情报学报　2018 年 11 月　第 37 卷　第 11 期
Journal of the China Society for Scientific and Technical Information, Nov. 2018, 37(11): 1077-1086

DOI: 10.3772/j.issn.1000-0135.2018.11.001

面向学科领域的学术文献语义标注框架研究

孙建军，裴　雷，蒋　婷

（南京大学信息管理学院，南京　210093）

摘　要　海量的学术文献为科研工作者的研究带来了困难。语义标注是实现学术文献的快速阅读和知识的快速获取的基础，因此，本文旨在构建一个面向学科领域的学术文献语义标注框架，以规范和丰富学术文献的标注体系。本文从三个方面进行了研究：一是学术文献标注本体的构建，二是学科领域本体的构建，三是标注本体与领域本体的关联实例。本文从学术文献内容定位、概念关联、方法流程标注及引文标注几个方面给出了标注的实例。

关键词　语义标注；学术文献；本体构建；标注本体

Research on Semantic Annotation in Academic Literature

Sun Jianjun, Pei Lei and Jiang Ting

(*School of Information Management, Nanjing University, Nanjing*　210093)

Abstract: Since the vast size of academic literature can present difficulties to researchers, semantic annotation is essential to rapid reading and knowledge acquisition. In order to regulate and enrich the semantic annotation system of academic literature, this paper focuses on the construction of an annotation ontology, the construction of a domain ontology of discipline domains, and the relationship between the terms in annotation ontology and domain ontology. This paper provides several instances of semantic annotation, such as labeling concepts, relationship of concepts, methods, processes and citations in academic literatures.

Key words: semantic annotation; academic literature; ontology construction; annotation ontology

1　引　言

20 世纪 80 年代起，随着互联网及计算机软硬件的发展，数字出版的基础设施逐步发展成熟，数字学术出版物应运而生，而随之带来的是数字学术出版物在数量上呈现爆发式增长。2015 年《STM 报告：科技及学术期刊出版概述》指出：截至 2015 年，CrossRef 数据库包含超过 7100 万个 DOI 号，Google 学术索引了 1 亿~1.6 亿的学术资源（包括期刊文献、书籍和灰色文献），Web of Science 数据库中包含了约 9000 万条记录；截至 2017 年 9 月，《中国学术期刊（网络版）》共收录接近 5000 万篇中文学术文献。在这种背景下，学术交流产生了重大的变革。

研究者可以从网络文献数据库中获取到大量的学术文献，这为研究者的研究工作提供了非常好的基础，但同时如此大体量的资源为学术工作的展开也带来了困难。首先，新概念的产生或者新涉足某一领域时，研究者需要学习大量的已有知识才能跟上现有的研究进展。而且，研究者的时间是有限的，获取到的文献越多，分配到单篇学术文献阅读的时

收稿日期：2018-06-08；修回日期：2018-08-14

基金项目：国家社会科学基金重大招标项目“面向学科领域的网络信息资源深度聚合与服务研究”（12&ZD221）。

作者简介：孙建军，男，1962 年生，博士生导师，教授，主要研究方向为网络资源管理，E-mail: sjj@nju.edu.cn；裴雷，男，1981 年生，副教授，主要研究方向为信息政策分析、信息资源管理；蒋婷，女，1988 年生，博士研究生，主要研究方向为网络信息资源管理。

间则相应减少，Tenopir 等[1]的研究就证实了这一假设研，研究者阅读文献不再是阅读全文，而是获取感兴趣的内容进行阅读：研究者通过浏览许多文章的部分来寻找、评估和利用一系列的信息[2]，这种阅读方式也被称作碎片化阅读。因此，第一个问题就是如何快速定位到文章的有用部分。另外，学术文献中的知识元存在大量的关联性，如引文关联、相关概念等，如何组织这些相关的知识元是研究者面临的第二个问题。

因此，Renear 等[3]提出了“策略阅读”的概念，采用学科本体来表示及链接科学数据可以提高研究者阅读学术文献的效率，即需要利用学科本体对学术文献中的相关内容进行语义标注（Semantic Annotation）。语义标注就是将本体或元数据中的概念与资源建立联系的一个过程。其中，语义标注的核心是学科领域本体，本体最广泛的定义是“本体是概念模型的明确的规范说明”[4]，它可以灵活地定义事物结构，以元数据的模式，提供概念受控词表，每个概念都包括一个明确定义的机器可理解的语义，且概念与概念之间的关联也显式地进行了定义，这样的结构能够让计算机进行推理应用。

学术文献的语义标注就是借助领域本体，将学术文献中的相关内容与本体中的知识元（概念或关系）进行链接，当读者需要获取文献中知识元对应的描述时，可以借助语义本体从对应的知识库中进行获取。例如，Textpresso[5]就是一个与本体关联的数据挖掘系统，它所包含的学术文献集依据本体中的术语分为了33个类别，用户输入一个或多个标记或关键词集合就可以定位到学术文献中特定的句子，并可获取本体中词对应的含义，支持语义查询。预先对学术文献的结构、内容或引文信息进行标注后，读者可以通过这些标注信息快速定位到文章的部分内容实现“策略阅读”。

目前，已有一些研究针对资源语义标注框架提出了标注本体的概念，标注本体旨在针对学术文献提出一个规范的本体框架，进而采用标注本体中的概念对学术文献的内容进行标注。目前已有的标注本体有 PAV[6]、PROV-O[7]以及 AO[8]本体等。其中，PAV 本体用于获取数字科技资源的出处、作者以及版本信息，用以区别资源被获取、转换以及消费的过程；PROV-O 是 W3C 小组制定的用于统一资源交换的本体；AO 本体提供了用于标注生物医学领域科技文献的概念及关系。

但是，现有的研究主要集中在标注本体的制定上，而如何对学术文献进行标注的研究比较少。为了实现学术文献的语义标注，首先需要明确学术文献所包含的知识元类型，在继承已有标注本体的基础上构建一个面向学术文献标注的标注本体，除了包含学术文献的一些标准元数据信息（作者、创建者、创建时间）以外，还包括了学术文献中的主题、发现、方法论等；其次，需要构建一个与某一学术领域相关专业术语的领域本体，包含该领域的概念及概念间的关联；最后，要将学术文献中的内容与本体中的概念一一对应，从而可以通过标注信息实现文献的快速浏览，也可以通过 URI 对相应概念作进一步了解。

