基于内容特征元数据的多源异构科技资源关联聚合研究

刘 伟
（中国科学技术信息研究所，北京 100038）

摘要：科技资源已成为推动科学技术进步、提升国家科技实力的关键性因素，但科技资源的孤岛问题严重阻碍了科技资源共享服务。在分析科技资源内容特征元数据的基础上，关联聚合研究多源异构的科技资源的方法，利用知识组织工具从内容特征元数据中抽取主题概念，利用主题概念建立科技资源之间的关联，继而对不同类型和来源科技资源进行聚合，利用真实的科技资源元数据进行实证分析，表明该方法在揭示科技资源共享服务方面的有效性。

关键词：科技资源；内容特征；元数据；主题抽取；关联聚合
中图分类号：G203 **文献标识码：**A **DOI：**10.3772/j.issn.1674−1544.2020.05.005

Research on Correlation and Aggregation of Scientific and Technical Resources Based on Content Feature

LIU Wei
(Institute of Scientific and Technical Information of China, Beijing 100038)

Abstract: Scientific and technical resources have become a key factor to promote scientific and technical progress and enhance national scientific and technical strength, but the problem of isolated islands of scientific and technical resources has seriously hindered the sharing of scientific and technical resources. By analyzing the characteristics of the content of scientific and technical resources metadata, multi-source heterogeneous resources of science and technology association polymerization method is proposed, which extract theme concepts from the content features of scientific and technical resource metadata with a knowledge organization tool, the subject concepts are used to establish the correlation between scientific and technical resources to implement the aggregation of different types and sources of scientific and technical resources. The empirical analysis on the real scientific and technical resources metadata has carried out to show that the effectiveness of the method.

Keywords: scientific and technical resource, content feature, metadata, subject extraction, correlation and aggregation

0 引言

科技资源是科技活动产生的各类资源的总称，是科技创新的要素集合，比如各种科技文献、科学仪器、科学数据、种质资源等，也包括科技人才、科研项目和科研环境设施等，广义上把科技资源划分为科技人力资源、科技财力资源、科技物力资源和科技信息资源四大类[1−2]。

作者简介：刘伟（1976—），男，中国科学技术信息研究所副研究员，研究方向：知识组织，信息检索。

基金项目：国家重点研发计划“分布式科技资源体系及服务评价技术研究”课题二“面向资源科技云的分布式协同机制及模式研究”（2017YFB1400200）。
收稿时间：2020 年 6 月 15 日。

