同学科词表后,也有学者基于知识服务平台“术语在线”对比分析不同学科术语含义。在词表应用方面,国内外都有丰富的实践,主要集中在提供术语服务和构建语义知识库^[7]等方面,例如“术语在线”可提供术语检索、术语标注和术语图谱等服务。总体来说,对于领域词表而言,提升关联性和互操作性对于提升词表结构化程度,推动词表的应用和普及具有较大的价值。

2 词汇基本特征分析

2.1 词频分析

词表中包含了“数据”“信息”等原始词汇以及经过概念组配的复合词汇,如“科学数据”“信息组织”等。使用中文分词工具jieba对原始词汇进行分词处理,对分词结果进行词频统计并选取词频排名前100的词绘制词云图,如图1(a)所示。另外,中文词缀一般可表达词汇的真实语义类型,例如“最大效用原则”一词释

义为“对有限的资源进行最优组合,使资源的有效利用达到最大的满足程度的原则”,其词缀“原则”即可揭示该词汇的本质内涵,对词缀的分析能够进一步挖掘学科领域的核心研究对象。因此,对所有词汇的词缀进行词频统计并绘制词云图,如图1(b)所示。



(a) 词汇词频的词云图



(b) 词缀词频的词云图

图1 词频词云图

从图1(a)中可以发现,“信息”一词的出现频次最高,这一点在词表的组织框架上也有体现:该词表包含了信息资源建设、信息组织、信息检索和信息服务与用户研究4个与信息的生产、组织、检索和应用密切相关的部分。信息是图书情报领域的基本对象,传统的图书文献与当前的数字信息资源都属于广义的信息范畴,同时对信息的研究与实践也充分体现在DIKW(Data-In-

formation-Knowledge-Wisdom)模型从数据到信息、知识、智慧的层级结构中。如图1(a)和图1(b)所示,“检索”一词的出现频次都位于前列,这不仅反映了检索是图情领域的核心研究对象,还体现了该领域以信息检索为核心业务的实践特征。进一步分析发现,“图书馆”“文献”“索引”“目录”“标引”和“服务”等的出现频次都位于前10,体现了图情领域研究与实践的核心内容。一方面,图书馆、文献等作为内容载体,在科学知识生产与消费中起到了推动知识交流与传播的作用;另一方面,图书馆的智慧化建设、文献的形态变革能够进一步推动图情事业的发展。“索引”“目录”“标引”围绕着信息资源的组织,体现了图情领域注重分类方法研究和分类工具研制的特点;“服务”一词则表明了学科研究的目的和导向,即面向用户需求提供各种信息服务方案。此外,在对词缀词频的统计上,发现“本”“法”和“分析”的词频排序比较靠前。其中,“法”有两个内涵指向:一是方法,如分类法、标引法、索引法、排架法等,这体现出图情领域较为成熟的信息组织方法体系;二是法律,如信息法、图书馆法等。“本”主要用于描述文献尤其是古籍的不同版本,如刻本、抄本、印本等,集中在文献学部分,这体现了词表的文献学构成特征。

2.2 词汇组织形态的不足

词汇组织形态主要体现在词汇所在的目录体系上,具体来说原始词表的8个一级类目下又另设多个更细粒度的类目,共有8个一级类目、44个二级类目,类目的设置主要依据词汇之间的主题相关性,例如在信息组织一级类目下的分类法二级类目列出了与分类方法、分类表、分类规则等相关的词汇。词表的原始组织方式虽然在一定程度上反映了词汇之间的相关性,但这种相关性粒度太粗,无法深度体现词汇的概念内涵及词汇间语义关系。例如,在图书情报工作管理一级类目下,图书馆联盟、科学数据共享同属于合作与共享二级类目,但实际上图书馆联盟是图书馆合作组织,科学数据共享则属于一类事件,这两个词汇对应不同的事物类型,而原始的组织方式无法体现词汇的本质属性。再如,在图书馆建筑与设施二级类目下,有书架、书架侧护板等并列的词汇,这两个词汇间具有包含的语义关系,但原始的组织方式同样无法体现这种语义关系,上述情况在整个词表中大量存在。为探究词表词汇的语义本质内涵,挖掘词汇间关联关系,通过引入顶层本体