因此，本文旨在构建学科领域学术文献语义标注框架，提出适用于学术文献语义标注的标注本体，以及针对学术文献具体内容（如引文信息、内容信息等）进行语义标注的方法。本文提出的学术文献语义标注框架也是实现文献语义检索的基础，通过语义标注，给予机器可以理解的语义，让使用者更方便更有效地利用学术文献，另外，提出的学术文献标注本体，可以被其他标注本体进行继承和扩展，具有较高的实践价值。

2 相关研究

2.1 学术文献语义标注方法相关研究

学术文献标注主要有两种方法：一是社会标注，研究者在学术文献阅读过程中使用辅助阅读或管理的软件进行标注；二是采用机器自动进行学术文献的标注。

社会标注，即 folksonomies，目前已有一些面向学术文献的标注软件，如 Utopia、Mendeley，这些软件可以自动获取到文章的一些元数据信息，如题名、作者、摘要、DOI、URL 等，也可以获取读者的统计数据以及读者对文章内容的标注。这类软件有利于资源的分类和组织，标签可以提升检索效率，也促进了以同一兴趣标签的社交网络生成。但是社会标注有一些缺陷，不同的表达、词的歧义、不同粒度，都为标签的共享和重用带来困难。

机器学习方法进行标注可以减少人工标注的成本。Boella 等[9]提出了一种结合语言学及机器学习的方法来进行语义标注，语言学方法主要依赖于 POS 标注以及句法分析，再将这些元素转化为特征集，

采用支持向量机来对文本进行语义标注。段宇锋等[10]结合朴素贝叶斯和弱监督学习方法 Bootstrapping 来迭代学习和标注中文物种领域的文本。Vidal 等[11]提出了一种基于图的方法来对 e-Learning 领域的教学资源文档进行标注，每个相关术语链接到本体中的子图，这一扩展过程中，排除与文档主题不相关的信息，因而有一系列本体子图标注文档，最后取这些本体子图的交集作为文档的语义标注。

2.2　学术文献元数据或标注本体相关研究

目前，针对资源描述出现了一些元数据以及标注本体。

都柏林核心元数据适用于描述和管理数字资源及馆藏资源，包括题名、创建者、主题及关键词、说明、出版者等 15 个广义元数据。PROV 本体（PROV-O）是针对不同系统不同内容生成的信息进行表示、交换或集成的本体，由 W3C 小组开发、管理和维护。PAV 本体是用于获取网络资源的出处、作者以及版本信息的本体。标注本体（AO）是与标注相关的本体，包括评论、实体标注（或语义标注）、文本标注（经典标记）、笔记等用于部分或全部电子文档（文本、图片、声音、表格等）的标注信息。SWAN[12]本体描述了艾滋海默症领域的知识，它作为一个知识支撑系统能够有效地支持艾滋海默领域研究，并且它与 SIOC 本体进行了本体对齐，为不同粒度级别的科学论述的表示提供了一个完整的模型。

SPAR 本体是用于描述出版领域的本体，它为语义出版和引文提供了一套可以机读的 RDF 元数据集，包括文档的描述，文献目录识别，引文的类型和相关内容，书目引文，文档的部分及状态，个体的角色及贡献，文献计量学数据及工作流程。SPAR 本体包括下述子本体：FaBiO 是用于描述出版或者潜在出版实体的本体；CiTO 是一种引文本体，用来描述引文的特性及类型，并允许标注者标记引文链接和引用意图；BiRO 是用于描述书目记录及参考文献的本体；C4O 是用于描述参考文献引文的本体，如文本内部参考文献指针、文本被引用文献引用的次数等；DoCO 提供了文档结构元素的词表，如段落、节或列表等；PSO 是用于描述文件出版状态或者出版过程中不同阶段的出版实体的本体，如提交、审稿中、拒稿、接收等；PRO 是用于描述个体出版过程中（如作者、编辑、评审等）的角色的本体；PWO[13]是用于描述出版实体在出版过程中的步骤的本体，如文章在审稿中、印刷、发表等；DEO 为文件中的修饰元素提供了一个结构化的词表，如引言、讨论、致谢、参考文献列表、附录等；SCoRO 是用于描述学术贡献及角色的本体；FRAPO 是用于描述研究项目信息的本体，如拨款申请、资助机构、项目合作者等；BiDO 是用于描述文献数据中数字和分类的模块本体，如期刊影响因子、作者 H-指数、研究类型分类等；Five*是描述网络期刊文章中五种属性的本体。

对于描述学术资源的数据的规范，学术文献语义标注本体可在继承现有元数据和标注本体的基础上加以扩展。

2.3　学术文献内容提取相关研究

目前，一些研究针对学术文献中的元数据元素、文献结构以及引文的提取提出了方案。

Constantin 等[14]设计了基于规则的系统 PDFX，利用设计的规则和特征集进行了元数据的抽取以及标注文本片断。Kovriguina 等[15]研究采用规则和模板匹配的方法从会议文献中提取元数据。

Tkaczyk 等[16]主要采用启发式规则及支持向量机方法实现了基本结构抽取，采用支持向量机以及简单的规则进行元数据抽取，采用支持向量机及条件随机场模型实现了引文抽取。Han 等[17]研究了采用支持向量机进行学术文献元数据（包括题名、作者、作者机构、作者地址、致谢、版权、引文、Email、出版时间、摘要、引言、联系方式、关键词、URL、程度、出版号、页面范围等）的抽取，该方法通过预测类标签进行迭代收敛来提升分类效果，再通过查找每行的块边界来进行元数据抽取。另一个采用支持向量机的学术文献元数据抽取方法是 Kovačević 等[18]提出的 CRIS 系统。

引文内容是学术文献中引用的与之相关的资源。Körner 等[19]采用线性条件随机场实现了参考文献字符串的抽取。目前的研究主要是基于规则、模板和一些学习方法，但是这些方法主要是基于领域内一些手工提取的特征，为了突破这一限制，An 等[20]探索了采用序列标注的深度神经网络模型进行引文元数据抽取。