科技资源属于国家科技发展的战略性资源，各国都在加大科技资源的投入。随着我国对科技投入的逐年增加，积累了非常丰富的科技资源，但目前我国科技资源松散孤立，缺少有效的互通协调和配置管理，导致“科技资源孤岛”现象的产生，不能充分发挥科技资源其应有价值与作用[3]。为了推动科技资源的共享服务，经过多年努力，已在多个领域建成一批国家和地方的科技资源共享服务平台[4]。但是，目前科技资源的共享服务仍然处于整合与重组的阶段，只是把不同来源和不同类型的科技资源元数据整合在一起，但科技资源之间仍然是孤立的，不能够协同为用户服务。这一问题已经引起研究者的关注，并开始科技资源关联聚合方面的研究。该方面的研究主要分为两个层面：一是科技资源的描述，二是科技资源的关联聚合方法。科技资源的描述是建立对通用型或专业型的科技资源元数据规范和标准，只有在统一描述的基础上才能对科技资源进行有效的关联聚合。由OCLC和NCSA在1995年提出的都柏林核心元数据（简称“DC元数据”）[5]是目前世界上应用最广泛、通用型资源描述的国际标准，可以揭示科技资源的最小元数据元素集。国家标准《科技平台资源核心元数据》[6]和宋佳等[7]提出的“科技基础性工作专项科技资源核心元数据规范”都从共享服务的角度针对科技资源的特点规定了描述科技资源基本信息的元数据最小集合。通用型科技资源的描述适用性强，有利于多类别、跨领域的科技资源进行数据交换，但同时也缺乏领域针对性以及科技资源特性。不同类型科技资源的差异化导致特定类型科技资源元数据的提出，比如面向科技报告的《科技报告元数据规范》[8]、针对科学数据的《中国科学院科学数据库核心元数据标准》[9]，而针对特定类型资源对象的元数据适用于同类型资源的数据交换与集成。此外，即使同一类型的科技资源也呈现专业化且精细化的趋势，比如生物多样性的达尔文核心（Darwin Core）元数据标准[10]、地理空间数据的FGDC/CSDGM元数据标准[11]等。这些科技资源元数据专业性较强，部分元数据元素设计具备鲜明的领域特点，不利于跨学科领域的关联。科技资源的关联聚合方法在早期是对科技资源按照的集成，科技资源之间是相互独立的，随着用户需求从信息层面向知识层面的提高，逐渐发展为基于语义的跨类型跨来源的关联聚合。毕强等[12]将数字资源语义化方法划分为组织语义化以及内容语义化两个方面，瞿辉等[13]提出利用语义化共词分析方法实现基于主题的馆藏资源多维度聚合，尚珊等[14]运用社会网络中的凝聚子群法实现专家团队的聚合，以及周姗姗[15]将分众分类法和复杂网络分析有机融合提出的数字资源多维度聚合方法。

综上所述，关于科技资源关联聚合研究的主要问题是不能很好地解决不同来源、不同类型的科技资源之间的关联聚合。其根本原因是不同来源、不同类型的科技资源之间描述的元数据标准规范不一，难以在元数据元素之间建立准确可用的映射。本文将针对上述问题，基于内容特征研究适用于不同类型、不同来源的科技资源之间关联聚合的方法，满足当前用户对科技资源协同服务的需求。本方法首先识别能够表达科技资源内容特征的元数据元素，从内容特征元数据元素中抽取主题概念，利用主题概念建立科技资源之间的关联，最终实现对多源异构科技资源的聚合。

1　科技资源的内容特征元数据

元数据的本质是关于数据的数据，科技资源是通过元数据进行描述的，我国建设的数量众多的科技资源共享服务平台对元数据进行整合。不同的元数据对描述科技资源起着各自的作用，研究者提出了各自的元数据分类，比如马珉[16]将元数据分为描述型元数据、结构型元数据、存取控制型元数据、评价型元数据，Murtha Baca[17]将元数据分为管理型元数据、描述型元数据、保存型元数据、技术型元数据和使用型元数据。这些元数据的分类主要是面向数字资源的管理与检索。

当前，随着科技资源越来越丰富，呈现出规模大、增速快、类型多样、异构分散、学科覆盖广泛等大数据的特点。同时，科技活动也日趋复

杂，单一的科技资源已无法满足其需求，需要通过关联聚合使科技资源能够为用户协同服务。因此，本文面向科技资源的共享服务，从关联聚合的角度提出了新的科技资源元数据分类，将科技资源元数据分类为外部特征元数据、内容特征元数据和服务特征元数据三类。面向共享服务的科技资源元数据的类别如图1所示。外部特征主要用于科技资源的标识，比如URI、DOI等唯一性标识和规范的科技资源名称等，以区分科技资源，建立科技资源之间的依属关系。尽管利用外部特征也可以在科技资源之间形成关联，但主要用于科技资源的组织管理，并不是面向用户的共享服务需求。内容特征是对科技资源内在特征的描述，能够直接或间接表现资源内容主题性质的属性，包括关键词、类型、功能、用途等方面。科技资源共享服务平台主要是根据内容特征来判断是否是一项科技资源就能满足用户当前需求的。服务特征是科技资源所有者为其科技资源在附加的元数据元素，用以描述科技资源服务的条件、约束和限制等，比如一台科学仪器只在工作日提供服务，就无法满足用户在节假日使用该科学仪器的需求。