并构建词表语义类型框架来解决上述问题,为学科词表的深度语义理解提供新的视角。

3 基于顶层本体的词表语义类型框架

3.1 顶层本体选择

本体作为概念的形式化说明,其通过领域的核心概念、概念的内涵和外延、概念之间的复杂关系来全方位地刻画领域事物,其中顶层本体用于描述独立于单个领域的通用性概念。基于顶层本体的分类思想,将词表中的词汇根据术语的语义内涵重新分类。目前顶层本体的构建工作较多^[8],较为著名的是SUMO (Suggested Upper Merged Ontology)^[9]、UCO (Upper Cyc Ontology)^[10]、IFF (Information Flow Framework)^[11]、DOLCE (Descriptive Ontology for Linguistic and Cognitive Engineering)^[12]、OCHRE (Object Centered High Level Reference)^[13]和BFO (Basic Formal Ontology)^[14]等本体。除上述本体之外,研究人员还陆续构建了其他顶层本体并将其投入大规模应用,如UFO (Unified Foundational Ontology)^[15]、GFO (General Formal Ontology)^[16]等。

顶层本体的语义框架较为抽象,为确保词表语义类型框架的科学性和适用性,筛选出契合图情领域词汇特征、应用较为广泛、形式较为规范的顶层本体作为可参考的语义框架。由于DOLCE融合了人类的感知和认知特征^[17],相较于BFO、SUMO等本体更适合描述包含人类认知产物的抽象对象,考虑到图情领域与社会活动和信息资源密切相关,最终选择DOLCE作为词表语义类型框架。

3.2 图情词表的语义类型框架设计

构建词表的语义类型框架是一个自顶向下和自底向上相结合的过程:一方面,根据顶层本体的分类思想重新组织和归类词表中的词汇;另一方面,基于词汇的内涵有选择地继承、调整、改造DOLCE中的概念。顶层本体映射流程如图2所示。为保证词表语义类型框架的客观性和合理性,一方面基于顶层本体相关研究中对类和属性的解释和实例来确定顶层本体的含义;另一方面结合词汇词缀和词汇释义确定词汇的语义标签,聚类语义标签之后与顶层本体分类框架进行对齐、关联,从而构建出最终的词表语义类型框架。

