DOI:CNKI:11.2746.G2.20120928.1143.003.

面向关联数据的语义数字图书馆资源描述与组织框架设计与实现*

欧石燕

摘　要　本文提出了一个面向关联数据的语义数字图书馆资源描述与组织框架,该框架具有四个层次:元数据层、本体层、关联数据层和应用层,其核心是RDF语义元数据的构建与关联。以"图书、情报与档案学"领域的数据为例对该框架进行了实现,实现的重点是在本体层和关联数据层,包括相关本体的设计、普通元数据到语义元数据的自动转换、不同数据集间RDF语义链接的建立、关联数据的发布等。图5。参考文献46。

关键词　语义数字图书馆　关联数据　元数据　本体

分类号　G250

Design and Implementation of a Linked Data-oriented Framework for Resource Description and Organization in Semantic Digital Libraries

Ou Shiyan

ABSTRACT　In this paper, the author proposed a linked data-oriented framework for resource description and organization in semantic digital libraries, which includes four layers—metadata layer, ontology layer, linked data layer and application layer. The core of the framework is the construction and linking of RDF-based semantic metadata. Using the data from the domain of "Library, Information & Archives" as an example, the author focused on the ontology layer and linked data layer to implement the framework, which involves the design of related ontologies, automatic transformation of traditional metadata into semantic metadata, construction of RDF semantic links among different datasets, and publishing of linked data etc. 5 figs. 46 refs.

KEY WORDS　Semantic digital libraries.　Linked data.　Metadata.　Ontology.

1　引言

自20世纪90年代以来,数字图书馆这一综合研究领域在世界各地蓬勃兴起并取得了巨大发展。随着数字图书馆基础设施建设和遗留资源数字化工作基本告一段落,研究人员和从业者更多地开始关注如何保证在分布式的异构数字环境中,人们能够准确而全面地获得所需的信息与知识,这涉及到当前数字图书馆关于资源组织与利用的几个难题:①信息在局部范围得到组织但在整体上并不相互联系,形成了许多分散独立的信息孤岛;②无法对不同的信息系统实现统一的访问;③无法通过机器对信息进行语义层面的操作。

进入本世纪以来,互联网技术得到了突破

* 本文系国家社科基金项目"基于SOA架构的术语注册和服务系统构建与应用研究"(编号:11BT0023)和教育部人文社科基金项目"数据关联的语义数字图书馆研究"(编号:10YJA870014)的研究成果之一。

通讯作者:欧石燕,Email:oushiyan@nju.edu.cn

性进展，以语义网为核心的各种技术与标准的出现正在逐步影响并改变着当前的 Web 以及基于 Web 的各种应用，这其中也包括数字图书馆。将语义网技术应用到数字图书馆是国内外计算机和图情界近年来的研究热点。较早的研究主要偏重于采用 RDF、OWL、SKOS 等技术分别解决数字图书馆中的某些局部问题，如元数据、知识组织、信息检索等，当前的研究则更致力于探索如何利用语义网技术对数字图书馆中资源的描述、组织和检索等问题进行一揽子的解决，打造具有语义功能的语义数字图书馆，具有代表性的项目有三个：JeromeDL[1]、SIMILE[2] 和 Bricks[3]。所谓语义数字图书馆，是指以机器可读可理解的 RDF 语言为介质，能够集成基于不同元数据的各种信息，支持与其他数字图书馆或信息系统之间在通信层面或元数据层面的互操作，并提供具有语义功能的浏览和检索服务的数字图书馆[4]。当前几个语义数字图书馆的原型系统（如 JeromeDL）虽然实现了元数据的语义化描述，解决了元数据语义互操作问题，并支持语义检索，但是它们并没有真正成为语义网的一部分。首先人们对图书馆数据的访问需要通过 Web 应用程序接口（即 API）来进行，不同的数字图书馆系统拥有各自不同的访问界面，它们之间的互操作往往需要采用某种机制（如 OAI - PMH①）才能实现，因此无法在不同的数据集间建立无缝连接，从而像浏览 Web 文档一样通过链接的 URIs 地址在分布式的结构化数据之间进行冲浪。其次，虽然语义数字图书馆在一定程度上解决了语义互操作问题，但是这种互操作主要是针对图书馆的文献信息资源，还无法在不同的知识单元（如文献资源、知识组织资源等）之间建立显性链接来揭示它们之间隐含的各种相关关系，因此不同的知识单元是分散而独立地存在着。此外，即使在同一知识单元内部，也无法有效揭示资源之间的深层次关系，如相同、相关的资源等。关联数据的提出为上述问题的解决提供了现实和可能。

关联数据是由语义网创始人伯纳斯·李于 2006 年 7 月首次提出的一个概念，是指在语义网上发布、共享、连接各类数据、信息和知识的一种方式[5]。它以 HTTP 协议可参引的 URI 地址命名所有资源，以 RDF 语言语义化地描述资源，以 RDF 链接指向相关资源并揭示资源间的语义关系，是一种推荐的语义网最佳实践。2007 年至今，许多机构和研究者已经开展了众多的关联数据项目，如 DBPedia②、DBLP Bibliography③、GeoNames④ 等，将不同领域的结构化数据发布到网络上进行关联和共享，构成数据之网。虽然当前在数字图书馆中还没能实现全方位的数据关联与发布，但是已经有了关联数据的局部应用，有两个代表性案例：一个是瑞典国家图书馆实现书目数据的关联[6]，另一个是美国国会图书馆将其主题词表 LCSH 进行语义化描述后以关联数据的形式发布到 Web 上[7]。但是这两个项目都没有对关联数据之上的应用（如浏览和检索）做进一步的探索。

本研究的目的是构建一个数据关联的语义数字图书馆原型，实现对数字图书馆各种资源的语义化描述和语义检索以及全方位的数据关联，其核心是基于本体的元数据语义化转换和关联数据的构建与发布。该数字图书馆将具有以下功能和特点：

（1）实现文献资源的语义化描述和不同元数据类型间的语义互操作；

（2）实现图书馆知识组织资源（如受控词表、规范档等）的语义化描述；

（3）实现图书馆不同知识单元间资源的关联，使图书馆的资源组织由传统的基于主题的

① 全称 Open Archives Initiative Protocol for Metadata Harvesting，是用于收割基于 XML 的描述性元数据记录，实现不同信息系统间互操作的协议标准。