2.4　相关研究综合述评

目前，针对学术文献或学术资源进行语义标注的研究主要还是基于人工标注的方法，通过设计标

注软件或网页，利用社会标注方法来对学术文献进行标注，这类标注方法目前仅对学术文献的一些信息（如关键词、作者信息、题名等）作浅层标注，并未涉及针对学术文献进行语义标注。相关研究主要集中在对标注本体的设计、学术文献中一些信息（包括关键词、结构信息等）进行机器自动抽取的研究。而如何实现对学术文献的语义标注，即如何将学术文献中的内容与领域本体进行语义关联，目前的研究较少涉及，因此，本文提出了学术文献的语义标注框架。

3 学科领域语义标注框架

整个学术文献的标注流程如图 1 所示，本文提出一种适合手工标注与机器自动标注的学科领域语义标注框架。整个语义标注流程如下：首先，计算机按照标注本体或者元数据对文献中对应项进行读取并预处理，抽取出对应的信息，将计算机确定的条目（如期刊信息、作者、题名等）抽取结果存入知识库；其次，针对标注本体中的一些复杂项（如文献主题、关键词、研究问题、解决方案等），计算机将领域本体中的概念或关系与对应的抽取项进行关联，并将抽取结果提交给专家（或用户）选择标注；有文献阅读需要的用户在阅读文献后，可以利用标注工具对文献中的某些标注项进行标注，对计算机的抽取结果进行筛选，最后标注工具将用户的标注结果提交到知识库。

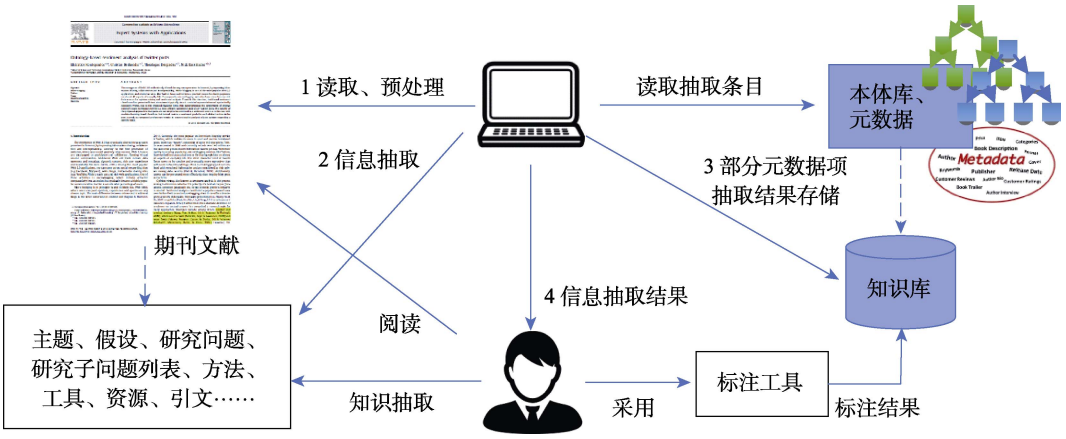

图 1 语义标注流程图

其中，知识库包括本体库、标注元数据以及文献语义标注后的知识，知识库可以辅助机器对文档的自动标注与标注人员的手工标注。本体库中包含标注本体和学科领域本体，例如生物医学领域已有较多的本体，在本体标注时可以借助已有本体进行使用，然后其他一些领域并不具备这种条件，没有已构建好的本体，因此，需要采用本体自动构建技术对该领域进行构建。

为了实现学术文献的语义标注，整个语义框架完成实现以下内容：

（1）面向学科领域的学术文献，从获取目前已有的元数据和标注本体出发，整理学术文献相关的概念和属性，构建标注本体。

（2）构建学科领域本体，获取学术领域的专业术语及关系。

（3）将学术文献中标注中的标记与学科领域本体中的概念关联。

3.1 标注本体设计

3.1.1 标注本体设计的目标

标注本体是为了描述需要标注的对象（即学术文献）中的元素、结构等信息，包括学术文献中的一些元数据，如题名、作者等；学术文献中的一些科学论述，如陈述、假设等；学术文献中的结构元素，如章、节、段落等；学术文献中的引文信息，如引用次数，引用文献等；某个领域的学术文献中的概念（专业术语）与关系。

3.1.2 标注本体中的元素

本文的标注本体继承了一些已有的本体（SPAR本体、FOAF、OA 等），并在此基础上进行了扩展。标注本体主要包含以下几类概念及属性：

（1）学术文献相关的个体（Agent）：包括学术文献在撰写、出版过程、检索利用中相关的人、机构及软件。

（2）学术文献相关的实体（Entity）：不同种类的学术文献，学术文献中的科学论述以及学术文献中的结构部分，学术文献参考文献记录及引文信息。

（3）与学术文献相关的活动（Activity）：如撰写、修改、提交、印刷等活动，以及与这些活动相关的时间节点或时间区间。

图 2 展示了学术文献标注本体的概念层级结构。

3.1.3　学术文献相关的个体

在学术文献中，有一些个体作为参与者，如人、软件、组织及机构。针对这些元素，我们继承了部分 FOAF 中的类，以一篇期刊文献来举例，可获取文献的作者，其在引用另一篇期刊文献时，被引的文献中包含的作者姓名也可被获取，这些作者都作为 FOAF 本体中 Person 类的实例存在。