不同类型和来源的科技资源元数据表达不统一，可以根据实际情况识别科技资源的内容特征元数据。如果科技资源的来源较少，以人工的方式将内容特征元数据标注出来即可；如果科技资源的来源广泛并且元数据元素较多，采用人工方式其效率可能较低。但可以利用元数据元素的名称特征和元数据元素的值特征制定启发式规则来识别内容特征元数据。内容特征元数据无论是在名称还是在值的特征上，与外部特征和服务特征都比较容易区分。在实际的科技资源服务中，用户的需求更加关注于科技资源的主题，其科研活动中使用的科技资源通常是主题相同或相近的，而科技资源的主题主要体现于内容特征元数据中。因此，本文认为利用科技资源的内容特征元数据进行聚合具有更加现实的应用意义。此外，将相同相近主题特征的不同类型和来源的科技资源聚合在一起，避免了不同类型和来源的科技资源之间元数据表达不统一的问题，为用户提供更好的科技资源达到协同服务的目标。

2　科技资源主题概念抽取步骤

为了将相同相近主题特征的不同类型和来源的科技资源聚合在一起，就要抽取主题概念。从科技资源中抽取主题概念需要经过两个步骤：第一步是从内容特征元数据中抽取主题概念；第二步是将抽取到的主题概念计算权重。

2.1　主题概念的抽取

尽管科技资源的元数据在语言表达上比普通的文本要规范严谨，但内容特征中的很多元数据仍然属于无结构的自由文本，比如摘要、简介和

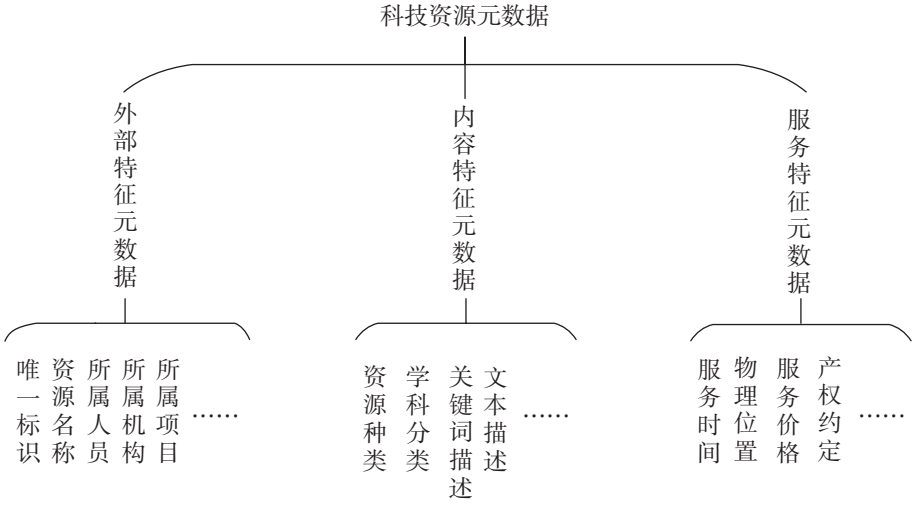

图1　面向共享服务的科技资源元数据的类别

备注等，因此从中抽取术语必须要准确识别主题概念的边界。简单直接的方法就是利用开源的中文分词工具实现在自由文本中识别主题概念的边界，即首先将每一项内容特征元数据分词，筛选出其中的主题概念。分词工具对分词的准确性主要依赖于所使用的词库。如果被切分的主题概念没有在词库中，就会出现将主题概念“切碎”和“切错”的情况。将主题概念“切碎”是由于未登录词而造成的问题，使一个长主题概念被切分成了若干个短词，比如“心脏植物神经”被切分为“心脏\植物\神经”3 个词。将主题概念“切错”是由于歧义切分而造成的问题，导致主题概念边界识别错误，比如“机器人类脑算法”正确的切分结果应是“机器人\类脑算法”，但如果分词工具使用的词库中没有主题概念“类脑算法”，则被错误地切分成“机器\人类\脑\算法”。