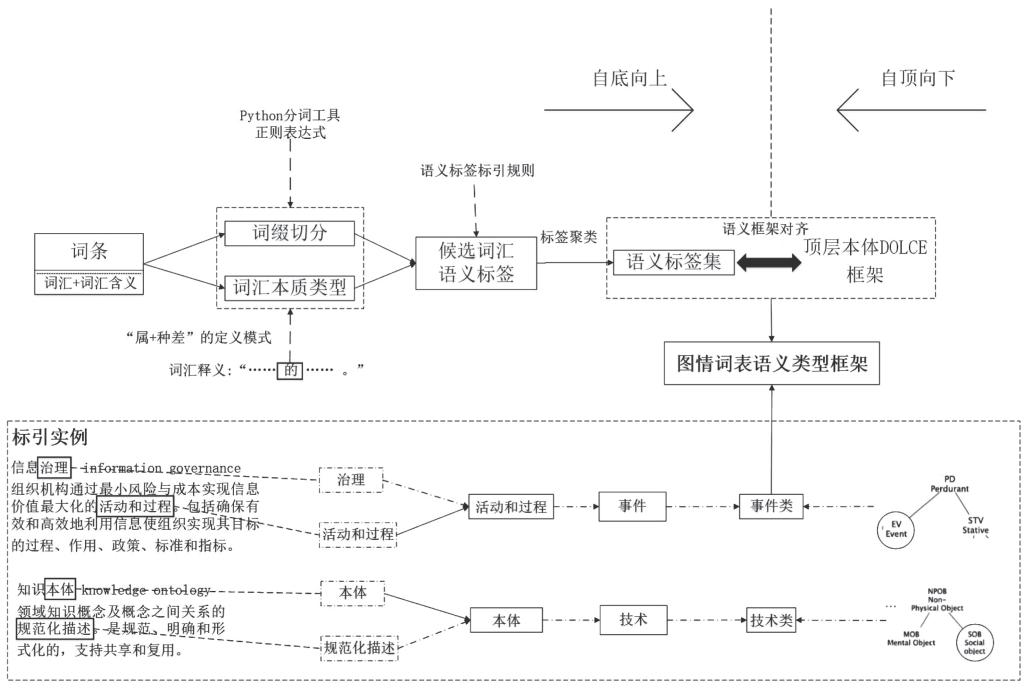


图2 顶层本体映射流程

全国科学技术名词审定委员会在名词审定章程中说明,在拟写名词定义时力争按“属+种差”模式拟出并推敲斟酌,以显示术语间的从属关系,即词汇释义文本中末位的“的”字和释义末尾句号之间的词汇可视为词汇的本质类型,而词汇后缀在一定程度上也能反映出词汇的本质,因此结合词汇释义中对词汇内涵的界定和词汇后缀共同确定词汇的语义标签。词汇的语义标签不宜过于抽象,否则可能放大词汇的真实含义,导致语义类型框架缺乏层次性;但也不能过于具体,否则

会给后续语义标签聚类工作带来负担,影响语义类型框架的简洁性。表1给出了语义标签的部分标引规则。在为词汇赋予语义标签的过程中发现,使用词汇后缀便于寻找词汇的同位类以及下位类,而使用词汇释义中的内涵界定便于寻找词汇本质也即词汇的上位类。由于对词汇赋予语义标签并进行聚类是一个自底向上的过程,在词汇词缀和本质类型都能体现词汇语义特征的情况下一般优先选用词缀,从而将具有相同本质类型的词汇聚类,以便于后续与顶层本体分类框架映射。

表1 词汇语义标签标引规则

序 号	标引规则	词汇示例	语义标签	语义标签集
1	必须结合词汇本质类型和词汇词缀共同确定词汇语义标签	装裱	技艺	技术
		全形拓	拓印技法	技术
2	在词汇词缀和本质类型都能体现词汇语义特征的情况下,选择中观粒度的语义标签,一般优先选择词缀	图书馆计划管理	管理	事件
		图书馆目标管理	管理	事件
3	在词汇词缀无法体现词汇含义时,选用词汇本质类型概念	图书馆学	学科	学科
		公共接入点	图书馆设备	图书馆设备用品
4	当同一词缀的词具有不同的词汇本质类型时,不应赋予不同的语义标签,一般统一以词缀为语义标签	学位论文	学位论文	文献资料
		硕士学位论文	学位论文	文献资料
		博士学位论文	学位论文	文献资料
5	当词汇较为抽象时,可结合词缀和词汇本质类型简单调整语义标签,使得语义标签能够真正体现词汇本质类型	图书情报一体化	方案	方法
		文献工作标准化	规范化	事件
6	对部分出现频次较低的词汇,考虑不赋予语义标签,直接保留词汇本身与核心词汇进行关联	书口	书口	文献特征
		版心	版心	文献特征
		书耳	书耳	文献特征

在标注词汇语义标签后,人工开展语义标签的聚类工作,得到32个标签集,标签集的名称、含义及其包含的语义标签示例如表2所示,将标签集与DOLCE的分类框架进行对齐。DOLCE的顶层类有静态实体(Endurant)、动态实体(Perdurant)、属性(Quality)和抽象域(Abstract)^[17],其中:静态实体(Endurant)类用于描述不受时间因素限制的实体,例如Agentive Physical Object类下的实例“a human person”;动态实体(Perdurant)类则相反,其必须存在于一定的时间情境下,例如实例“跑步”“写作”等。在对静态实体(Endurant)类的细分方面,主要依据是否具有物理存在分为“physical”和“non-physical”,以及依据是否存在意图、信念和欲望分为“agentive”和“non-agentive”。

此外,属性(Quality)和抽象域(Abstract)类作为分类框架的补充,属性(Quality)内生于静态实体(Endurant)和动态实体(Perdurant),具体有时间、地点和抽象属性。对齐的过程中发现,在借鉴DOLCE分类思想的基础上,应根据图情词表词汇特征进一步细化对齐过程。例如,在对数字资源的分类上,“数字图书馆”“网页”“计算机系统”等词汇因存在于网络空间而不具备物理属性,但是其作为数字化资源的载体具有承载数字化信息资源的功能并且具有唯一标识符作为其存在的位置指向,因此依然将其划分为物理对象(Physical Object)。按照上述标引规则与对齐路径,词表中所有词汇都可归入语义类型框架,词表的语义类型框架如图3所示。