② 该项目将维基百科中的数据作为关联数据在 Web 上发布，见 http://dbpedia.org。

③ 该项目将 80 万个科学论文书目数据作为关联数据在 Web 上发布，见 http://www4.wiwiss.fu-berlin.de/dblp/。

④ 该项目将全世界超过 650 万个地名信息作为关联数据在 Web 上发布，见 http://www.geonames.org/。

层次化组织结构扩展到多方位、多层次的网络状组织结构；

(4)支持在网络上通过 RDF 链接浏览语义相关的资源，实现不同信息系统间或不同数据集合间信息的无缝过渡；

(5)实现对数字图书馆资源的统一检索和访问；

(6)支持语义检索和自然语言检索。

2 相关研究综述

本研究涉及数字图书馆和语义网领域的两个热门主题“语义数字图书馆”和“关联数据”。语义数字图书馆是由爱尔兰 DERI 研究所(Digital Enterprise Research Institute)的 Kruk 等人首先提出的一个概念[4]，是建立在传统数字图书馆、语义网、社会网络和人机交互研究之上的一个新事物。语义数字图书馆系统将传统图书馆中的知识组织系统与语义网和社会网络技术相结合，支持对信息的语义标注和与其他信息系统间的语义互操作，并允许用户参与到信息标注和知识共享中来，使信息发现变得更加容易。相对于普通数字图书馆，语义数字图书馆有两个主要优点：①提供了对信息空间新的搜索范式，如基于本体的搜索/分面搜索；②提供了数据层面的互操作，如集成各种不同来源的元数据，在不同的数字图书馆系统之间建立连接[4]。目前具有代表性的语义数字图书馆项目有 JeromeDL、SIMILE 和 Bricks。JeromeDL 是波兰 Gdansk 理工大学图书馆与爱尔兰 DERI 研究所合作进行的一个社会语义数字图书馆项目，它采用一个共享的书目本体 MarcOnt 作为中介实现不同类型元数据(即 Dublin Core、BibTeX 和 MARC21)的语义化转换以及它们之间的互操作，从而在同一个数字图书馆内部实现对各种资源的语义搜索和浏览[8]。SIMILE 是麻省理工学院、万维网联盟(W3C)和 HP 实验室联合研制的一个数字图书馆项目，其目的是支持和扩展 DSpace 数字资源管理系统，提高它对分布存储在不同地点和环境中的各类数字资产、概念体系(包括词表和本体等)、元数据之间语义互操作的支持[9]。通过对 RDF 和语义网技术的应用，SIMILE 提供了一系列用于转换、浏览、检索和映射异质元数据的工具，首先针对不同类型的元数据构建元数据本体，并在它们之间建立映射关系，然后依据各个本体对相应的元数据类型进行语义化转换，最后通过元数据本体间的映射关系实现不同元数据间的互操作[9]。此外，SMILE 还将不同类型的数据(包括数字资产的元数据、OCLC 人名规范档、维基百科中的人物生平信息)进行了关联，可以看作是关联数据的雏形；但是因为没有采用可参引的 HTTP URI 地址将关联的数据在 Web 上发布，还不能算作是真正的关联数据[9]。Bricks 是一个欧盟研究项目，目的是建立分布式文化遗产数字图书馆网络基础结构并实现互操作[10]。Bricks 与 SMILE 实现元数据语义互操作的方法大致相同，都是采用元数据本体间相互映射的方法，但是 Bricks 是采用 OAI - PMH 协议在不同数字图书馆系统之间实现互操作，而 SIMILE 则是在同一数字图书馆系统内部实现不同元数据间的互操作。

本研究除了属于语义数字图书馆范畴，也属于关联数据在图书馆领域的一种应用。关联数据自提出以来受到了计算机和信息领域的极大关注，许多个人和组织机构采用关联数据作为发布数据的一种途径，从而构成了一个称之为数据之网的全球数据空间。数据之网的出现源自于语义网研究社区的努力，特别是得益于万维网联盟(W3C)的关联开放数据项目(Linking Open Data)。至 2011 年 8 月，以关联数据形式在万维网上发布的数据集，即构成“关联开放数据云”(Linking Open Data Cloud)的数据集，已达 295 个，其中图书馆及其相关领域的关联数据集有 87 个，约占整个数据云的 9.33%[11]。图书馆拥有并一直在不断生成大量高质量的结构化数据，譬如书目数据、知识组织数据等，这些数据的发布、集成、发现是图书馆的核心工作之一，因此图书馆具有成为关联数据实践者和提供者的天然特性，可以利用关联数据发布资源，扩展资源发现服务，进行数据融合，促进异构数据的开放与复用，实现数字图书馆系统之间以

及与其他信息系统之间的集成等。

图书馆采用关联数据发布最多的是知识组织资源。在关联开放数据云中，具有代表性的词表数据有：美国国会图书馆发布的美国国会图书馆标题表 LCSH[7]，联合国粮农组织发布的多语言农业词表 AGROVOC[12]，OCLC 发布的杜威十进制分类法 DDC[13]，欧盟研究项目 TELplus 发布的法国国家图书馆主题词表 RAMEAU[14]，德国国家经济图书馆发布的经济学词表 STW[15]等。这些关联数据化的词表通常采用标准 SKOS 语言或 SKOS 标签扩展（SKOS - XL）语言表示，采用 RDF 存储器存储，支持基于 HTML 和 RDF 浏览器的浏览，并能通过 SPARQL 终端进行查询。图书馆发布的第二大类关联数据是书目数据，代表性项目是瑞典国家图书馆将瑞典联合书目 LIBRIS 发布为关联数据[6]，这是首个实现图书馆书目数据关联数据化的实例。2012 年 6 月，OCLC 将 WorldCat. org① 中的书目元数据发布为关联数据，是目前 Web 上最大的关联书目数据[16]。此外，RDF Book Mashup 提供了一种虚拟的书目数据关联数据化的发布和访问模式，它将来自多个不同 Web APIs 的书目信息集成到一个语义网界面中，其实质是通过构建一个包装器使得需要用户通过各个不同 Web APIs 访问的书目信息能够统一以关联数据的虚拟形式进行访问[17]。除了词表数据和书目数据，一些科技论文数据也被语义网实践者以关联数据的形式发布为数据之网的一部分。德国柏林自由大学和汉诺威大学的研究者采用 D2R② 服务器将著名的计算机科技文献书目数据库 DBLP 发布为关联数据[18-19]。英国南安普顿大学的研究者采用 RKB Explorer 将 DBLP 发布为关联数据[20]。RKB Explorer 是欧盟 ReSIST 项目开发的一个能够将来自多种异质数据源的数据进行集成并在语义网上统一发布的工具。除了 DBLP，RKB Explorer 还能够发布来自 Citeseer、ACM、NSF 和部分 IEEE 会议的学术资源。此外，爱尔兰和英国的研究者们共同开发了一个以关联数据形式发布的语义网学术会议资料库 Semantic Web Dog Food[21]。

我国对关联数据的认识比较晚。最早将关联数据介绍到国内的是上海图书馆的刘炜和华裔学者曾蕾[22-23]。自 2010 年起，国内出现了大量关于关联数据的论文[24-27]，但主要是对关联数据这一概念及其研究与应用现状进行介绍和综述，对关联数据应用与实践的研究还几乎没有。

3 语义数字图书馆资源描述与组织框架

在本文中，作者提出了一个层次化的语义数字图书馆资源描述与组织框架，将数字图书馆的资源描述、组织、发布和应用分为四个层次（见图 1）。该框架基于本体对图书馆的各种资源（如文献资源、知识组织资源、人名/组织机构名、地名等）进行语义化描述，采用关联数据原则发布数据，提供统一的数据访问机制，实现异构数据之间的语义互操作。