为了解决上述问题，必须使用专业术语库与分词工具的通用词库结合达到对主题概念准确切分。因为以主题概念抽取为目标，在切分词时要以专业术语库中的术语优先，即如果存在多种可能的分词方案，对每种分词方案中出现的专业术语进行评估，选出其中的最优方案，再利用最优方案切分出来的专业术语即为主题概念的抽取结果。主题概念抽取的核心是对多种可能的分词方案进行评估选优。本节提出的主题概念抽取方法首先将一个元数据元素S的切分进行形式的表达如下：

$$C=\{P_1,P_2,\ldots,P_n\}$$

其中，$P_i\left(1\leqslant i\leqslant n\right)$是指对元数据元素$S$的$n$种切分方案。设$T_i$为$P_i$中切分出来的主题概念集合，$P_i$可以用$T_i$来代替，即：

$$C=\{T_1,T_2,\ldots,T_n\}$$

这样对每种切分方案P_i的评估就转变成了对相应主题概念集合T_i的评估。事实上，即使都是主题概念，对科技资源的代表能力也是有强弱的区分的。为此，从主题性角度对科技资源的表达能力进行评估。主题概念的主题性借鉴信息检索中TF−IDF模型来计算。计算公式如下：

$$T\left(t_i\right)=\left|t_i\right|\times\log\frac{|R|}{|R_i|}$$

其中，$\left|t_i\right|$是指主题概念t_i在一项科技资源的内容特征元数据中出现的频次，$|R|$是指所有科技资源的数量，$|R_i|$是指科技资源内容特征元数据中出现主题概念t_i的科技资源的数量。进一步，$E\left(T\right)$为T中所有主题概念的$T\left(t_i\right)$之和，而最优方案为：

$$\text{Max}\{T_i\,|\,E\left(T_1\right),E\left(T_2\right),\ldots,E\left(T_n\right)\}$$

这样，对科技资源中的每个内容特征元数据最优方案中的主题概念集合，就实现了从一项科技资源的内容特征元数据中对主题概念的抽取。现实中通常存在多个主题概念相互同义的情况，比如“色谱仪”“色谱系统”和“色谱分析仪”。从科技资源的内容特征元数据中抽取主题概念之后，需要进一步将相互同义的主题概念合并为一个主题概念才能真正实现主题概念的抽取。

2.2　主题概念的加权

从科技资源的内容特征元数据中进行主题概念抽取后，每项科技资源得到了一个主题概念的集合。主题概念集合中每个主题概念对科技资源主题特征的代表性并不相同，比如一篇关于原子能应用的文献，经过抽取得到“原子能”“应用”和其他主题概念，显然“原子能”比“应用”更加能够代表这篇文献的主题，因为“应用”是一个使用广泛的通用词，而“原子能”则是一个专业性非常强的词。因此，需要量化主题概念集合中每个主题概念的主题代表性。

TF−IDF是一种用于信息检索中广泛使用的计算词权重的技术，主要根据词频和该词汇命中的文档数进行计算。本文对TF−IDF技术进行改进，用于计算主题概念的主题代表性。改进之处主要有 3 个方面：一是对通用词预先进行过滤；二是用主题概念频度代替词频；三是用主题概念命中的科技资源数量代替词汇命中的科技资源数量。第一项改进是考虑到通用词对科技资源没有实际代表性，比如“方法”“数据”“设备”等。第二项改进是将同一主题概念的所有主题概念的词频作为这个主题概念的词频，比如一项科技