表2 标签集名称及其含义

标签集名称	标签集含义	语义标签示例
文献特征	描述文献资料尤其是古籍文献的特征组成部分	书口、版心
文献资料	描述不同形式、不同种类的文献、资料	文献、出版物、刻本
系统平台	描述图书情报工作中可能使用的计算机系统、社交平台、应用软件等	数据库、系统、软件
建筑场所	描述用于容纳人员、文献资料和设备等的建筑、地点,如储存文献资料的图书馆	图书馆、书库、场所
文献载体	描述文献等的承载实体,包括传统和数字化形式载体	纸张、光盘、磁带
图书馆设备用品	描述图书馆日常运维使用的设备、物品	书架、设备、卡片
心理状态	描述人员的感受、意识等心理活动的产物	需求、兴趣、倾向
人员	描述具有社会意义的人	用户、图书馆员、学生
组织机构	描述承担一定社会功能的组织、机构	机构、信息中心、学会
语词	描述图书情报学中的基本语言、词汇	词、叙词、概念
符号	描述由语言、数字等组成的用于表征或标识的词汇	编号、号码、符号
目录	描述起到分类作用的目录、索引等	目录、索引、款目
词表	描述词表及其组成部分	词表、叙词表、复分表
方法	描述方法、策略、计划、手段等具有指导作用的词汇	方法、标引方法、排架方法
分类系统	描述基于分类符号、原则和方法等的分类系统	分类体系、标引系统
原则标准	描述用于规范类目、图书情报工作的原则、标准、协议等	原则、标准、标记制度
信息对象	描述以数据、信息、知识、情报、信息资源为核心的图书情报学核心概念	信息、竞争情报、情报
理论学说	描述图书情报学的核心理论、学说	理论、定律、原理
学科	描述图书情报学的分支和交叉学科等	学科、分支学科
经济要素	描述与图书情报业务活动相关的经济要素	图书馆基金、资产
现象	描述图书情报工作中的现象	现象、状态
技术	描述信息技术和相关技艺	本体、技术、拓印技法
历史	描述发展历史	图书馆史
角色	描述社会人员承担的角色	责任者、著者
道德素养	描述图书情报工作中的基本道德标准	职业道德、网络素养
事件相关概念	描述难以归类的与检索、标引事件紧密相关的部分概念	检索结果、检索提问单、引文矩阵
抽象总合	描述由多个要素共同构成的复杂性综合概念	电子政务、文化、信息源
事件	描述过程、活动、工作等动态概念	过程、工作、活动
时间属性	描述事件发生的时间特征	时间、生命周期
地点属性	描述位置特征	地理位置、书架方便区
抽象属性	描述除时间、位置属性外的其他属性,典型的有指标属性、格式属性、装帧形式和图书馆属性等	价值、要素、能力