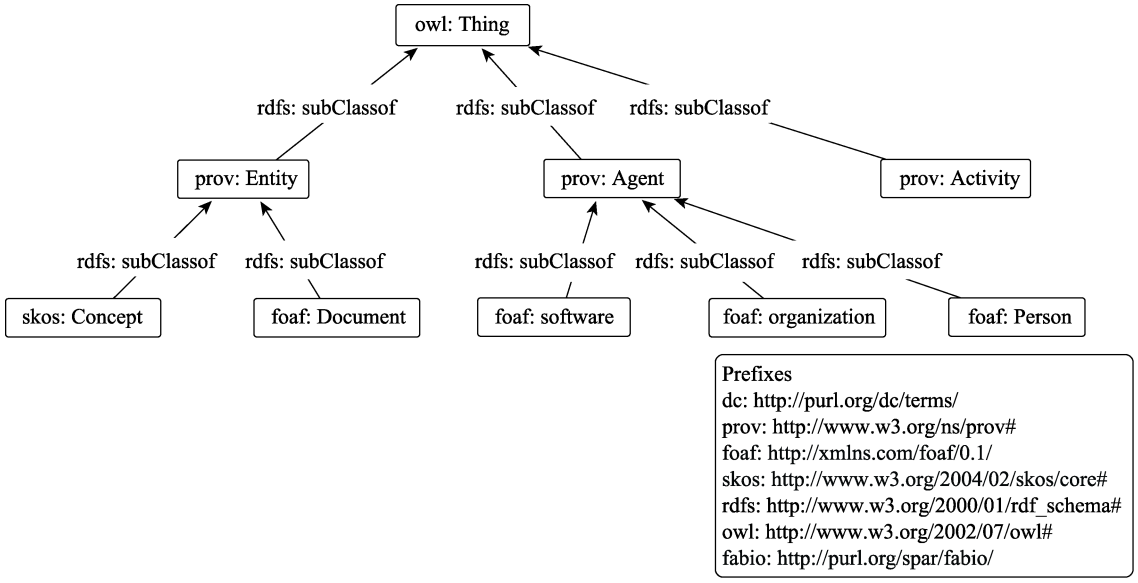

图 2　学术文献标注本体包含的概念

3.1.4　学术文献相关的实体

1）学术文献的种类

本文对学术文献的种类进行了分类并总结，不同类型的学术文献的撰写规范、包含元素、结构、内容不一致，本文主要将学术文献分为：书籍、文章、报告、会议文章等 13 个大类，并在此基础上又进行细分，例如，文章又可以分为综述类文章、新闻类文章、杂志文章以及期刊文章。本文的学术文献类型继承了 Fabio 本体中的一些概念，其概念层次关系如图 3 所示。

2）学术文献的科学论述及结构元素

学术文献的科学论述元素是指单篇学术文献所提的观点，在国外称为 scientific discourse，包括断言、提出问题、假设、支持的证据以及它们之间的论证关系，每个科学论述元素可以与学科领域本体或者社会标注中的术语或者断言进行链接。断言在学术文献中通常指一些主观性比较强的言论，例如对某个术语下的定义。提出问题通常是一个研究或者实验开展的主题。学术文献中的参考文献及引文就为科学论述元素提供支持的证据。

学术文献中的结构元素是组成学术文献的部分，包括引言、背景、相关研究、方法、讨论、数据等期刊学术文献的结构，也包括前言、后记、附录等书籍修饰部分，以及章节、段落、句子等学术文献粒度。

为了本体的共享和重用，上述的元素继承了 doco 本体、deo 本体以及 fabio 本体，其主要概念层次关系如图 4 所示。

3）学术文献参考文献及引文元素

通常情况下，学术文献中的参考文献通常是与当前文章相关的研究，或者为学术文献中的论述提供证据。关于参考文献及引文元素主要包括参考文献集合描述、引用行为（其属性包括引用意图及情感倾向）以及引文计量。

关于参考文献及引文的集合、记录、列表等元素继承自 biro 本体中的概念及属性。

根据学者引用文献的意图可以将引用行为分为：作为权威描述引用、作为数据源引用、作为证据引用、作为潜在方案引用、作为推荐阅读引用、作为相关文章引用、作为原始文档引用、作为信息