第一层是元数据层。在数字图书馆中，针对不同类型（如普通图书、学位论文、期刊等）、不同时期（如遗留资源、新建资源）、不同来源（如数字化的实体资源、网络资源）的文献资源一般采用不同的元数据规范（如 MARC、DC、BibTex 等）进行描述，这导致同一数字图书馆内部往往并存着多种元数据规范，不同数字图书馆之间使用的元数据规范更是千差万别。这些元数据规范之间可能存在着某些相似之处（如共享相同的核心元素），但并不完全兼容。此外，元数据主要是为人而设计的，元素的语义缺乏明确的、形式化的定义，无法利用机器的强大功能对元数据直接进行理解和处理。因此元数据虽然提供了数字图书馆的语义基础，但却无法解决资源描述的异构性和语义性问题[28]。

① WorldCat. org 是 OCLC 的全球图书馆和其他资料的在线编目联合目录，是世界最大的联机书目数据库。
② D2R Server 是一个将关系型数据库发布在语义网上的工具。

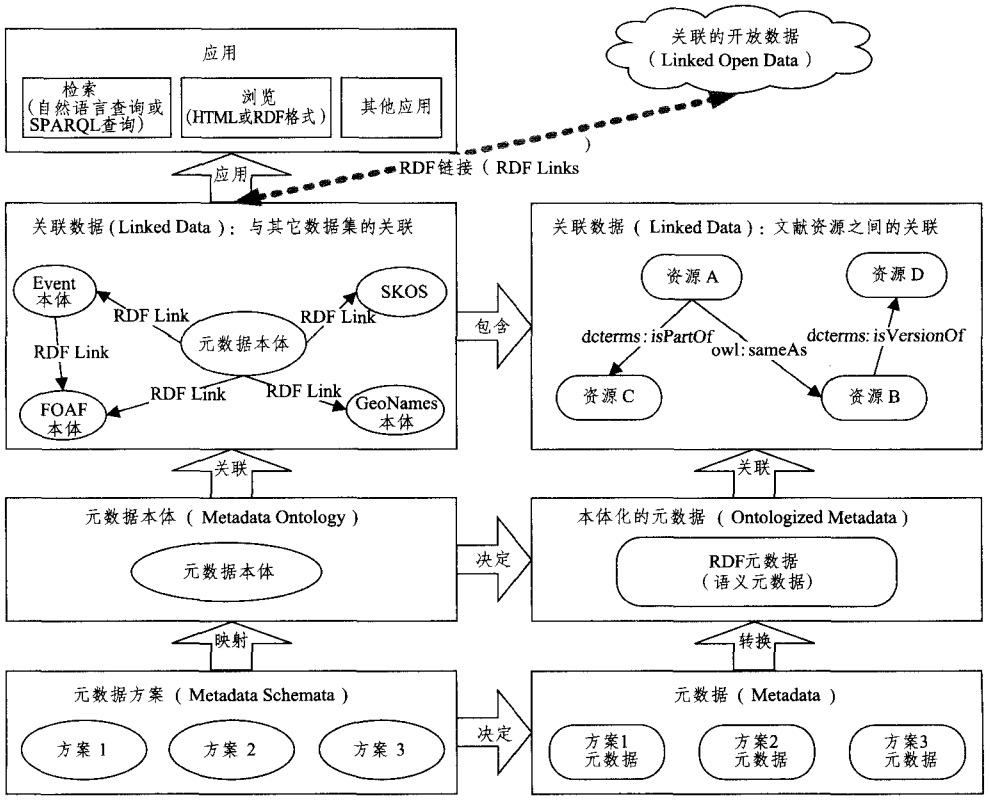


图1 语义数字图书馆资源描述与组织框架

第二层是本体层。鉴于元数据的上述局限性，需要在资源元数据描述的基础上构建某种机制，实现不同元数据类型和格式间的语义互操作，这就是本体层的作用。目前通过本体实现元数据的语义化描述和语义互操作，主要有两种方法：一种是对不同元数据规范中的概念和属性进行整合，采用本体描述语言（通常是OWL语言）构建一个集成元数据本体，如JeromeDL项目中的MarcOnt本体[29]，并基于该本体实现对各种元数据的语义化转换，转换成为统一的具有相同语义的RDF格式；另一种是采用本体描述语言对每种元数据规范进行本体化描述，并基于构建的元数据本体将相应的元数据转换成为RDF格式，然后通过不同元数据本体之间的映射关系，实现不同语义的RDF元数据之间的语义互操作，如SIMILE和Bricks项目中的做法。这两种方法各有优缺点。对于第一种方法，如果有新的元数据规范出现，必须修改和扩展集成元数据本体，使其能够容纳所有元数据规范中的概念和属性，因此灵活性比较差，但是在应用层面更容易实现不同类型元数据间的语义互操作。第二种方法更具有灵活性，因为当有新的元数据规范出现时，只需对它进行本体化并增加与其他元数据本体的映射，而无需改动已有的元数据本体及它们之间的映射关系，但是在应用层面实现元数据间的语义互操作比较复杂，需要借助OWL语言的推理功能。考虑到DC元数据是目前描述绝大多数文献资源的基本规范，作者综合上述两种方法的优缺点，提出基于DC元数据规范构建一个各种文献资源共享的核心元数据本体。之所以称之为核心元数据本体，是指该本体并不试图容纳各种元数据规范的所有元素，而是形式化地描述各种元数据规范所共有的核心元素（即DC元

数据元素)。特定文献资源类型(如会议论文)所特有的元数据元素(如所属的会议)或相互间关系(如会议论文与会议论文集之间的关系)可以动态地加入到核心元数据本体中来,通过对核心元数据本体进行定制化扩展生成针对特定文献资源类型的专门元数据本体。这样做的原因是既保证元数据本体具有灵活的适应性,可以针对不同类型的文献资源,又能使不同的元数据本体之间具有核心共享部分(即核心元数据本体),容易实现不同类型元数据之间的语义互操作。

第三层是关联数据层。虽然通过元数据本体,可以在语义层面上描述文献资源的元数据信息,并揭示它们之间的显性关系(如两个资源是整体和部分的关系),但是这些资源仅限于书目元数据,无法与图书馆中的其他资源(如知识组织资源)或外界的相关信息相沟通,也无法揭示资源间深层次或隐含的相互关系(譬如两个资源属于同一主题),更无法被读者直接浏览和访问①。因此作者提出采用关联数据的形式对本体化的元数据进行再组织,并采用关联数据原则在网络上进行发布。在类层面上,通过在不同领域的本体间建立联系,可以将图书馆不同知识单元的资源(包括书目元数据、知识组织资源、名称规范档等)在语义层面上相互关联起来;在数据层面上,可以将同一知识单元中的相同或相似资源进行关联,从而使图书馆中的各种资源构成一个有机联系的统一整体。为了实现数据关联,需要使用 HTTP 协议可解引用的 URI 地址命名每个资源,使用 RDF 链接连接相关的资源并语义化地揭示关系的类型(譬如作者关系、主题关系、相同的资源等),使用内容协商机制解决非信息资源(见 5.1 节)的访问问题,使得每个资源都可以使用浏览器通过 URI 地址进行直接访问,并可沿着 RDF 链接爬行从而访问其他相关资源,自由地在不同数据集中进行切换,使图书馆的各种数据构成一张立体的数据之网。此外,图书馆的关联数据还可进一步与其他图书馆的关联数据或其他领域的关联数据(如 DBPedia)相关联,成为整个数据之网的一部分,更容易被读者所发现。