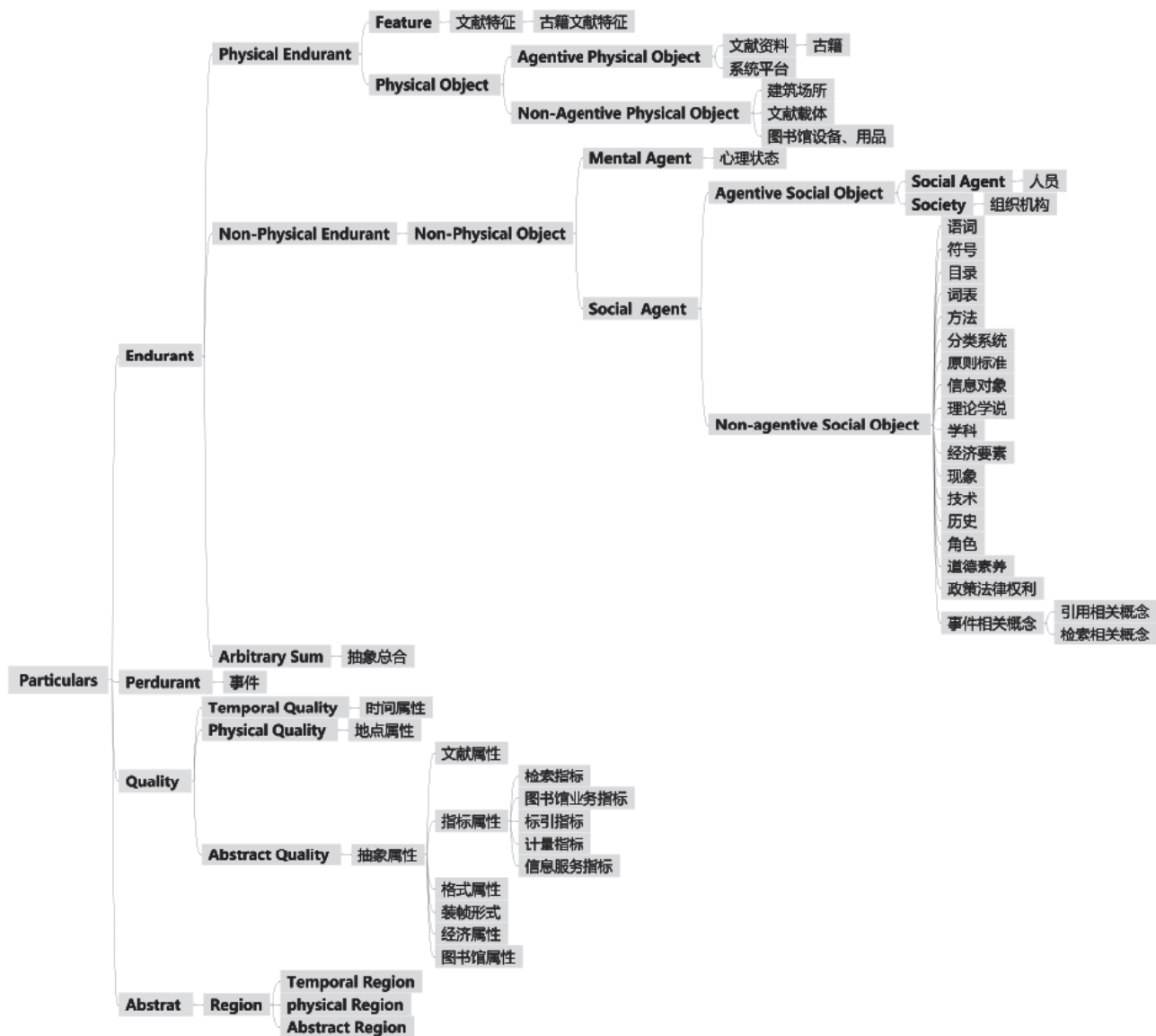
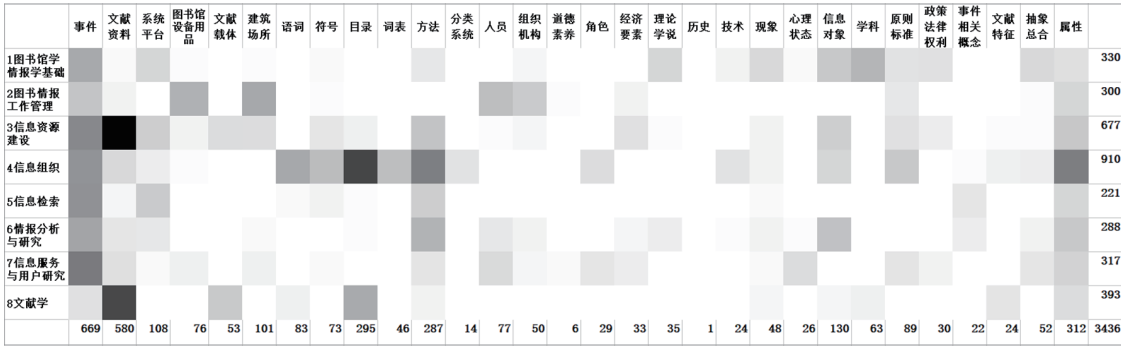


图3 图书情报学词表语义类型框架

在此基础上对原始的8个一级类目和图3所示的32个语义类型进行交叉分析, 其中将抽象属性、时间属性和地点属性统一为属性(Quality), 如图4所示。具体来说, 图4中交叉程度较高的部分主要集中在信息组织一级类目和事件语义标签集上, 整体的交叉程度不高, 表明原始的词汇组织方式和词表语义类型框架是两类差异较大的词汇组织体系, 词表语义类型框架可为理解词表提供新的视角。具体来看, 在8个一级类目中, 图书馆学情报学基础类目主要包括学科、信息对象、理论学说和系统平台等标签集; 图书情报工作管理类目主要包括建筑场所、图书馆设备用品、人员、组织机构等标签集, 即围绕着图书馆及其内部人员、设备情况, 并收录了相关的图书情报机构; 信息资源建设类目主要包括