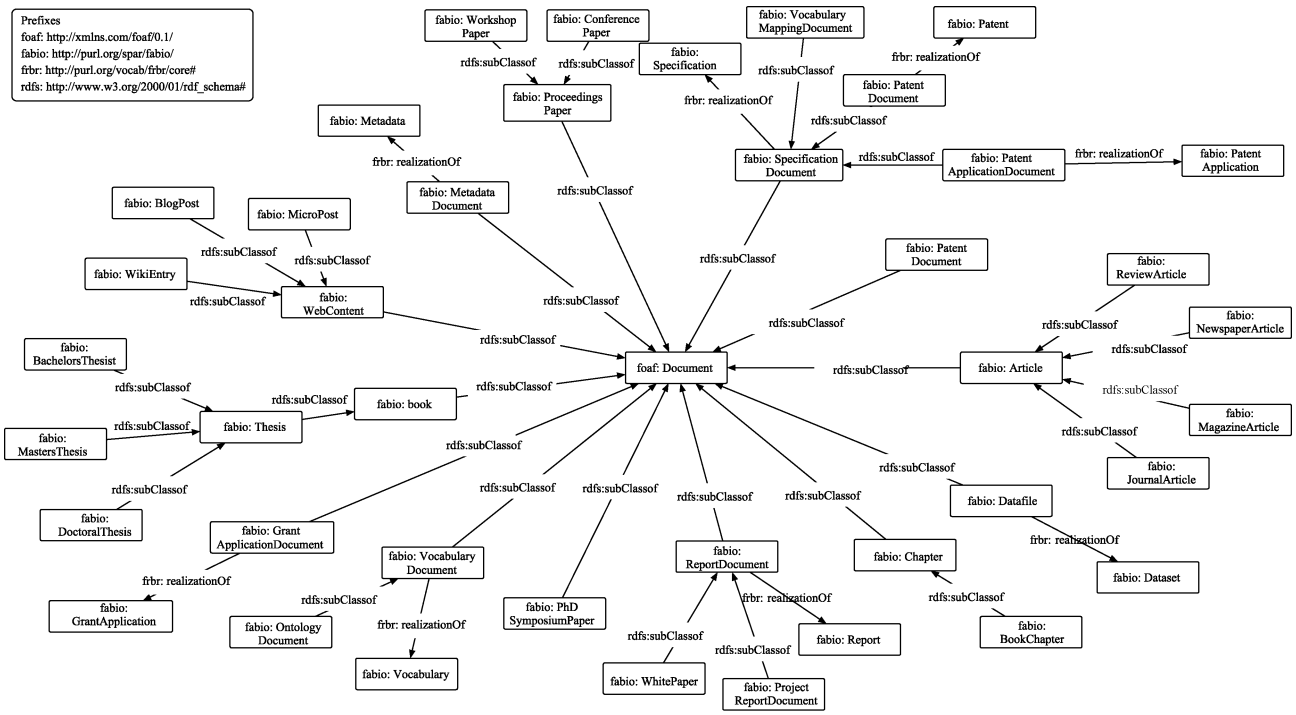

图 3 学术文献类型的概念层次关系

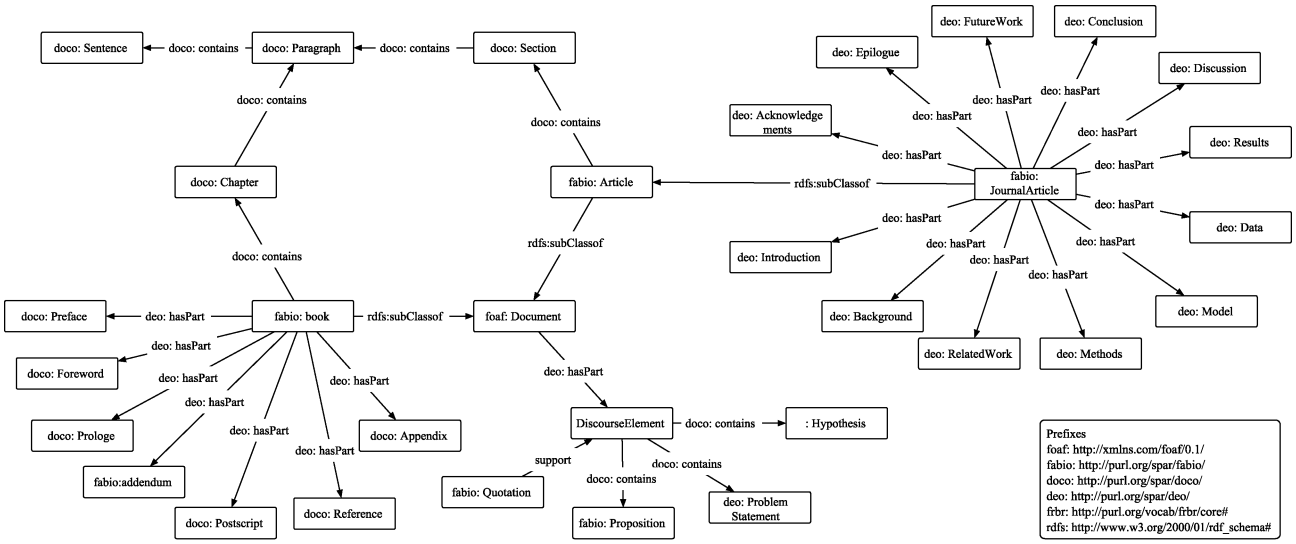

图 4 学术文献中的科学论述及结构元素图

源引用等几类。根据学者引用文献时对文献的情感倾向，可将引用行为分为：同意、不同意、认为正确、批判、嘲讽、奚落、驳斥这几类。这些概念及属性继承自 cito 本体中的一些概念及属性。

关于引文计量的概念及属性，如总被引次数，主要继承自 c4o 本体。

3.1.5 学术文献相关的活动

与学术文献相关的活动主要包括学术文献创造、加工、修改、使用过程中相关的活动，继承 PROV 本体中的 Activity 类。这些活动主要有作者生产、提交、修改、接受、退回、出版、预印本发布、发行、撤回、勘误等，主要继承自 Fabio 本体。

3.2 学科领域本体构建

为了将学术文献中的专业术语与学科领域本体中的概念相关联，首先需要构建学科领域本体，该领域本体中包含的概念是某一学科领域中的专业术语，这些术语也可以是领域词表中的术语转化而来，本节介绍一种学科领域本体半自动构建方法。

（1）定义需要获取学科的范畴，收集该学科领域相关的本体、词表，考虑复用的可能。

（2）获取领域内的术语：首先确定领域内术语的类型，如任务、方法、工具、资源这几个类别。收集领域内的语料，对语料进行文本转化、去噪、分词（英文语料包括词根化）、词性标注等预处理。结合语言学、统计学或机器学习方法自动地从语料中抽取术语，语言学方法需要按照领域内术语的规律提炼出词性模板，机器学习方法首先需要获取用于抽取术语的特征。最后抽取出领域内的术语。

（3）获取术语间的等级关系：首先定义一些等级关系的规则模板（例如，A 是一种 B，则 A 是 B 的子类），从网页或语料中获取到等级关系概念对，再利用基于图的方法获取等级关系图模型，最后利用图剪枝方法去除冗余的关系。

（4）获取术语间的非等级关系：针对领域内的知识，定义术语间非等级关系的类型（如部分-整体关系）；再到语料中获取具有非等级关系的三元组，采用统计学方法判定非等级关系三元组中概念对、动词与概念对之间的关联程度，取阈值内的非等级关系三元组；再提取特征，采用机器学习的方法判断提取三元组的类型。

最后对生成的本体进行评价，或者重复上述过程。

4　学术文献语义标注实例

学术文献的语义标注可以是手工标注或是机器自动标注，无论是采用何种标注方法，均是对学术文献或者其中某一部分，添加注释或者进行语义链接。本节针对学术文献中语义标注的常见类型进行区分，并给出学术文献语义标注的实例。

根据 OA 本体[21]中的规定，标注实例可以描述为类 oa: Annotation 的成员（实例），包含标注主体（oa: hasBody）以及标注对象（oa: hasTarget）。针对标注实例，可以添加相关描述，如标注者、创建时间等，标注者是 FOAF 本体中 Person 类的实例，如图 5 左部分所示。同时，可以对标注动机进行描述，本文继承了 OA 本体中的 oa: motivatedBy，这些动机有评论、描述、分类、链接、标注等。