第四层是应用层。在应用层需要实现的是对关联数据的统一浏览和检索以及其他语义互操作。除了提供传统的基于关键词的检索方式,还可进一步提供界面更为友好的问答式检索,使用户可以采用自然语言的提问在语义层面精确地表达自己的信息需求,并获得精确的查询结果。此外,通过利用元数据本体间的映射关系以及本体的推理功能,或者利用术语服务机制中的查询词扩展与精炼功能[30],能够更进一步提高检索的智能性与复杂性。

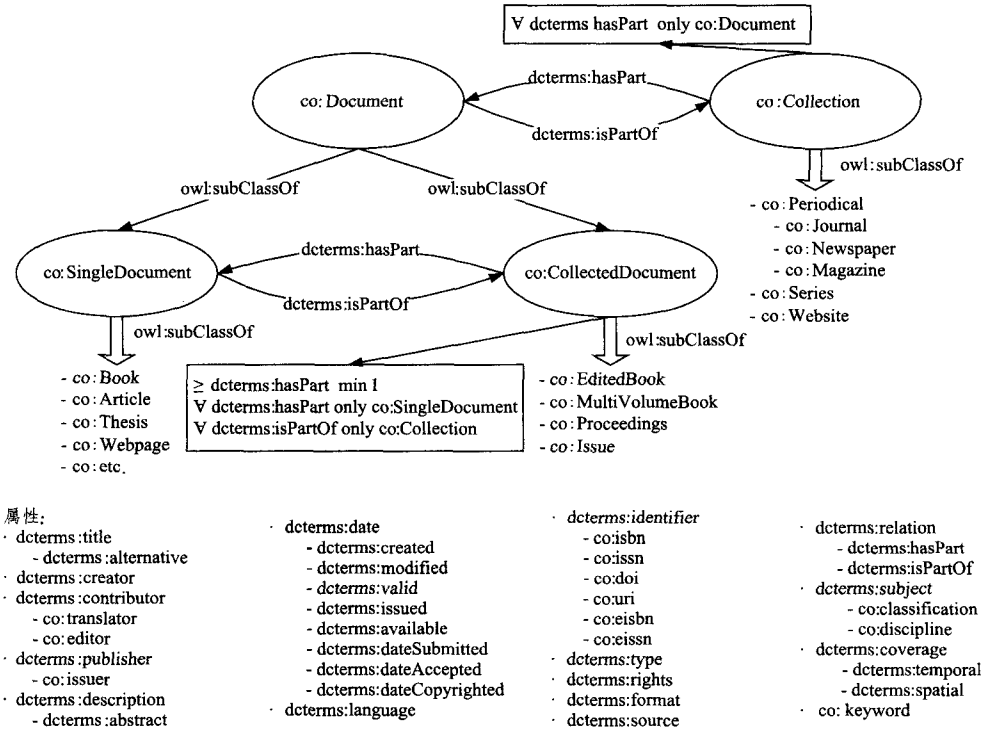
4 本体的设计

图书馆面对的是多学科、多领域的文献资源,需要建立一种通用的、与领域无关的文献资源描述和组织方法。元数据,作为对文献资源的一种通用描述方式,已经在图书馆领域广泛应用,并且逐渐向着规范化的方向发展,即采用规范化的本体描述语言(如 RDFS、OWL)对元数据规范进行形式化描述,譬如著名的 DC 和 DCTERMS 元数据规范均采用 RDFS 语言进行描述。近年来国外逐渐出现了采用 OWL 语言描述的书目本体、数字资源结构本体等,如 MarcOnt[29] 和 BIBO (Bibliographic Ontology)[31] 本体,对文献资源的书目元数据或者结构特征进行规范化描述。在本研究中,作者提出了基于元数据规范构建元数据本体的思想,其目的是对文献资源的属性以及文献资源之间、文献资源与其他资源之间的相互关系进行精确的语义化描述,从而将不同格式、不同类型的元数据转换成为统一的以 RDF 格式表示的语义元数据,以实现不同类型元数据之间的语义互操作。

如上节所述,鉴于 DC 和 DCTERMS 是描述文献资源的通用元数据规范,因此作者首先基于 DC 和 DCTERMS 构建了一个通用的 OWL 核

① 在传统 Web 中,对数据资源的访问需要采用 API 等方式进行,不能像对 Web 文档那样直接访问。

心元数据本体(命名空间为 co)。在该本体中,按照粒度不同将文献资源分成两大类——文档(Document)和文档集合(Collection),其中文档又分为单一文档(SingleDocument)和合集文档(CollectedDocument)。单一文档是指相对独立的一个文档,譬如一本图书(Book)、一篇文章(Article)、一本学位论文(Thesis)、一个图片(Image)、一份报告(Report)、一张网页(Webpage)等。合集文档是指紧密合并在一起进行出版和发行的一组文档,譬如由不同作者的论文汇编而成的编辑图书(EditedBook)、多卷书(MultiVolumeBook)、期刊中的一期(Issue)、会议论文集(Proceedings)等。文档集合是指一组相关文档的总称,但是并不绑定在一起出版和发行,譬如期刊(Periodical)、丛辑(Series)、网站(Website)等。文献资源的属性主要复用自 DCTERMS 中的 15 个核心元素,此外还根据需要自行扩展了一些新的属性或子属性,如 co:keyword(非规范化的关键词)。在核心元数据本体中,主要采用 dcterms:relation 属性的一对互逆子属性 dcterms:isPartOf 和 dcterms:hasPart 来描述文献资源之间整体与部分的关系,譬如论文(Article)是某期期刊(Issue)的一部分,而期刊中的一期(Issue)又是整个期刊(Journal)的一部分。核心元数据本体中主要的类和属性如图 2 所示。



注:@prefix co: <http://purl.org/ontology/core#>. @prefix dcterms:<http://purl.org/dc/terms/>.

图 2 核心元数据本体中主要的类和属性

核心元数据本体是各种类型文献资源共享的一个通用本体。特定类型的文献资源(如期刊论文、会议论文、学位论文等)往往还具有各自特殊的属性,是核心元数据本体中所没有容纳的,此时可通过定义新属性或者为现有属性添加子属性来扩展核心元数据本体,生成针对某种特定文献资源的专门元数据本体。譬如对于学位论文,扩展了三个新属性:cox:advisor(指导

教师），cox:degree（被授予的学位），cox:degreeConferredOrganization（授予学位的机构），这里命名空间 cox 指对核心元数据本体进行扩展的部分。