文献资料、方法和信息对象等标签集, 列举了典型的文献资料、图书馆对文献资料等的整理方法以及信息、资源两类信息对象; 信息组织类目主要包括目录、方法、语词、符号、原则标准和信息对象等标签集, 该类目的词汇数量占比较大, 这表明信息组织在图书情报学领域的重要地位, 也反映出信息组织研究的成熟度较高; 信息检索类目主要包括平台和方法等标签集; 情报分析与研究类目主要包括方法和信息对象等标签集, 围绕情报研究方法和情报对象; 信息服务与用户研究类目主要包括人员和心理状态等标签集, 这也在一定程度上反映出信息服务与用户研究的重点指向; 文献学类目主要包括文献资料、目录和文献载体等标签集, 这与文献学的学科内容紧密相关。



注：颜色深浅表示不同词汇分类框架的交叉程度高低，颜色越深则交叉程度越高；数字表示不同分类框架包含的词汇数量。

图4 两种词汇分类框架的交叉分布图

4 词表分类框架的学科差异分析

考虑到图情领域具有明显的交叉学科特征，其发展离不开信息技术，图情学科中的理论方法又可被应用于医学、生物学、国家安全学等学科，因而尝试从学科差异角度分析图情词表与其他学科词表分类框架设置的差异，从而为图情学科的定位以及图情词表的后续编制提供参考。参考莱顿大学的科学研究领域分类^[18]，收集分属于基础科学、工程与技术科学、生命与地球科学、生物医学与健康科学和人文与社

会科学领域的共16个词表（见表3）的分类框架，并总结不同领域词表的分类体系特征，与图情词表进行比较，探索不同学科在知识组织方面的逻辑差异。总体来说，学科词表围绕概念体系的基本框架收录科技名词，其体系设置遵从中图法分类原则：从理论到实践，从基础理论到应用；按事物发展从简单到复杂、从低级到高级的次序，按社会及社会现象的发展过程；按从一般到具体、从总论到专论的次序。在此基础上，不同学科领域在分类框架的设置上也表现出明显的差异。

表3 学科领域与典型的学科词表

学科领域	基础科学	工程与技术科学	生命与地球科学	生物医学与健康科学	人文与社会科学
词表名称	《物理学名词》 《数学名词》 《化学名词》	《计算机科学技术名词》 《测绘学名词》 《电力名词》	《农学名词》 《土壤学名词》 《地理学名词》 《古生物学名词》	《精神医学名词》 《烧伤学名词》 《微生物学名词》	《管理科学技术名词》 《教育学名词》 《经济学名词》 《图书馆·情报与文献学名词》

基础科学领域的词表基于学科概念体系，直接以分支研究方向为词表的分类框架，其类目之间具有显著的区别，这依赖于数学、物理学等基础性学科的悠久学科历史，而图情学科尚不具备该特征，其一、二级类目设置存在明显的交叉。生命与地球科学领域词表同样基于学科的核心构成划分，添加了交叉学科性质的主题，并呈现出从静态概念到动态应用的趋势。工程与技术科学领域的词表按照从通论（总论）到具体研究分支主题、从理论到实践、最后补充新兴交叉学科概念的框架组织，体现出工科类学科立足于基础学科原理、面向生产实践的特性。相比于这两个领域的词表，图情学科现有的词表框架中尚且缺乏能够体现交叉学科性质的词汇^[19]，如数字人文、元宇宙、安全情报学、健康信息学

等，这不仅表现出当前词表在类目组织上的不足，还表明当前词表编撰更新的滞后性。生物医学与健康科学词表面向医学研究与实践，体现出较强的实践性和业务逻辑。例如，《精神医学名词》下设置了总论、症状与体征、诊断分类、治疗和常用量表5个一级类目，围绕着精神医学诊断治疗展开，其类目具有明显的逻辑关系，而《图书馆·情报与文献学名词》目前主要以主题罗列的方式组织。人文与社会科学领域的词表总体上也按照核心主题划分，特别之处在于其注重梳理学科发展历史脉络，如《经济学名词》《教育学名词》专门设置章节介绍经济思想史、教育学史相关名词，而《图书馆·情报与文献学名词》相对弱化学科历史，仅有“图书馆史”一词表明图书馆和图书馆事业产生和发展的历史。