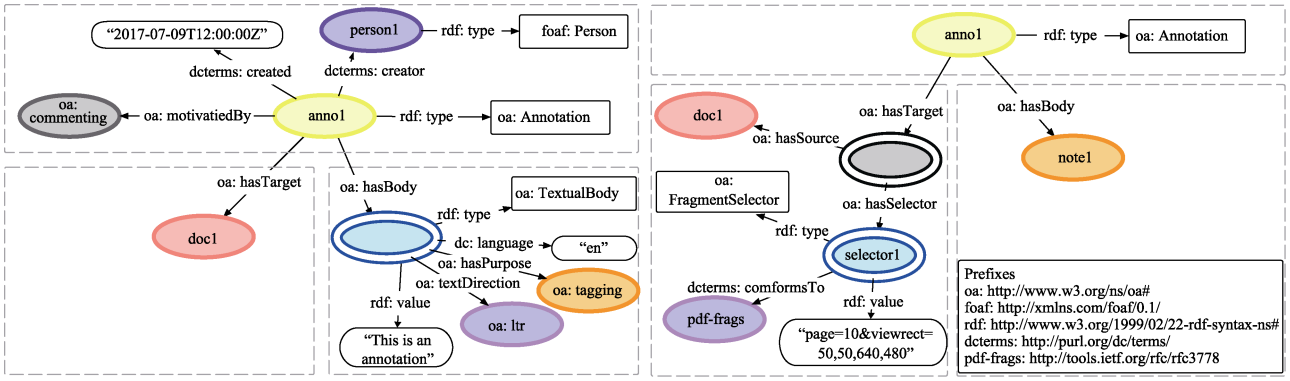

图 5　学术文献标注本体标注实例示意图

标注对象是指学术文献语义标注实例中需要进行标注的对象，可以是整个学术文献或其部分。学术文献的部分可以是学术文献中的论述元素、某个结构部分甚至是一个句子、一个词语。例如，图 5 中右图采用 OA 本体中的片断选择器指示到 PDF 学术文献中的部分片断；又如，文本类型的文档，可以定位到文本中的某个位置的字符中间的片断或者某个具体的词，如图 6 所示。

标注主体是标注本身，可以是一个文本类型的注释，如图 5 左部，还可以对标注主体进行描述，如文本方向、标注目的、语言、标注类型、值等。除了针对学术文献进行注释以外，还可将学术文献的部分与领域本体或社会标注中的概念或专业术语进行关联。将文章的术语、论述元素、结构片断或者全文链接到领域本体或者社会标注中的一个术语或概念。例如，图 7 左部分将学术文献与领域本体中的一个主题词术语进行了关联，表明该术语是学术文献的主题词，右部分将学术文献中的术语与领域本体中的一个概念进行了关联。

对学术文献的方法流程进行语义标注时，本文将其作为流程类的一个实例，继承 pwo 本体中的相关概念及属性，流程中所含的步骤单独标出作为步骤 pwo: Step 类的实例，如图 8 所示。

