

NSTL的关联数据构建与应用场景设想*

□ 乔晓东 白海燕 梁冰 / 中国科学技术信息研究所 北京 100038

摘要: NSTL关联数据构建的内容主要包括NSTL结构化数据的RDF关联表达与发布、NSTL的关联数据消费及与外部数据集的关联构建,以及基于科技知识组织系统构建科技知识数据网络。关联数据在资源组织、检索及服务方面的应用场景主要包括检索结果扩展、异类资源整合检索、多维分面组织与检索、基于关系的复杂和智能检索、基于关联数据的融合与混搭服务、基于数据节点和关系的嵌入式服务等。

关键词: 关联数据, 信息组织, 信息检索

DOI: 10.3772/j.issn.1673—2286.2011.12.008

1 引言

NSTL的文献服务平台,经过多次技术改造升级,已经发展成为集海量科技文献检索与服务的大型公益性国家科技数字图书馆,并采用嵌入第三方系统、接口开放、知识库以及集成揭示等多种方式,实现资源的开放与整合^[1]。NSTL服务系统未来的发展方向包括实现资源组织的深度序化、语义化和知识化,从传统检索查询模式向具有知识导航、自动聚类、语义检索和双语查询等功能的智能检索模式转变,并基于知识节点和知识关系嵌入用户的信息环境和构建科研信息情境;从文献检索、文献传递服务向科技对象实体的关联发现、科技热点监测、科技趋势分析等知识服务方向

转变。

关联数据是一种轻量级的语义网技术。近年来发展迅速,许多海量信息机构,如BBC、路透社、维基百科、美英政府信息部门、美国国会图书馆等,纷纷将其资源“文档”标注、解析为以URI标识的地点、人物、事件、主题等数据节点,通过RDF连接的三元组形式在Web上发布和提供查询,并与其他数据集相互关联;通过统一、标准、自助、去中心化的关联数据整合机制与混搭平台,催生和驱动了许多功能丰富和具有创新性的应用,且易于将原生资源和增值资源嵌入用户的信息环境^[2]。

关联数据是一种推荐的最佳实践,用来在语义网中使用URI和RDF发布、分享、连接各类数据、信

息和知识,发布和部署实例数据和类数据,从而可以通过HTTP协议揭示并获取这些数据,同时强调数据的相互关联、相互联系以及有益于人机理解的语境信息。关联数据提出的目的是构建具有结构化和富含语义的数据网络,以便于在此之上构建更智能的应用。关联数据也被人们视为轻量级的语义网,使用了部分语义网技术,遵循早期语义网的发展路线,建立已有信息的语义标注和实现数据之间的关联,特别是在语义网自上而下的推广面临困难和受挫的情况下,关联数据技术所具有的框架简洁、标准化、自助化、去中心化、成本低的特点,为构建人机理解的数据网络,提供了根本性的保障,为实现语义网远景奠定了坚实的基础^[3]。

* 本文系“十二五”国家科技支撑计划“信息资源自动处理、智能检索与STKOS应用服务集成”(2011BAH10B05)研究成果之一。

NSTL关联数据构建及应用的目的在于通过大规模科技文献的精细化揭示、深度组织、深度序化和知识化整序,充分发挥国家大量财力购买、大量人力物力加工建设的科技资源的最大作用,为提供更加高效、智能的信息服务奠定基础;通过文献资源与科技知识组织系统的关联构建,为知识服务提供坚实的基础,更有效地提高科技信息资源的发现和重用,更有助于构建用户科研情境;通过与外部科技信息资源的关联构建,可以扩展资源范围,从而扩大我国科技文献资源体系的服务范围和资源范围,并促进科学数据、地理信息、科研管理信息等与科技文献的融合,实现基于数据融合、服务融合的跨平台的开放与复用。

2 NSTL关联数据的主要构建内容

2.1 NSTL结构化数据的RDF关联表达与发布

2.1.1 NSTL元数据的RDF表达转换

NSTL的结构化数据资源是关联数据构建的重要基础。所谓结构化是指基于信息对象的类型抽象与划分,确定每一信息对象的属性和属性值,并依据一定的数据模型进行编码。如CSV序列化文档、关系型数据库、RDF三元组等都是具有不同数据模型的结构化数据格式。NSTL的结构化数据主要是指各类科技文献的书目和篇名元数据,目前NSTL外文科技期刊文献约2.6万