5 结论与启示

5.1 结论与不足

从词表内部语义特征和外部学科差异入手重新组织和认识了图情领域词表: 先从词频等角度认识《图书馆·情报与文献学名词》的词汇分布情况和组织形态, 再从语义视角借助顶层本体的分类体系构建《图书馆·情报与文献学名词》的语义类型框架, 最后在学科交叉融合的背景下, 跳出图情领域视角, 比较全学科词表分类框架的差异。在词汇分布方面, 《图书馆·情报与文献学名词》围绕“信息”“检索”“图书馆”“文献”等核心词汇, 体现出图情学科的应用性、方法性特征; 在词表组织方面, 《图书馆·情报与文献学名词》的原始组织方式虽然在一定程度上反映了词汇之间的相关性, 但无法深度体现词汇的概念内涵及词汇间语义关系。在词汇的基本组成单元即“词汇+含义解释”的基础上, 为总共3 436个词汇赋予语义标签并得到32个语义标签集, 在与顶层本体DOLCE对齐后得到词表的语义类型框架, 提供了从语义视角认识词表的新途径。通过与全学科词表的比较分析, 梳理了不同的词表分类体系的特征与差异。不过, 研究也存在一些不足: 词表词汇量较大且语义关系类型多样, 在实际操作过程中难以穷尽, 后续可结合具体场景来设置抽象的词汇语义关系类型。

5.2 启示

词表编制和发布有助于推进图情学科研究体系的构建和学科名词规范化事业的发展, 但是通过分析发现主要基于主题相关性的词表组织方式未能体现词汇的语义本质, 词汇间缺乏语义关联。本文在词表语义化的方向上进行了初步探索, 未来可进一步探究结合词汇主题特征和语义特征的词表组织方法, 改善词表的编制与更新机制, 推动词表的标准化发布与应用。

具体来说, 当前基于词汇主题特征的术语组织方式的主要优势在于可简便快捷地将候选术语组织在一起, 在组织层面操作性较强, 而基于词汇语义特征的组织方式则需要进行详细的领域分析, 构建领域概念框架及类属结构, 操作成本较高但对领域的揭示程度更高。两种组织方法各有优劣, 未来可探索结合这两种方法的术语词表组织方式。

通过前述分析发现, 《图书馆·情报与文献学名词》中尚且缺乏具有交叉学科特征的前沿性词汇, 未来可思考更新词表的新途径, 如通过构建学科语料库、借助机器学习算法自动化抽取学科新兴词汇, 再由学科专家集体决策来完善词表的知识体系。例如, CSO (Computer Science Ontology)^[20]基于包含约1 600万种出版物的数据集, 使用Klink-2算法自动生成树状结构的领域词表, 可通过在新发布出版物上运行该算法进行自动更新。相较于专家人工编制的词表, CSO不用经历漫长的词表编撰历程并且能够自动化更新本体词汇, 还添加了relatedEquivalent、superTopicOf和contributesTo等语义关系来描述研究主题之间的等价、父子和贡献等关系。这种以智能化算法为主构建学科词表的方式值得在未来深入探索, 并尝试应用在图情学科词表的编制及学科新名词的发现之中。

学科词表相关工作的目的之一是推动科学技术名词的宣传推广和应用普及, 但词表目前所依托的“术语在线”平台仅提供术语查询服务, 而不具有数字化版本词表的浏览、交互、引用等功能, 这在一定程度上限制了用户对词表的使用。今后应当全面布局词表的数字化、标准化发布与共享工作, 以更好地促进学科研究和领域实践。

参考文献

- [1] 图书馆·情报与文献学名词审定委员会. 图书馆·情报与文献学名词(2019)[M]. 北京: 科学出版社, 2019: 7.
- [2] 张士男. 国外叙词表宏观结构分析: 以理学领域叙词表为例[J]. 图书馆论坛, 2014, 34(4): 5-11.
- [3] 郝嘉树. 我国印发型公文主题词表分析及分面改造[J]. 图书馆工作与研究, 2016(6): 63-67, 76.
- [4] 李传芳, 余军, 林竞. 商务英语词汇: 描述、认定与教学: 基于《财富》商务英语词表的数据分析[J]. 外语界, 2015(3): 83-88, 96.
- [5] 石泽顺, 肖明. 基于网络叙词表的图情学科SKOS构建与可视化研究[J]. 情报学报, 2018, 37(3): 274-284.
- [6] 吴霞, 曾建勋, 吴雯娜. 汉语主题词表生物、医学、农业领域顶层本体语义类型框架研究[J]. 情报科学, 2022, 40(1): 94-101.
- [7] 陶俊. 词表语义组织研究的演进(1998—2018)[J]. 图书情报工作, 2018, 62(21): 140-148.
- [8] GUINARD A. A survey of top-level ontologies[EB/OL]. [2023-05-17]. <http://www.cnterm.cn/sdgb/sdyzjff/>.