对于数字图书馆中的其他资源（如人名、地名、知识组织资源等），主要利用现有本体或者现有本体的扩展进行描述。对于个人/组织机构/团体，采用 FOAF 本体（命名空间为 foaf）中的 foaf:Agent 类及其相关属性进行描述。FOAF（Friend of A Friend）本体是在 FOAF 项目中创建的一个描述个人、团体和组织机构的本体，是目前应用最广泛的描述人及其行为的本体[32]。foaf:Agent 类包含了 3 个子类：foaf:Person（个人）、foaf:Organization（组织机构）和 foaf:Group（团体），描述 foaf:Agent 类的属性有 foaf:name（姓名），foaf:title（头衔），foaf:homepage（主页），foaf:phone（电话），foaf:mbox（邮箱）等。对于会议等事件，采用 EVENT 本体（命名空间为 event）中的 event:Event 类及其相关属性进行描述。EVENT 本体是由伦敦大学玛丽皇后学院数字音乐中心于 2004 年开发的一个 OWL 本体，主要用于描述事件（如会议、演出、音乐会、节日等）以及与之相关的时间、地点、主体、因素、产品等信息[33]。在该本体中定义了一系列描述事件的属性，如 event:place（事件地点），event:time（事件时间），event:Agent（事件主体），event:product（事件产品）等。对于时间的描述目前有两个比较著名的本体：一个是伦敦大学玛丽皇后学院数字音乐中心于 2004 年构建的 TimeLine 本体[34]，另一个是南加州大学信息科学研究所于 2006 年开发的 Time 本体[35]。这两个本体有一些重合之处，都包含了描述不同时间单元的两个类 Instant（时刻）和 Interval（时间段）类，但是 TimeLine 本体对这两个类的描述更加简单易读，因此在本研究中选择 TimeLine 本体描述时间概念。对于地名，作者直接从 GeoNames 地理数据库中获取其描述。GeoNames 地理数据库包含了约 620 万个地名，每个地理名称都有一个唯一的 URI 标识符，基于 GeoNames 本体进行了语义化描述，并且已经发布为关联数据[36]。对于规范术语，采用 SKOS 语言进行描述。但是对于中文知识组织系统，如《中国图书馆分类法》和《汉语主题词表》，标准 SKOS 语言并不能够完全胜任，因此作者对 SKOS 核心模型进行了定制化扩展使其能够无损地应用于中文知识组织资源。

5 关联数据的构建

5.1 关联数据中资源的命名及访问机理

在关联数据中，所有实体对象或抽象概念（如文献资源、个人、组织机构、地点、事件、术语等）都必须采用唯一的 HTTP URI 标识符进行命名，但是它们的 URI 地址不能够被 HTTP 协议直接解引用。这些实体对象或抽象概念在 Web 架构中被称为非信息资源，以区别于传统 Web 中 URI 地址能够被 HTTP 协议直接解引用的信息资源（如网页、图片或其他数字媒体格式等）[37]。对于非信息资源，Web 架构提供了两种方式来解决其在 Web 上的访问问题：一种是 Hash URIs，另一种是 303 URIs[38]。

Hash URIs 方式是采用带有“#”分隔符的 URI 标识符命名非信息资源，如将元数据本体中定义的抽象概念 Book 命名为 < http://hostname/ontology/core#Book >。当使用浏览器访问一个非信息资源的 Hash URI 地址时，HTTP 协议会自动将 URI 地址中“#”符号之后的部分剥离掉，服务器返回的是剥离后的 URL 地址指向的信息资源的一个表示（如 OWL 文档），该表示包含了对被请求的非信息资源（如 Book 概念）的描述[38]。Hash URIs 访问方式适用于小型的 RDF 词表，浏览器可以很快显示整个词表文档，而且因为文档长度较小易于浏览，但是对于含有大量三元组的 RDF 文档则不适用。

303 URIs 方式是采用带有“/”分隔符的 Slash URI 标识符命名非信息资源，如将一本图书命名为 < http://hostname/document/book/isbn9787301149034 >。当使用浏览器访问一个非信息资源的 303 URI 地址时，服务器根据客户端浏览器的类型将其重定向到描述它的一个信息

资源的URI地址,然后浏览器再向服务器请求这个新的URI地址,服务器返回HTML或RDF/XML文档,它提供了对被请求的非信息资源的描述[38]。因此,对于一个非信息资源需要命名三个相关的URI地址:①资源本身的URI地址;②资源元数据的RDF/XML表示;③资源元数据的HTML表示。采用303 URIs方式的一个主要缺点是需要两次HTTP请求才能获取一个非信息资源的描述,因此会造成访问延迟。

在本研究中采用Hash URIs方式命名本体中的类和属性,采用303 URIs方式命名所有的实例(即个体)。

5.2 语义数据的构建与关联

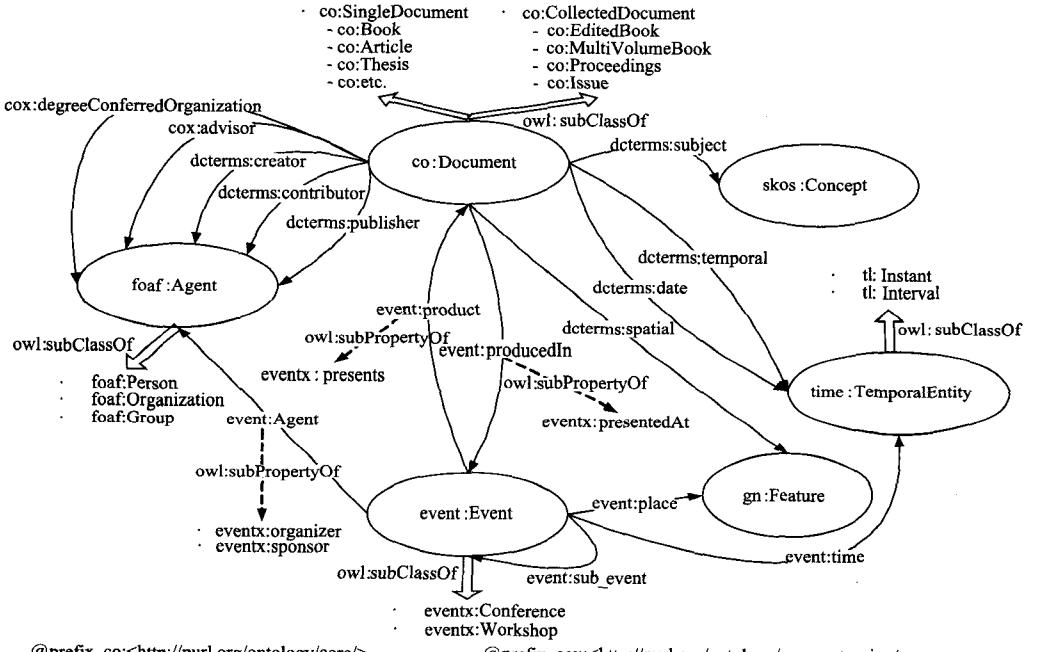
本研究选择国家图书馆书目数据库和万方数据库作为数据源构建关联数据。为了使来源数据之间具有较强的关联性,下载的数据主要集中在"图书、情报与档案学"领域。作者从国家图书馆下载了以文本格式表示的30本中英文图书的MARC记录;从万方数据库中下载了以XML格式表示的100篇中英文期刊论文、50篇国内外学术会议论文和70篇国内学位论文的书目数据以及相关的学术机构和科技专家信息。关联数据的构建分为两个阶段:第一阶段是基于本体构建RDF格式的语义元数据,第二阶段是将不同数据集中的RDF数据进行关联。