资源中出现了“色谱仪”“色谱系统”和“色谱分析仪”3个主题概念，在合并为一个主题概念“色谱仪”后，那么主题概念“色谱仪”的频度就是“色谱仪”“色谱系统”和“色谱分析仪”3个主题概念词频的总和。第三项改进的含义与第二项改进类似，主题概念“色谱仪”命中的科技资源数是“色谱仪”“色谱系统”和“色谱分析仪”3个主题概念命中的科技资源数的总和。一个主题概念的主题代表性的计算公式如下：

$$S(t)=\sum_{j} tf_j \times \log \sum_{j} \frac{|R|}{df_j}$$

其中，t是指要计算主题代表性的主题概念，j是指c由多少个主题概念合并，tf_j是指t中每个主题概念的词频，df_j是指c中每个主题概念命中的科技资源数。这样就为每个主题概念计算出对应的主题代表性，也就是主题概念的权重。

3　科技资源的关联聚合

在为每一项科技资源抽取了主题概念后，就可以利用加权主题概念在任意两项科技资源之间判断是否存在关联以及关联的强度，然后基于建立好的关联对总体的科技资源进行聚合。判断两项科技资源之间是否存在关联的方法是这两项科技资源之间是否存在相同的主题概念。如果有相同的主题概念，则这两项科技资源存在关联，否则不存在关联。对于存在关联的科技资源，关联的强度是不同的，因为不仅相同主题概念的数量不同，而且主题概念的权重也是不同的。因此，对于存在关联的两项科技资源，进一步计算它们之间的关联强度，计算公式如下：

$$L(r_i,r_j)=\sum_{k=1}^{m} w_k$$

其中，科技资源r_i和r_j有m个相同的主题概念，第k个主题概念的权重为w_k。从该公式可以看出，科技资源之间的关联强度与相同主题概念的数量以及主题概念的权重成正比。

根据科技资源之间的关联强度就可以对科技资源进行聚合，关联强度越大，科技资源关系越密切。为了达到多个关联强度大的科技资源聚合在一起的目的，本文采用力导向算法对所有的科技资源进行同步聚合。在力导向算法中，如果两个科技资源不存在关联，就表现为一个斥力常量；如果两个科技资源存在关联，就表现为一个引力变量。变量值为这两个科技资源之间的关联强度。算法的主要步骤如下：

（1）对每一项科技资源执行步骤2—步骤4。

（2）计算当前科技资源受到的斥力合力。

（3）计算当前科技资源受到的引力合力。

（4）计算当前科技资源受到的总合力。

（5）根据总合力计算当前科技资源位置偏移量，包括方向和位移距离。

（6）对步骤1—步骤5经过多次迭代，直到每项科技资源的当前总合力为0。

力导向算法可以使每项科技资源尽可能与其关联强度大的科技资源聚合在一起。聚合在一起的科技资源在主题上彼此非常接近。

4　实证分析

本文依据真实的科技资源元数据，对基于内容特征的科技资源关联聚合的效果进行实际验证。科技资源元数据的数据集由中国科技资源共享网[18]提供，包括科学仪器、种质资源和科学数据3类科技资源，共随机选取1254条元数据记录。在实证开展前对数据集进行清洗，包括无效值缺失值处理、格式统一、一致性检查等。实证所需要的分词工具为开源的HanLP。HanLP能够很好地支持中文分词，并允许用户自定义词典，把专业术语库结合进来。选择《汉语主题词表》[19-20]作为本实证的知识组织工具。《汉语主题词表》是我国目前最大的综合性中文主题词表，主要利用其中庞大的专业术语库和术语之间精准的同义关系。无论是分词工具、专业术语库还是同义词典都比较容易免费获取，因此本文提出的科技资源关联聚合方法可根据实际应用进行灵活弹性配置。