余种,拥有40余个数据库,集科技期刊、图书、会议文献、学位论文、科技报告、专利、标准和计量规程等文献信息于一体,文摘、引文、题录数据总量达1.1亿条^{①[1,4]}。

NSTL现有的文献资源组织模式主要是基于关系型数据库的二维表结构,即通过表和行数据来表达文献信息对象及其属性,实现资源的描述、组织和检索。例如,期刊篇名、参考文献、期刊母体等是三类不同的信息对象,在关系型数据库中通过不同的二维表来表达和存储;每个信息对象具有各自的属性特征,通过每个表的字段来表达,其数据类型和基本约束是在创建表结构时确定的;信息对象的属性集可通过数据字典获得。因此,NSTL文献组织的颗粒度是由信息对象和属性划分的粒度决定的,文献组织、序化的手段主要依赖于索引(基于属性特征)的构建和对对象间关系的构建,对象之间的关系是通过表的主键和外键引用来实现的。

根据Berners-Lee的关联数据四项原则^[2],关联数据使用URI和RDF发布、分享、连接各类数据、信息和知识,发布和部署实例数据和类数据,从而可以通过HTTP协

议揭示并获取这些数据。关联数据依据RDF模型的“资源-属性-属性值”的形式进行表达,使用URI来标识不同的对象(包括资源节点、属性类或属性值),并将不同的URI连接起来,清楚地表达对象间的关系。为揭示对象间关系而由URI连接而成的RDF有向图摆脱了XML文档所隐含的树形资源结构的限制,可以更加灵活地表达网络上的知识或资源,提示它们之间的相互关系。同时,文档标注中为程序或人理解所使用的标签已转换成了定义清晰的词汇,并可显式地表达机器可理解的形式化的语义^[5]。

因此,NSTL关联数据的构建,主要包括以下内容:

(1) 构建NSTL的资源组织本体。将不同的信息对象类型,如期刊、会议、专利、标准、报告、工具书、引文等的母体和篇名数据,抽象为不同的类,依据现有的数据内容、数据结构和约束,确定类、属性的表达规范和约束。

(2) 基于上述NSTL资源组织本体,将关系型数据库中数据表的各类篇名元数据转换为基于RDF的表达,其主要转换内容见表1^[6,7]。

(3) 基于关键词,构建不同来

表1 关系型数据库和RDF表达的对应关系

关系型数据库	RDF表达
表结构	RDF Scheme
表	RDF类
字段	RDF属性 (data property)
字段值	RDF属性值
表间关系	RDF链接 (object property)

^① www.nstl.gov.cn.

源、不同类型的科技文献,如期刊、会议、标准、专利、科技报告等的关联关系。关系型数据库的二维表只提供了基于外键引用的有限的对象间关系,如引用、出处等。基于关键词的关联关系能够表达不同类型数据之间的内容联系,因此,还需要对分离的dc:subject属性关系进行集成,如通过重组关键词索引表、增加相应的数据库与RDF的映射规则等实现不同资源基于相同关键词的关联关系。

2.1.2 NSTL关联数据的发布

科技文献资源服务的发展方向,是面向最终用户和代理,提供嵌入科研过程和构建科研信息“情境”的服务。这不仅仅意味着资源的开放、数据接口的开放和各类平台系统的集成和混搭,而更重要的是通过数据关联网络提供基于数据节点和数据关系的开放和关联机制,即通过HTTP URI、RDF、SPARQL的标准接口提供对外的开放和互联,其优势在于:(1)简化了现有对NSTL资源集整合、调用的技术难度,提高了通用性;(2)将经过有效组织、深度序化和知识化整序后的NSTL关联资源,即NSTL数据集(dataset),以更细的颗粒程度,更丰富的数据关系,动态实时、简单透明地提供给用户,来达到嵌入用户环境与过程、融入用户科研情境的目的。因此,NSTL作为全国最大的公益性、普惠型科技文献信息资源保障与服务系统,通过关联数据的方式发布科技文献元数据,为外部环境提供了一种通用、简便、低成本的使用和消费NSTL关联数据的机制。

构建NSTL的关联数据发布机

制,主要采用以下四种基本方式:

第一种是关联数据直接参引方式。每个对象提供可参引(dereferencable)的HTTP URI,即通过URI在Web上可获得相应资源的信息。当URI参引时要求的媒体类型(MIME-type)为Accept: Application/RDF+XML,则数据源返回被标识资源的RDF/XML描述;当HTTP头的媒体类型请求为Accept:Text/HTML时,则需要提供基于HTML的适合阅读的格式。当参引的URI标识的是信息资源时,服务器成功响应返回给客户端的HTTP代码,即HTTP200;对于非信息资源被参引时,服务器无法返回一个表达,而是根据内容协商机制,数据源会返回给客户端HTTP303代码,即重定向到描述该非信息资源的信息资源地址,客户端再次请求,得到一个描述非信息资源的信息资源^[5,8]。

第二种是SPARQL查询服务方式。SPARQL是由RDF数据访问工作组提出的一种RDF查询语言,经过几年的发展,已成为W3C的推荐标识。SPARQL主要具有图模式匹配功能,即用变量代替RDF主语、谓词或对象的限制匹配能力,允许查询返回更多变量绑定和原始查询图的子图,每一组绑定集是满足子图查询条件的一种情况^[10]。SPARQL查询服务通过支持SPARQL协议的语言处理,接收客户端的查询请求,将RDF数据结果返回给客户端。用户或第三方系统可以通过SPARQL查询服务,像使用SQL语言查询关系型数据库一样,在WEB上精确地获得所需要的数据和关联关系。

第三种是RDF转存文件方式。在SPARQL查询服务的基础上,基

于检索结果或条件限定,提供RDF数据输出,并将数据集中的全部或部分RDF数据以某种序列化方式,如RDF/XML、N-Triples等格式输出并提供下载^[5]。

第四种方式是对NSTL已有API或者Web化的服务接口如OAI-PMH协议的封装实现的。各种不同的API提供了多样化的查询、检索接口以及多种格式的返回结果(如XML、JSON等),造成了搜索引擎抓取和一般数据浏览器访问的困难。封装器可以将HTTP URI分配给由API提供的非信息资源,当被参引的URI请求为Application/RDF+XML时,封装器能够重写客户端的请求为相应的API,并将API请求的结果转换为RDF格式并发送回客户端^[5]。

在具体实现方面,可提供关联数据界面供浏览器和各种爬虫获取,同时支持SPARQL服务和RDF文件下载,并可提供相应的用户界面和功能来促进NSTL数据的多样化使用。

2.2 NSTL的关联数据消费及与外部数据集的关联构建

NSTL是目前全国最大的科技文献资源服务平台,尽管资源数量巨大,仍然需要集成第三方组织、机构、个人的开放资源和数据,扩展和丰富其资源体系,以覆盖和满足各科技领域用户日趋复杂化和多样化的信息需求。

关联数据具有基于RDF链接的关联可扩展性。基于该特性,既可以实现本地异构数据源之间的整合,也提供了与外部数据集进行关联的可能性,从而支持资源的扩展发现、数据的融合与开放复用,以

及实现各类数据和服务的混搭。因此,NSTL的关联数据构建不仅包括其自身资源的RDF发布和服务,还应该使用、操作和消费机构外部的理、工、农、医领域的各类关联数据集,通过众多的关联数据的仓储和搜索引擎,对关联数据进行监测、采集、存储和整合,并构建NSTL关联数据与外部相关数据集的关联,以扩展资源范围,扩大我国科技文献资源体系的服务范围和资源范围,并促进科学数据、地理信息、科研管理信息等与科技文献的融合。

截至2010年9月^[11],Linking Open Data项目已收录203个关联数据集,250亿条RDF三元组,以及3.95亿条RDF链接。其中包括了大量的科技资源数据集^[12],如医学文献资源PubMed Article、PubMed Author^②,MESH主题词表,生物、基因与医疗数据集Bio2RDF^③,农业叙词表AGROVOC^④,地理数据GeoNames^⑤,计算机科学论文与作者编目数据DBLP^⑥,全学科的文献出版物如CiteSeer、EPrints等,全领域的百科全书DBpedia^⑦,通用本体和主题概念数据集Yago^⑧、UMBEL^⑨等等。

NSTL使用关联数据及与外部数据集的关联构建内容主要包括以下方面:

(1) 与科技文献数据集的关联

通过与外部科技文献数据集的关联,能够补充自身资源不足,达到扩展资源范围的目的。例如,

NSTL以科技期刊和会议录为收藏主体,相应的科学专著和图书类文献收录较少,通过主题、参考引用等信息,生成与RDF Book Mashup、DBLP等书目服务数据的链接,指引用户或机器使用这些开放的结构化数据,扩展用户资源发现。

(2) 与知识组织工具的关联

通过与各类SKOS化的规范文档、主题词表、叙词表、分类表关联链接,以及与通用本体、百科全书构建基于关联词的关联,其最直接的作用是获得某一主题、术语、关键词的知识点注释、信息扩展说明和更为丰富的知识关系链接;此外,通过各类知识组织工具的链接与嵌入,能够增加NSTL资源组织的知识化程度,实现词义消歧、范畴分类、知识分面、扩检/缩检等更丰富的功能。

(3) 与科学数据集的关联

对NSTL科技文档中的科学数据,如基因、蛋白、地理名称、分子式等进行抽取,与各类开放的科学数据集建立关联,从而获得完整、权威、准确的科学数据,并保证更新同步。例如,通过文档中的某一地理名称的扩展链接,可获得其经度、纬度、地图、水文等各类信息。其作用不仅达到了补充、扩展资源的目的,还可以通过科学数据集本身之间的关联关系,进一步实现语义挖掘、语义检索等智能应用等。

与外部数据集的关联关系,主要是owl:sameas的等同和映射关系,与外部数据集的关联方式主要

有生成RDF链接、直接使用返回数据、嵌入数据页面等。关联的构建方式主要有基于实体的文本映射(简单文本查找、扩展文本查找)、基于图相似度的映射和基于规则的互联^[13,14]。对于NSTL关联数据,通过与外部数据集的关联链接,也使自身的原生资源获得了增值,通过再次发布服务和开放接口,能够让原生科技资源及其增值资源,供人类用户、机器和代理使用。

2.3 基于科技知识组织系统构建科技知识数据网络

NSTL科技文献资源的知识数据网络由三部分组成,第一层为本体层,即知识组织系统层,由覆盖理、工、农、医四大领域的统一超级科技词表、领域本体和科研本体组成;第二层是实例层,是基于本体层,通过对科技文献的文本内容分析、知识抽取、自动标注、相关计算、数据挖掘形成的文献标引层,是知识组织系统层的实例映射;第三层是文献层,由NSTL的科技文献资源文档组成,是标引数据在文献资源中的实例映射。关联数据网络,就是将彼此分离的三个层面的相互映射关系进行显性表达,集成到同一层面上,连通共同的概念和关系,形成知识关联网络。通过知识关联网络的构建,将实现对大规模科技文献资源的结构化深度揭示、知识化整序和语义化关联,能够从整体上提升我国科技信息资源组织的语义化和知识化程度,为语

^② <http://esw.w3.org/TaskForces/CommunityProjects/LinkingOpenData/DataSets>.

^③ <http://bio2rdf.wiki.sourceforge.net>.

^④ <http://aims.fao.org/website/Web-services/sub>.

^⑤ <http://www.geonames.org/ontology/>.

^⑥ <http://www4.wiwiw.fu-berlin.de/dblp/>.

^⑦ <http://dbpedia.org>.

^⑧ <http://www.mpi-inf.mpg.de/yago-naga/yago/>.

^⑨ <http://www.umbel.org/>.

义和智能检索奠定基础。

知识数据网络同时构建关联发布和检索输出机制,以满足知识网络的独立对外发布与服务,同时也为上层知识服务和智能服务提供智能接口。主要包括:

- 对所有数字对象使用唯一、持久的数字标识,兼顾可用性和持久性,支持资源发现、获取、利用的一致性和持久性;

- 通过统一的数据模型、一致的语义描述方法,提供统一的存取,屏蔽了复杂的接口和标准规范,支持多种发现模式;

- 支持节点级应用,不需绑定特定应用,不必为特定应用定制,具有动态性、灵活性和可扩展性。

知识数据网络的构建依赖于本体层(统一超级科技词表、领域本体、科研本体)的完整构建和实例层海量数据的抽取与标注质量。作为知识数据网络本身,也存在相当的管理和实现难度。主要表现在:

(1) 层次复杂:由知识组织系统、科技文献实例和知识数据三个网络层次互相连接而成;(2) 数据规模巨大:节点数目众多,节点之间的连接数目众多,且节点之间的关系种类众多;(3) 存在数据噪音:如实体的歧义、重复,知识关系有待校验和一致性检查等;(4) 具有进化性:节点或连接具有动态产生与消失特性等。因此,由海量知识数据组成的复杂网络要支持上层的知识服务和智能检索,还需要通过有机的组织、集成、存储和控制,构建成为知识组织体系、知识对象/知识关系和科技文献实例不同层次集成和互映射的、知识节点相互联通的、可管理、可维护、具有检索输出机制的知识网络。

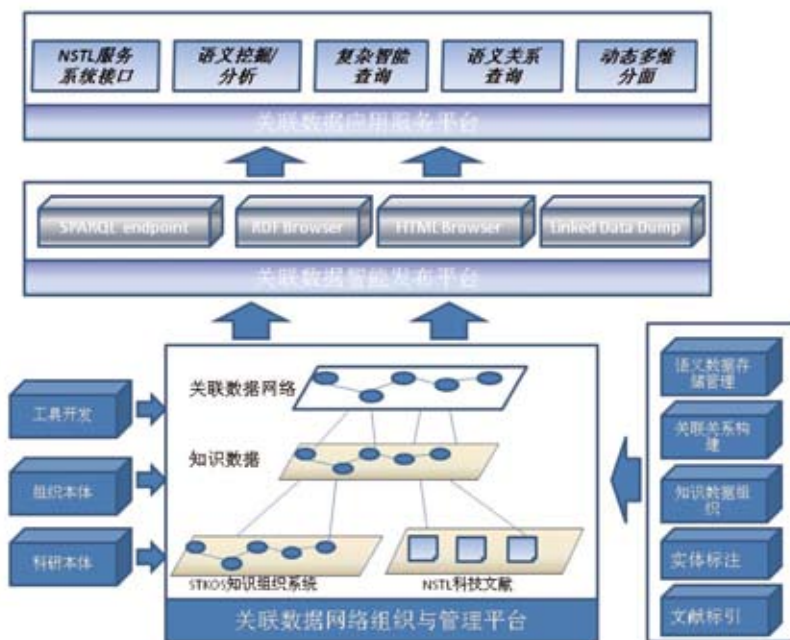


图1 关联数据各平台关系图

3 NSTL关联数据的应用场景

3.1 资源组织与检索

(1) 检索结果扩展

基于关联数据的资源扩展是一种数据融合(data mashup)应用。假设在NSTL现有资源中检索命中一篇有关“基因治疗”的医学文献,可以得到基于其题名、作者、主题词、母体刊名等元数据,以及与外部数据源的关联关系,扩展、集成NSTL资源体系中没有的、更为丰富和动态的数据内容。例如,通过检索结果生成的RDF链接,可以获得RDF Book中关于该主题的科学专著的图书信息、图书封面、图书评论;CiteSeer中的参考引用信息;Pubmed医学数据库中的作者信息;BBC、纽约时报等大众媒体的报道;WikiPedia百科全书中基因治疗的术语信息等。也可以利用题名或文摘中的基因名称,将向外部

数据源发出的检索请求结果,将完整的基因、蛋白、药物信息等嵌入检索结果页面,以保证每次检索命中浏览该文献时,可获得与数据源的同步更新。

(2) 异类资源整合检索

异类资源是指资源类型和体裁不同的资源,异类资源整合检索是通过一次检索,将来自不同资源类型的信息对象,按知识节点进行整合,返回给用户关于该对象的所有相关信息的统一视图。

例如,用户检索“干细胞”一词,可以获得NSTL期刊论文、会议论文、科技报告、标准、专利等各类资源中基于同一关键词的各类相关信息,同时,利用该关键词在知识组织系统中的结构和关系,可以进一步扩展得到干细胞治疗、干细胞工程、干细胞分离等在各类文献中的相关信息,从而提供给用户关于“干细胞”的一个完整信息视图;同时,用户无论从检索干细胞的文献、从专门从事干细胞研究的