在第一阶段,基于前文中构建的本体(见第4节),作者采用JAVA语言实现了从MARC文本格式记录和XML格式记录到RDF/XML格式的转换,分别生成了文献、个人/组织机构和学术会议三种资源的语义元数据,其中文献资源又包含了图书、论文、学位论文、期刊、会议录、丛书等子集,个人/组织机构数据集又包含了个人、组织机构和团体三个子集。在这一阶段,对资源的描述基本上全部采用数据类型属性,即属性值为文本字符串。

在第二阶段,作者在不同数据集中的数据间建立RDF链接。链接类型有两种:数据层面的链接和语义层面的链接。数据层面的链接是指相同资源间的链接。在开放的Web空间里,经常会出现不同的信息提供者提供同一个资源的情况,他们通常采用各自的命名规则对资源进行命名,如 <http://hostname1/book/isbn9781608454303> 和 <http://hostname2/resource/doi10.2200_S00334ED1V01Y201102WBE001>,这两个不同的URI地址指向的其实是同一个资源。对于指向同一个非信息资源的不同URI地址,被称为URI别名(URI Aliases),通常采用RDF链接"owl:sameAs"连接两个URI别名,从而识别不同数据源中的相同资源。在本研究中只从两个数据源中获取文献资源,重复资源的情况很少出现,因此主要面对的是语义层面的链接而非数据层面的链接。语义层面的链接是指在同一本体的类之间、不同本体的类之间以及本体与概念体系之间的链接。通过在前文所述的元数据本体、EVENT本体、FOAF本体、GeoNames本体和SKOS概念体系间建立RDF语义链接(见图3),能够实现数字图书馆中文献资源、个人/组织机构/团体、地点和知识组织资源的相互关联,构成图书馆的关联数据。

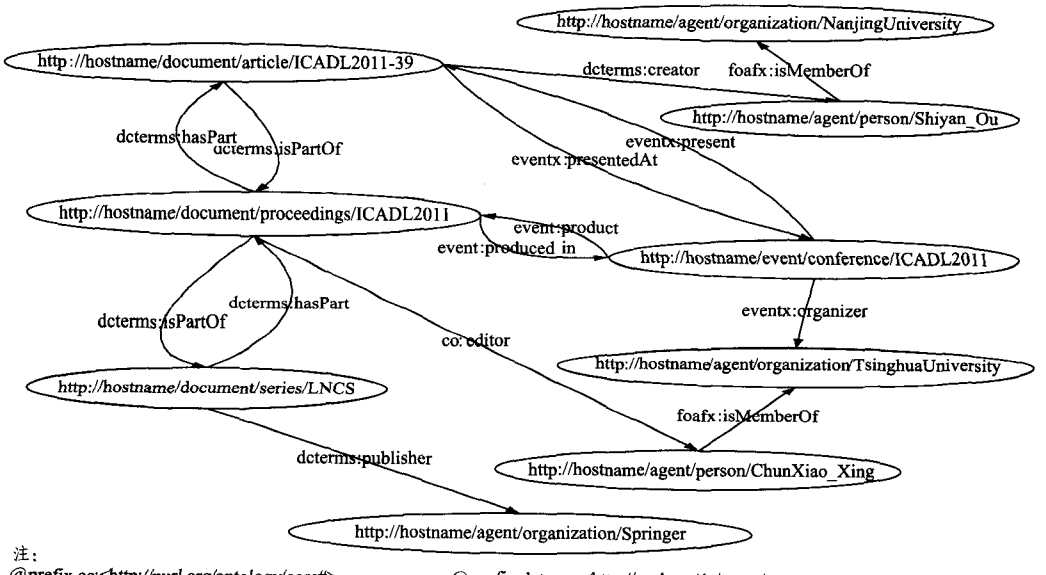
在第一阶段生成的RDF数据中,数据之间的关联关系是隐含地存在于数据类型属性中的。在第二阶段需要将这种隐性的语义关系转换为显性的RDF语义链接,即采用URI标识符替换原有的文本字符串属性值,将数据类型属性转换为对象属性。通过图3中设定的RDF链接的值域,定位相应的数据集,然后采用字符串模糊匹配的方法自动从该数据集中查找与原有属性值相匹配的实体,用其URI标识符替换原有的文本字符串值。图4显示了基于万方数据库中的一篇会议论文记录生成的论文、会议录、丛书、会议、个人、团体、组织机构、地点相互关联的关联数据。

此外,图书馆内部的数据还可进一步与外界的关联开放数据相关联,如DBPedia、DBLP Bibliography等,使图书馆中的各种资源成为整个Web空间数据之网的一部分,更易于被用户发现和浏览。



@prefix co:<http://purl.org/ontology/core/>.
@prefix cox:<http://purl.org/ontology/core_extension/>.
@prefix foaf:<http://xmlns.com/foaf/0.1/>.
@prefix skos:<http://www.w3.org/2004/02/skos/core#>.
@prefix event:<http://purl.org/NET/c4dm/event.owl#>.
@prefix eventx:<http://prul.org/NET/c4dm/event_extension.owl#>.
@prefix time:<http://www.w3.org/2006/time#>.
@prefix tl:<http://purl.org/NET/c4dm/timeline.owl#>.
@preifx gn:<http://www.geonames.org/ontology#">.
@prefix dcterms:<http://purl.org/dc/terms/>.

图3 本体间的语义关联示意图



注：
@prefix co:<http://purl.org/ontology/core#>.
@prefix dcterms:<http://purl.org/dc/terms/>.
@prefix event<http://purl.org/NET/c4dm/event.owl#>.
@prefix eventx:<http://purl.org/NET/c4dm/event_extension.owl#>.
@prefix foaf:<http://xmlns.com/foaf/1.0/>.
@prefix foafx:<http://xmlns.com/foaf/1.0/extension/>.

图4 关联数据示意图

6 关联数据的发布

目前关联数据的发布主要有以下六种方式：

(1)以静态 RDF/XML 文件发布关联数据：这种方式通常用于发布小型的 RDF 词表，对于大数据量并不适用，因为需要预先生成大量的 HTML 或 RDF/XML 文档。

(2)通过服务器端脚本发布关联数据：譬如 Semantic Web Dog Food 就是采用“Jena RDF 存储器 + Jena Joseki SPARQL 终端 + PHP 脚本”的组合发布 RDF/XML 格式的会议信息[21]。

(3)以 RDFa 格式发布关联数据：譬如德国经济学图书馆采用该种方式发布了经济学词表 STW[15]，其缺点是构建内嵌所需 RDF 三元组的网页比较复杂。

(4)从 RDF 存储器发布关联数据：该种方式是在 RDF 存储器的 SPARQL 终端的前端放置一个关联数据界面(如 Pubby[39]、Elda[40])，将不可解引用的 URI 地址转换为能够被 HTTP 协议解引用的，实现关联数据显示，譬如 AGROVOC 词表就是采用 Pubby 发布的[12]。

(5)从关系型数据库发布关联数据：该种方式是利用现成的工具(如 D2R、Triplify[41]、OpenLink Virtuoso[42])将存储在关系型数据库中的关系型数据直接发布为关联数据，譬如 DBLP 就是采用这种方式发布的[18-19]。