在对学术文献的引文进行标注时，标注实例中标注主体为引用文献指向被引文献的引用，文本内的引用指针作为标注对象。标注学术文献中引文的

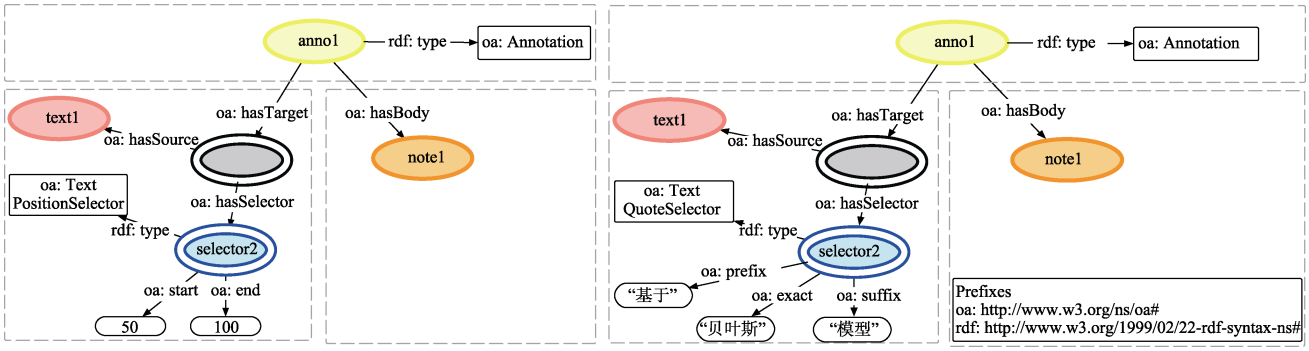

图 6 学术文献标注对象位置选择及词定位示意图

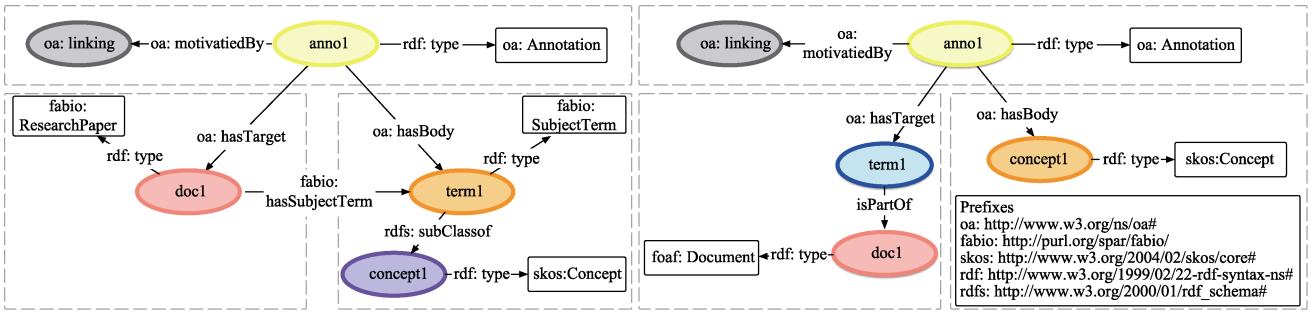

图 7 学术文献中术语与领域本体或社会标注中的本体链接示意图

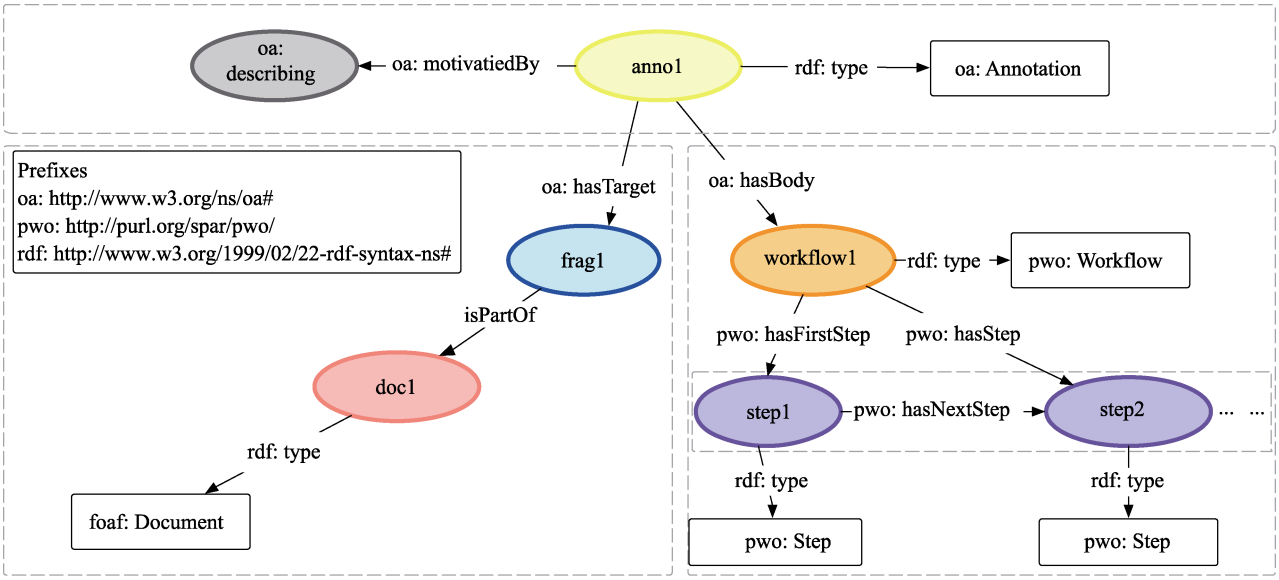

图 8 学术文献中方法流程的标注概念及属性示意图

情感时，继承 cito 本体中的类和属性，标注主体为文本，采用类 cnt: ContentAsText 进行描述，标注对象是类 cito: CitationAct 的实例，如图 9 所示。

5 结论与展望

针对学术文献进行语义标注是将学术文献中有意义的单元进行语义化组织的过程，有利于实现学术文献“策略阅读”的目标。为了实现学术文献语义标注，本文从三个方面来进行研究：一是研究学术文献的知识类型、结构等信息，在继承现有的标注本体元素的基础上构建了学术文献语义标注本体；二是获取学科领域的专业术语和关系，构建领域本体；三是将标注本体、领域本体中的概念与学术文献中的知识元相关联，并且本文针对学术文献标注过程中的片断选择、概念关联、方法流程及引文标注等内容给出了语义标注实例。