对随机选取的1254条科技资源元数据记录进行关联聚合，得到了如图2所示的聚合效果，其中科学仪器、种质资源和科学数据3类科技资

源分别使用不同几何形状区分（圆形节点表示科学仪器，矩形节点表示科学数据，菱形节点表示种质资源），当两项科技资源存在关联关系时用无向边连接。从图 2 中可以看出，这些科技资源比较明显地聚合成 5 个团簇，而且每个团簇中都聚合了不同类型的科技资源。经过人工对各个团簇细节的分析，团簇A主要是化工化学分析方面的科技资源，团簇B主要是物理特性测量方面的科技资源，团簇C主要是食品成分分析方面的科技资源，团簇D主要是地质矿产分析方面的科技资源，团簇E主要是地球环境测量方面的科技资源。在团簇内部，处于核心（中心）位置的科技资源与本团簇的其他科技资源关联密切，而与其他团簇的科技资源关联较弱甚至没有关联。由于对科技资源聚合的目的是为了向用户提供更好的协同服务。因此，利用关联聚合的结果：一是可以提高团簇中核心科技资源的优先服务级别；二是用户确定使用一项科技资源时可以推荐该项科技资源关联密切的其他项科技资源，尤其是不同类型的科技资源。

为了更清楚地展示关联聚合的效果，图 3 对团簇E进行放大，并隐藏科技资源之间的连接，

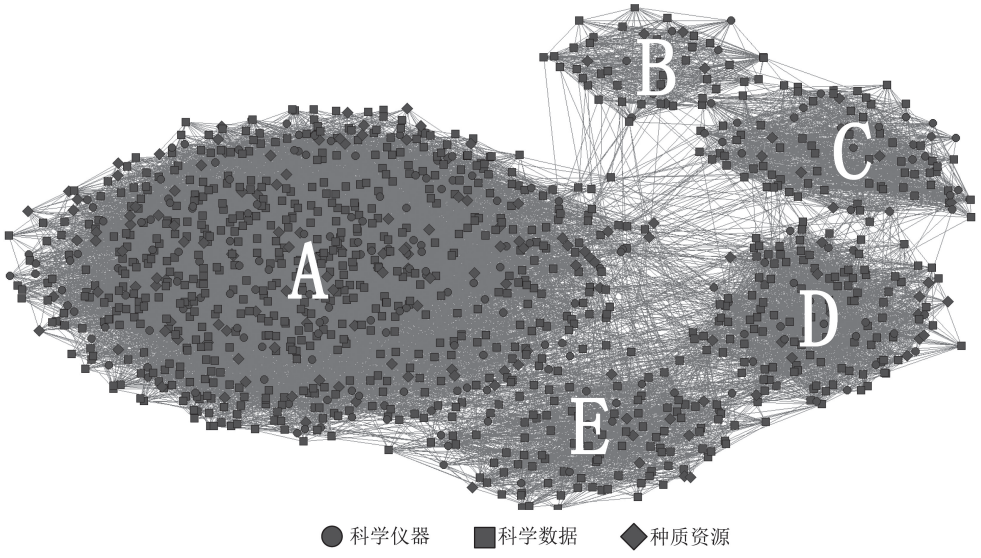

图 2　对科学仪器、种质资源和科学数据 3 类科技资源的关联聚合的效果展示

图 3　团簇E放大后的细节展示

展示团簇的内部细节。团簇E内的科技资源主要是地球环境测量相关的科技资源，高权值的主题概念主要有生态、气候、土壤、有机碳、降雨量、植被等，其中的科技资源既有科学数据“华东地区30m森林分布数据”等，也有种质资源“氮中二氧化碳气体标准物质”等，还有科学仪器“双频激光干涉仪”等。不同类型的科技资源可以很好地聚合在一起，比如“氧分析仪和氧气检测报警器检定装置”和“贵州省公里格网植被释放氧气量数据”。这样就可以为用户优化科技资源的检索排序并提供异类科技资源推荐服务等，达到科技资源协同服务的目的。