(6)通过包装已有的应用或 Web APIs 发布关联数据：譬如 RDF Book Mashup[17]。

在本研究中，选择以方式(1)和方式(2)发布关联数据。方式(1)用于将语义关联的各种本体(包括元数据本体、FOAF 本体、EVENT 本体、TIMELINE 本体等)发布为关联数据，因为本体的数据量较少，适于发布为静态的 RDF/XML 文件。方式(2)则用于将真正的图书馆数据(即相互关联的文献资源和其他相关资源的语义元数据)发布为关联数据。因为在本研究中已经预先生成了文献资源及相关资源(包括个人/组织机构/团体、地点、事件、知识组织资源等)的 RDF/XML 语义元数据，适于直接将 RDF 数据存储在 RDF 存储器中进行发布。

在方式(1)中，选用目前广泛使用的 Apache HTTP Server(安装版本 2.0.64)作为 Web 服务器，采用该方式的关键是在 Web 服务器的主配置文件 httpd.conf 和(或)分布式配置文件.htaccess 中添加对 RDF/XML 内容类型的支持，开启重写功能并建立好 URL 重写规则。

在方式(2)中，采用“Jena TDB + Jena Joseki + PHP 脚本”的组合来发布 RDF 关联数据，其架构如图 5 所示。Jena 是由 HP 实验室开发的一个开源语义网框架，提供了一整套用于开发语义网和关联数据应用的工具和 Java 类库，目前由 Apache 基金会接管。TDB(安装版本 0.8.10)是 Jena 提供的一种持久化的 RDF 数据存储模式，具有速度快，效率高的优点，适用于大数据量的存储。Joseki(安装版本 3.4.4)是 Jena 提供的一个支持 SPARQL 协议和 SPARQL RDF 查询语言的 HTTP 引擎，能够为 RDF 数据提供一个独立的 SPARQL 查询终端，但因为该终端 SPARQL 查询结果中的 URL 地址是不能被 HTTP 协议解引用的，无法进一步在 Web 上进行浏览，因此这里采用服务器端脚本基于存储的 RDF 数据自动生成描述非信息资源的 HTML 或 RDF/XML 文档，并将其不可解引用的 URI 地址重定向到描述它的文档。Joseki 需要一个 Servlet 容器来运行，这里选择 Tomcat 7.0.25 作为 Servlet 容器，服务器端脚本则采用 PHP 语言来实现。

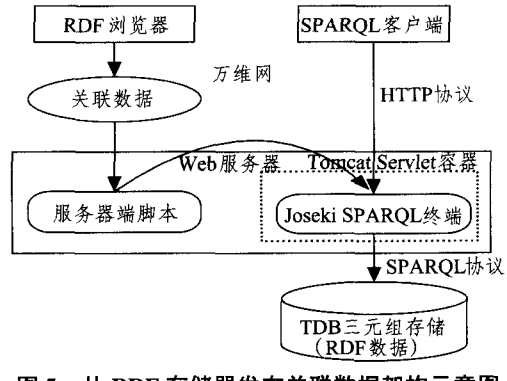


图5 从 RDF 存储器发布关联数据架构示意图

对于关联数据的浏览，目前已经有了众多

的关联数据浏览器(也即 RDF 浏览器),如 Disco[43]、Tabulator[44]、Marbles[45]等。据作者调查,绝大多数浏览器都只有在线版本,只有 Tabulator 具有本地版本,而且是以 Firefox 插件的形式存在的①。因此作者在 Firefox 3.6.26 中安装了 Tabulator 插件,使之成为一个 RDF 浏览器,用于关联数据的浏览。

与传统的基于应用程序接口(即 API)的数据访问方式相比,关联数据提供了一种统一的、标准的数据访问机制,避免了访问界面和结果格式的纷繁复杂。通过采用关联数据,能够在来自不同数据源的数据间建立链接,使数据源更容易被搜索引擎抓取,而且能够采用通用的数据浏览器(即 RDF 浏览器)访问不同的数据源[46]。

7 结论与展望

本文提出了一个面向关联数据的语义数字图书馆资源描述与组织框架,该框架具有四个层次:元数据层、本体层、关联数据层和应用层,其核心是 RDF 语义元数据的构建与关联。本研究以"图书、情报与档案学"领域的数据为例对语义数字图书馆资源描述与组织框架进行了实现。实现的重点是在本体层和关联数据层,包括本体的设计、普通元数据到语义元数据的自动转换、不同数据集间 RDF 语义链接的建立、关联数据的发布等。在后续研究中,还将对应用层进行实现,其关键是如何对涉及多个本体的关联数据实现基于自然语言的问答式检索。这也是当前语义网和自然语言处理领域的最新关注热点。基于语义数字图书馆资源描述与组织框架,能够构建一个数据关联的语义数字图书馆原型,实现对数字图书馆各种资源的语义化描述和语义检索以及全方位的数据关联。

参考文献:

[1] JeromeDL[OL]. [2012-05-10]. http://www.jeromedl.org.
[2] SIMILE[OL]. [2012-05-10]. http://simile.mit.edu.
[3] Bricks[OL]. [2012-05-10]. http://www.brickscommunity.org.
[4] Kruk S R, McDaniel B. Goals of semantic digital libraries[G]//Kruk S R, McDaniel B. Semantic Digital Libraries. Heidelberg: Springer, 2009: 71-76.
[5] Berners-Lee T. Linked data. Personal notes on design issues for the World Wide Web[OL]. [2012-05-10]. http://www.w3.org/DesignIssues/LinkedData.html.
[6] Malmsten M. Making a library catalogue part of semantic web[C]//Proceedings of the 2008 International Conference on Dublin Core and Metadata Applications. Singapore: Dublin Core Metadata Initiative, 2008: 146-152.
[7] Summers E, Isaac A, Redding C, et al. LCSH, SKOS and linked data[C]//Proceedings of the 2008 International Conference on Dublin Core and Metadata Applications. Singapore: Dublin Core Metadata Initiative, 2008: 25-33.
[8] Kruk S R, Cygan M, Czella A, et al. JeromeDL—The social semantic digital library[G]//Kruk S R, McDaniel B. Semantic Digital Libraries. Heidelberg: Springer, 2009: 139-150.
[9] Butler M H, Gilbert J, Seaborne A, et al. Data conversion, extraction and record linkage using XML and RDF tools in project SIMILE[R]. HP Labs Technical Report HPL-2004-147. Bristol: HP Laboratories, 2004: 2-15.
[10] Haslhofer B, Hecht R. Metadata management in a heterogeneous digital library[C]//Proceedings of the eChallenges 2005. Amsterdam: IOS Press, 2005: 1251-1258.
[11] Bizer C, Jentzsch A, Cyganiak R. State of the LOD cloud(version 0.3)[OL]. [2012-05-10]. http://www4.wiwiss.fu-berlin.de/lodcloud/state.
[12] Caracciolo C. et al. Thesaurus maintenance, alignment and publication as linked data: The AGROVOC use case

① 据测试,Tabulator 插件只能安装在 Firefox 3.X 版本。

[C]//Proceedings of the 5th International Conference on Metadata and Semantics Research. Heidelberg:Springer, 2011:489 - 499.

[13] OCLC. Dewey summaries as linked data[OL]. [2012 - 07 - 30]. http://www.oclc.org/dewey/webservices/default.htm.