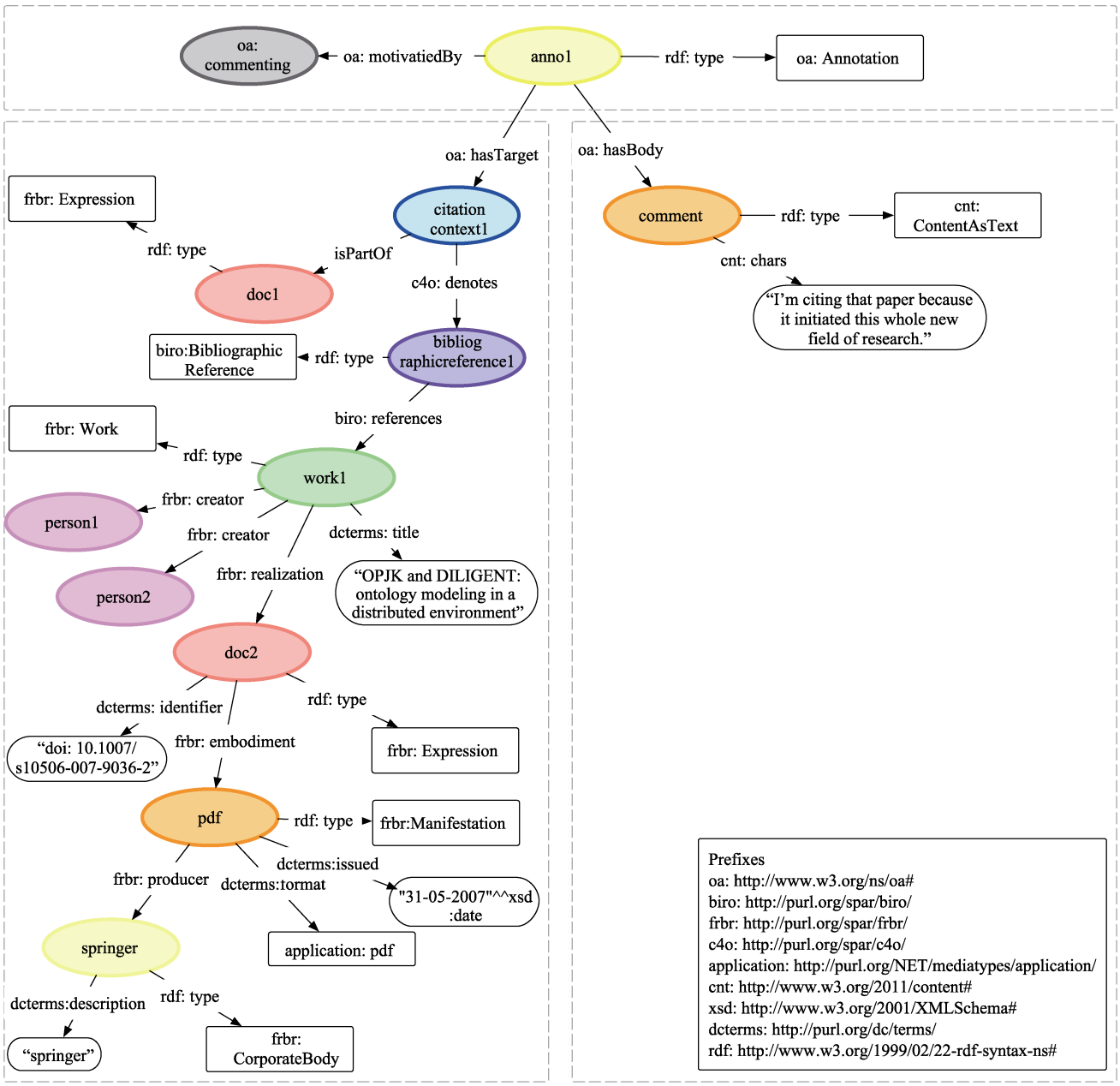

图 9　学术文献中引文标注相关的概念及类示意图

本文提出的学术文献语义标注本体可由各类使用者（包括科研人员、读者、期刊编辑等）进行学术文献标注，标注后的数据可以采用一些本体查询语言（如 SPARQL）或者一些推理机（如 SWRL）进行查询或推理应用，该标注本体继承了现有的本体类和关系（如 AO 本体、SPAR 本体等），具有可扩展性。此外，本文提出了学科领域本体自动构建的方法，能够实现领域专业术语及其关系的自动抽取。

实现学术文献的语义标注不仅包括标注本体、领域本体的构建，还需要针对学术文献中的知识元进行获取，将其与标注本体、领域本体中的概念及关系进行对应。现有的方法主要采用用户手工标注的形式，在未来的工作中，我们将研究学术文献中知识元的自动抽取进而实现机器自动语义标注。并解决领域本体体量过大时，利用本体模块化的思想来自动标注学术文献。此外，我们将扩展学科领域本体并开发与学科领域语义标注的系统，方便科研人员快速有效地利用学术文献和获取领域知识，以便实现学术领域的知识共享。

参　考　文　献

[1]　Tenopir C, Wang P L, Zhang Y, et al. Academic users' interac-

tions with ScienceDirect in search tasks: Affective and cognitive behaviors[J]. Information Processing & Management, 2008, 44(1): 105-121.

[2] Rowlands I, Nicholas D, Jamali H R, et al. What do faculty and students really think about e-books?[C]// Aslib Proceedings. London: Emerald Group Publishing Limited, 2007: 489-511.

[3] Renear A H, Palmer C L. Strategic reading, ontologies, and the future of scientific publishing[J]. Science, 2009, 325(5942): 828-832.

[4] Gruber T R. A translation approach to portable ontology specifications[J]. Knowledge Acquisition, 1993, 5(2): 199-220.

[5] Müller H M, Kenny E E, Sternberg P W. Textpresso: an ontology-based information retrieval and extraction system for biological literature[J]. PLoS Biology, 2004, 2(11): e309.

[6] Ciccarese P, Soiland-Reyes S, Belhajjame K, et al. PAV ontology: provenance, authoring and versioning[J]. Journal of Biomedical Semantics, 2013, 4: 37.

[7] Lebo T, Sahoo S, McGuinness D, et al. Prov-o: The prov ontology[J/OL]. W3C Recommendation, 2013. [2017-09-15]. http://www.w3.org/TR/prov-o/.

[8] Ciccarese P, Ocana M, Garcia-Castro L J, et al. An open annotation ontology for science on Web 3.0[J]. Journal of Biomedical Semantics, 2011, 2: 2-4.

[9] Boella G, Di Caro L, Ruggeri A, et al. Learning from syntax generalizations for automatic semantic annotation[J]. Journal of Intelligent Information Systems, 2014, 43(2): 231-246.

[10] 段宇锋, 朱雯晶, 陈巧, 等. 朴素贝叶斯算法与 Bootstrapping 方法相结合的中文物种描述文本语义标注研究[J]. 现代图书情报技术, 2014(5): 83-89.

[11] Vidal J C, Lama M, Otero-García E, et al. Graph-based semantic annotation for enriching educational content with linked data[J]. Knowledge-Based Systems, 2014, 55: 29-42.

[12] Ciccarese P, Wu E, Wong G, et al. The SWAN biomedical discourse ontology[J]. Journal of Biomedical Informatics, 2008, 41(5): 739-751.

[13] Gangemi A, Peroni S, Shotton D, et al. A pattern-based ontology for describing publishing workflows[C]// Proceedings of the 5th International Conference on Ontology and Semantic Web Patterns, Aachen, 2014: 2-13.

[14] Constantin A, Pettifer S, Voronkov A. PDFX: fully-automated PDF-to-XML conversion of scientific literature[C]// Proceedings of the 2013 ACM Symposium on Document Engineering. New York: ACM Press, 2013: 177-180.

[15] Kovriguina L, Shipilo A, Kozlov F, et al. Metadata extraction from conference proceedings using template-based approach[C]// Semantic Web Evaluation Challenge. Cham: Springer, 2015: 153-164.

[16] Tkaczyk D, Szostek P, Dendek P J, et al. Cermine–automatic extraction of metadata and references from scientific literature[C]// Proceedings of 2014 11th IAPR International Workshop on Document Analysis Systems. IEEE, 2014: 217-221.

[17] Han H, Giles C L, Manavoglu E, et al. Automatic document metadata extraction using support vector machines[C]// Proceedings of 2003 Joint Conference on Digital Libraries. IEEE. 2003: 37-48.

[18] Kovačević A, Ivanović D, Milosavljević B, et al. Automatic extraction of metadata from scientific publications for CRIS systems[J]. Program, 2011, 45(4): 376-396.

[19] Körner M, Ghavimi B, Mayr P, et al. Evaluating reference string extraction using line-based conditional random fields: A case study with german language publications[C]// Advances in Databases and Information Systems Cham: Springer, 2017: 137-145.

[20] An D, Gao L, Jiang Z, et al. Citation metadata extraction via deep neural network-based segment sequence labeling[C]// Proceedings of the 2017 ACM on Conference on Information and Knowledge Management. New York: ACM Press, 2017: 1967-1970.

[21] Sanderson R, Ciccarese P, Van de Sompel H, et al. Open annotation data model[R]. W3C Community Draft, 2013: 8.

（责任编辑 车 尧）