专家、机构入手,都可以获得干细胞治疗的专利信息、标准信息,并且得到多样化的链接,这些链接构成了该主题的信息情境,具有导航和支持用户认知的作用。

(3) 多维分面组织与检索

利用目前的NSTL资源检索与服务系统,对stem cell一词进行检索,可获得约4万余条检索结果,检索结果命中过多,对于用户的信息选择和信息吸收形成很大干扰,Google等搜索引擎也有类似的问题。分面检索技术试图对上述问题提供一定程度的解决,主要是以事物的属性特征为维度,对信息集合按一定逻辑关系和结构,进行有助于信息获取的划分,以改善用户体验,提高信息的易获取性。

关联数据网络对于分面的实现,有诸多支持能力。首先,通过RDF检索,比较容易地得到某一对象的所有对象特征和属性特征,属性特征不仅包括像时间、类型、数据库、作者、母体等外在特征,还包括主题、分类等内容特征,作为分面的维度;其次,基于知识组织系统,能够获得主题和分类的知识结构和知识关系,以此作为分面依据,能按知识的内容特征对信息资源进行组织。例如,当检索stem cell命中4万条检索结果时,不仅可以按命中文档的时间、作者、期刊、数据库等进行分面,还可以按全能干细胞、多能干细胞、单能干细胞或者胚胎干细胞、成体干细胞,或者按干细胞治疗、干细胞工程、干细胞进展等进行多维分面。

(4) 基于关系的复杂和智能检索

关联数据适于使用基于图模式的关系查询。例如,以“治疗”为谓词进行检索时,不是检索含有

“治疗”的文本,而是检索命中存在“治疗”关系的各种“药物”和“疾病”,或者通过“病状”与“疾病”的关系,检索“治疗”某些“疾病”的“药物”,且一次检索可以使用多重关系,关联多个对象,从而实现和支持基于多种关系的复杂检索。智能检索的实现,则是依赖现有语义和知识关系,通过增加规则描述,利用OWL进行表达,从而在一定程度上实现推理检索,挖掘隐含知识。

3.2 资源服务

(1) 融合与混搭服务

资源的融合与混搭(Mashup)是指在资源整合的基础上,通过增值的信息服务,创造出新的服务形式和服务内容。关联数据有助于实现一次写入、多次复用和跨机构、多类型数据的融合与开放。科学实验数据、生命科学数据、地理信息、文献信息、多媒体信息,通过关联数据这种语义化技术的介入,能够简单、快速地衍生出众多富有创新性的融合和混搭应用。

例如,针对自然灾害的应急处理情报服务项目,可利用NSTL的关联数据及其他LOD云,通过数据整合和相关分析,快速构建某一自然灾害的各类相关信息,如提供地震、台风的预警、时间轴、地图、文档、用户概要、专家信息、三维地图、三维语义空间、风险分析图表等。

此外,基于对知识数据网络中科研领域的研究分析,可以监测知识领域的变化发展情况,分析知识的结构和演化发展过程,揭示科研过程中的学术研究活动关系等,从而得到增值的信息内容,创造出包

括分析评价、趋势预测、决策支持的新型科技文献服务。

(2) 基于数据节点和关系的嵌入式服务

关联数据网络以SPARQL引擎支持上述基于知识节点和知识关系的查询和多维分面的实现,同时,以HTTP URI、RDF、SPARQL的标准接口提供对外的开放使用:1)其他机构可以动态调用上述关联数据,如参引“心脏病”的URI,可以得到其全部RDF链接,不必解析API接口,即可简单地将这一概念及其全部相关链接,实时、无缝地嵌入到自己的信息查询系统、病症诊断系统等应用中,而不必考虑其内容新增或关联变化的维护;2)其他机构也可以通过DUMP接口,下载关联数据网络的所有资源,将原始文献数据和标引数据、本体数据及其关联关系,以RDF格式全部下载,嵌入自己的应用中,可直接开展知识服务和语义查询,而不必再进行知识标注;3)第三方信息机构以同样的方式开放自己的关联数据网络,而由NSTL的关联数据网络发出动态调用请求,将其他数据源嵌入自身的资源体系中,如NSTL的关联数据网络,可以通过参引DBPEDIA百科全书的“心脏病”注释,来扩展自己原有的知识说明。