5　总结及展望

对科技资源关联聚合是解决当前科技资源松散孤立、使用效率低、重复浪费等问题的有效手段。由于科技资源元数据的多源异构、类型多等特点，给关联聚合带来了挑战。本文从共享服务的角度将科技资源元数据划分为内容特征元数据、外部特征元数据和服务特征元数据三类，基于内容特征元数据提出科技资源关联聚合的方法。首先从内容特征元数据中抽取主题概念并计算权重，然后利用加权的主题概念建立起科技资源之间量化的关联，最后根据关联强度实现对科技资源的聚合。在真实的科技资源元数据上验证了关联聚合的效果。

本文是从主题的角度对科技资源关联聚合，解决了不同来源、不同类别科技资源元数据表达不统一带来的挑战。在未来工作中，将开展基于服务特征元数据开展关联聚合的研究，丰富科技资源关联聚合的方法体系，更好地满足用户对科技资源协同服务的需求。

参考文献

[1] 周寄中. 科技资源论[M]. 西安: 陕西人民教育出版社, 1999.

[2] 刘玲利. 科技资源要素的内涵、分类及特征研究[J]. 情报杂志, 2008, 27(8): 125−126.

[3] 徐晓霞. 中国科技资源的现状及开发利用中存在的问题[J]. 资源科学, 2003, 25(3): 83−89.

[4] 袁伟, 赵辉, 石蕾.国家科技资源共享服务平台管理模式的熵效应分析[J].中国科技资源导刊, 2017, 49(4): 1−6.

[5] The Dublin Core Metadata Element Set: ANSI/NISO Z39. 85−2012[S]. Baltimore: NISO, 2013.

[6] 科技平台资源核心元数据: GB/T 30523−2014 [S]. 北京: 中华人民共和国国家质量监督检验检疫总局 中国国家标准化管理委员会, 2014.

[7] 宋佳, 高少华, 杨杰, 等. 科技资源元数据的关联与推荐方法[J]. 中国科技资源导刊, 2017, 49(5): 37−44, 103.

[8] 科技报告元数据规范: GB/T 30535−2014 [S].北京: 国家质量监督检验检疫总局, 2014.

[9] 浦燕妮, 刘琪, 耿骞. 通用型科学元数据标准研究[J]. 数字图书馆论坛, 2016(12): 33−39.

[10] Darwin Core [EB/OL]. [2020−03−30]. http: //rs.tdwg.org/dwc/index.htm.

[11] 土壤科学数据元数据: GB/T 32739−2016[S]. 北京: 中华人民共和国国家质量监督检验检疫总局 中国国家标准化管理委员会, 2017.

[12] 毕强,尹长余,滕广青,等. 数字资源聚合的理论基础及其方法体系建构[J]. 情报科学,2015,33(1):9−14, 24.

[13] 瞿辉,邱均平. 基于语义化共词分析的馆藏资源聚合研究[J]. 情报科学,2016,34(02):15−20.

[14] 尚珊,孟琦. 基于凝聚子群法的专家团队聚合[J]. 图书馆理论与实践,2014,(8):12−16.

[15] 周姗姗. 基于Folksonomy模式的数字资源多维度聚合研究[D]. 长春: 吉林大学, 2014.

[16] 马珉. 元数据: 组织网上信息资源的基本格式[J]. 情报科学, 2002(4): 42−44.

[17] MURTHA B. Introduction to metadata[M]// Transactions on High−Performance Embedded Architectures and Compilers I. Germany: Springer−Verlag, 2008.

[18] 中国科技资源共享网. 科技资源元数据的数据集[EB/OL]. [2020−03−30]. http: //www.escience.net.cn/.

[19] 中国科学技术信息研究所. 汉语主题词表: 工程技术卷[M]. 北京: 科学技术文献出版社, 2014.

[20] 中国科学技术信息研究所. 汉语主题词表: 自然科学卷[M]. 北京: 科学技术文献出版社, 2018.