[14] Meij L, Isaac A, Zinn C. A web-based repository service for vocabularies and alignments in the cultural heritage domain[C]//Proceedings of the 7th European Conference on the Semantic Web: Research and Applications-Volume Part 1. Heidelberg:Springer, 2010:394 - 409.

[15] Neubert J. Bringing the "thesaurus for economics" on to the web of linked data[C]//Proceedings of the WWW 2009 Workshop on Linked Data on the Web. CEUR - WS. org, 2009. [2012 - 09 - 20]. http://ceur - ws. org/Vol - 538/ldow2009_paper7. pdf.

[16] WorldCat linked data[OL]. [2012 - 07 - 30]. http://www.oclc.org/data.html.

[17] Bizer C, Cyganiak R, Gauss T. The RDF book mashup: From web APIs to a web of data[C]//Proceedings of the 3rd Workshop on Scripting for the Semantic Web. CEUR - WS. org, 2007. [2012 - 09 - 20]. http://ceur-ws.org/Vol - 538/ldow2009_paper7. pdf.

[18] D2R server publishing the DBLP bibliography database[OL]. [2012 - 05 - 10]. http://www4.wiwiss.fu-berlin.de/dblp.

[19] D2R server publishing the DBLP bibliography database, hosted at L3S research center[OL]. [2012 - 05 - 10]. http://dblp.l3s.de/d2r.

[20] Glaser H, Millard I, Jaffri A. RKB Explorer.com: A knowledge driven infrastructure for linked data providers [C]//Proceedings of the 5th European Conference on the Semantic Web: Research and Applications. Heidelberg: Springer, 2008:797 - 801.

[21] Moller K, Heath T, Handschuh S, et al. Recipes for semantic web dog food—the ESWC and ISWC metadata projects[C]//Proceedings of the 6th International Semantic Web Conference and 2nd Asian Semantic Web Conference. Heidelberg:Springer, 2007:802 - 815.

[22] 刘炜. Semantic interoperability of linked data[OL]. [2012 - 05 - 10]. http://www.kevenlw.name/downloads/uploads/presentations/liuwei-jdasdl.ppt. (Liu Wei. Semantic interoperability of linked data[OL]. [2012 - 05 - 10]. http://www.kevenlw.name/downloads/uploads/presentations/liuwei - jdasdl.ppt.)

[23] 曾蕾. 关联的图书馆数据[OL]. [2012 - 05 - 10]. http://202.114.9.60/dl6/pdf/26.pdf. (Zeng Lei. Linked library data[OL]. [2012 - 05 - 10]. http://202.114.9.60/dl6/pdf/26.pdf.)

[24] 沈志宏,张晓林. 关联数据及其应用现状综述[J]. 现代图书情报技术,2010,26(11). (Shen Zhihong, Zhang Xiaolin. Linked data and its applications: An overview[J]. Modern Technology of Library and Information, 2010,26(11).)

[25] 黄永文. 关联数据在图书馆中的应用研究综述[J]. 现代图书情报技术,2010,26(05). (Huang Yongwen. Research on linked data-driven library applications[J]. Modern Technology of Library and Information,2010,26(05).)

[26] 刘炜. 关联数据:概念、技术及应用展望[J]. 大学图书馆学报,2011,29(2). (Liu Wei. Linked data: Concepts, technologies and application perspective[J]. Journal of Academic Libraries,2011,29(2).)

[27] 李琳. 关联数据在图书馆界的应用与挑战[J]. 图书与情报,2011(4). (Li Lin. Application and challenge of linked data in libraries[J]. Library and Information,2011(4).)

[28] 刘炜,李大铃,夏翠娟. 元数据与知识本体[J]. 图书馆杂志,2004,23(6). (Liu Wei, Li Daling, Xia Cuijuan. Ontology-based metadata application for digital libraries[J]. Libary Journal,2004,23(6).)

[29] Kruk S R, Synak M, Zimmerman K. MarcOnt: Integration ontology for bibliographic description formats[C]//Proceedings of the 2005 International Conference on Dublin Core and Metadata Applications. Singapore: Dublin Core Metadata Initiative,2005:231 - 234.

[30] 欧石燕. 基于SOA架构的术语注册和服务系统设计与应用[J]. 中国图书馆学报,2011(5). (Ou Shiyan. Design and application of SOA based terminology registry and terminology services[J]. Journal of Library Science in China,2011(5).)

[31] Giasson F, D'Arcus B. The bibliography ontology[OL]. [2012 - 05 - 10]. http://bibliontology.com.

[32] Brickley D, Miller L. FOAF vocabulary specification 0.98[OL]. [2012 - 05 - 10]. http://xmlns.com/foaf/

spec/.

[33] Raimond Y, Abdallah S. The event ontology[OL]. [2012-05-10]. http://motools.sourceforge.net/event/event.html, 2007.

[34] Raimond Y, Abdallah S. The timeline ontology[OL]. [2012-05-10]. http://motools.sourceforge.net/timeline/timeline.html, 2007.

[35] Hobbs J R, Pan F. Time ontology in OWL[OL]. [2012-05-10]. http://www.w3.org/TR/owl-time/.

[36] GeoNames ontology[OL]. [2012-05-10]. http://www.geonames.org/ontology/documentation.html.

[37] W3C Technical Archiecture Group. Architecture of the World Wide Web, volume one—W3C recommendation 15 December 2004[OL]. [2012-09-20]. http://www.w3.org/TR/2004/REC-webarch-20041215.

[38] Sauermann L, Cyganiak R. Cool URIs for the semantic web—W3C interest group note 03 December 2008[OL]. [2012-09-20]. http://www.w3.org/TR/cooluris.

[39] Cyganiak R, Bizer C. Pubby: A linked data frontend for SPARQL endpoints[OL]. [2012-05-10]. http://www4.wiwiss.fu-berlin.de/pubby.

[40] Elda: Epimorphics linked data API implementation[OL]. [2012-05-10]. http://code.google.com/p/elda.

[41] Auer S, Dietzold S, Lehmann J. Triplify-light-weight linked data publication from relational databases[C]//Proceedings of the 18th International Conference on World Wide Web. New York: ACM, 2009: 621-630.

[42] Virtuosouniversal server[OL]. [2012-09-20]. http://virtuoso.openlinksw.com.

[43] Chris Bizer C, Gauß T. Disco-hyperdata browser: A simple browser for navigating the semantic web[OL]. [2012-05-10]. http://www.wiwiss.fu-berlin.de/bizer/ng4j/disco.

[44] Berners-Lee T, et al. Tabulator redux: Writing into the semantic web[OL]. [2012-09-20]. http://eprints.soton.ac.uk/264773/1/tabulatorWritingTechRep.pdf.

[45] Becker C, Bizer C. Marbles[OL]. [2012-05-10]. http://marbles.sourceforge.net.

[46] Heath T, Bizer C. Linked data: Evolving the web into a global data space(1st edition)[M]//Hendler J. Synthesis Lectures on the Semantic Web: Theory and Technology. Morgan & Claypool, 2011: 1-136.

欧石燕 南京大学信息管理学院教授。
通讯地址：江苏省南京市鼓楼区汉口路22号。邮编：210093。

（收稿日期：2012-06-15）