- [9] The Suggested Upper Merged Ontology (SUMO) [EB/OL]. [2023-05-17]. <https://www.ontologyportal.org/>.
- [10] MATUSZEK C. An introduction to the syntax and content of cyc[C]//2006 AAAI Spring Symposium on Formalizing and Compiling Background Knowledge and Its Applications to Knowledge Representation and Question Answering, 2006.
- [11] KENT R E. The information flow framework: a descriptive category metatheory[EB/OL]. [2023-05-17]. <https://arxiv.org/abs/1108.4133>.
- [12] CNR-ISTC Laboratory for Applied Ontology. DOLCE: descriptive ontology for linguistic and cognitive engineering[EB/OL]. [2023-05-17]. <http://www.loa.istc.cnr.it/dolce/overview.html>.
- [13] SCHNEIDER L. Designing foundational ontologies[C]//SONG I Y, LIDDLE S W, LING T W, et al. Conceptual Modeling-ER 2003, 2003: 91-104.
- [14] SMITH B. Basic Formal Ontology (BFO) [EB/OL]. [2023-05-17]. <https://basic-formal-ontology.org/>.
- [15] GUIZZARDI G. The Unified Foundational Ontology (UFO) [EB/OL]. [2023-05-17]. <https://nemo.inf.ufes.br/en/projetos/ufo/>.
- [16] HERRE H. General Formal Ontology (GFO) [EB/OL]. [2023-05-17]. <https://www.onto-med.de/ontologies/gfo>.
- [17] BORGIO S, FERRARIO R, GANGEMI A, et al. DOLCE: a descriptive ontology for linguistic and cognitive engineering[J]. Applied Ontology, 2022, 17 (1) : 45-69.
- [18] The CWTS Leiden ranking 2023 five main fields of science[EB/OL]. [2023-09-28]. <https://www.leidenranking.com/information/fields>.
- [19] 马费成, 张帅. 我国图书情报领域新兴交叉学科发展探析[J]. 中国图书馆学报, 2023, 49 (2) : 4-14.
- [20] ANGELO A S. The computer science ontology: a large-scale taxonomy of research areas[C]//International Semantic Web Conference 2018, 2018.

作者简介

叶均玲, 女, 硕士研究生, 研究方向: 知识组织与知识服务, E-mail: 2538802980@qq.com。

徐雷, 男, 博士, 副教授, 研究方向: 科学交流与知识建模。

刘瑜, 男, 硕士研究生, 研究方向: 知识挖掘与数据分析。

Analysis of Semantic Type Framework and Disciplinary Differences in Vocabulary: A Case Study of Library and Information Science

YE JunLing XU Lei LIU Yu

(School of Information Management, Wuhan University, Wuhan 430072, P. R. China)

Abstract: The Science Press published *Library and Information Science Nouns* in 2019, which is a collection of scientific and technological terms in this field. Multidimensional semantic type analysis of the vocabulary content can further provide a new perspective for understanding the disciplinary characteristics and knowledge system of the field. This article takes *Library and Information Science Nouns* as analysis object, and analyzes the basic features of vocabulary, the construction of lexical semantic type framework, and the subject differences of lexical classification framework. Firstly, the frequency characteristics and the problems of vocabulary organization forms are clarified, and the classification system of vocabulary is re-recognized and organized from the semantic perspective. Finally, by comparing and analyzing the classification system of vocabulary of whole subject, relevant suggestions for optimizing the subject vocabulary of library and information science are obtained. For example, the semantic correlation among words should be further reflected, the words with interdisciplinary nature should be supplemented and perfected, and the publicity and popularization should be comprehensively promoted.

Keywords: Library and Information Science; Vocabulary Analysis; Top Level Ontology; Semantic Framework

(责任编辑: 王玮)