4 结束语

NSTL关联数据构建与应用意义和作用重大,但在实施过程中,仍需解决和突破一些技术和难点,例如:

(1) 需要解决和突破大规模、海量知识数据的知识关联网络构建问题。其存储、组织、索引和检索的效率与性能,需要不断突破和提

高,具有一定的技术难度。

(2) 知识关联网络的构建,对于知识实体的抽取质量、知识关系的有效性,以及知识组织系统的整体质量水平有很强的信赖性,同时还需建立良好的协同更新维护机

制,是影响整体实现的难点所在。

(3) 有赖于基础性技术的支持,如自然语言处理技术、文本挖掘技术、知识表达和数字对象标识技术、语义数据存储与管理技术发展水平与新进展等。

(4) 关联数据技术本身的难点,如关联数据发布的技术路线选择、关联数据集间的自动关联构建、关联关系的维护更新、关联数据的浏览、使用的跟踪、统计、检索结果排序等。

参考文献

- [1] 乔晓东,梁冰,李颖.National Science and Technology Document Service System in China [J].情报管理,2010(5).
- [2] 刘炜.关联数据的意义与实现[EB/OL].[2010-02-22]. <http://202.114.9.60/dl6/pdf/24.pdf>.
- [3] BERNERS-LEE T. Linked data. World wide web design issues [EB/OL].[2010-02-22]. <http://www.w3.org/DesignIssues/LinkedData.html>.
- [4] QIAO XIAODONG, LIANG BING, YAO CHANGQING.China National Science and Technology Digital Library (NSTL) [J]. D-Lib Magazine,2010(5/6).
- [5] BIZER C, CYGANIAK R,HEATH T. How to Publish Linked Data on the Web [EB/OL].[2010-02-22]. <http://www4.wiwiwiss.fu-berlin.de/bizer/pub/LinkedDataTutorial/20070727/>.
- [6] 张静,马春娥.如何利用D2R发布Linked Data [EB/OL].[2010-04-25].http://www.ibm.com/developerworks/cn/web/1003_zhangjing_d2r/index.html?ca=drs-cn-0325.
- [7] 白海燕,乔晓东.基于本体和关联数据的书目组织语义化研究[J].现代图书情报技术,2010(9).
- [8] 白海燕.DBpedia及其实例分析[J].现代图书情报技术,2010(3).
- [9] SPARQL Query Language for RDF [EB/OL]. [2010-02-22]. <http://www.w3.org/TR/rdf-sparql-query/>.
- [10] 金海,袁平鹏.语义网数据管理技术及应用[M].北京:科学出版社,2010.
- [11] [EB/OL]. [2010-02-22]. <http://esw.w3.org/topic/SweoIG/TaskForces/CommunityProjects/LinkingOpenData>.
- [12] 黄永文.关联数据在图书馆中的应用研究[C]//2010图书馆信息技术的应用、服务和创新学术研讨会,2010.
- [13] RAIMOND Y, SUTTON C, SANDLER M. Automatic Interlinking of Music Datasets on the Semantic Web [C]// LDOW 2008, Beijing, China, 2008.
- [14] 白海燕,朱礼军.关联数据的自动关联构建研究[J].现代图书情报技术,2010(2).

作者简介

乔晓东, 中国科学技术信息研究所总工程师、信息技术支持中心主任, 研究员, 信息学硕士。E-mail: qiaox@istic.ac.cn
白海燕, 中国科学技术信息研究所信息技术支持中心副研究馆员, 管理学硕士。E-mail: bhy@istic.ac.cn
梁冰, 中国科学技术信息研究所科研处副处长, 高级工程师, 计算机博士。E-mail: liangb@istic.ac.cn

Construction of Linked Data and Design of Application Scenes in NSTL

Qiao Xiaodong, Bai Haiyan, Liang Bing / Institute of Scientific and Technical Information of China, Beijing, 100038

Abstract: The construction of linked data in NSTL mainly includes the linkage description in RDF and publishing structured data of NSTL, the consumption of linked data and building linkage with outside datasets, as well as building data network based on science and technology knowledge organization system. The application scenes of linked data in the fields of resource organization, retrieval and service of NSTL mainly include the extending of retrieval results, the integrating and retrieval of heterogeneous resources, the multi-facet organization and searching, the complex and intelligent retrieval based on relationships, mashups of data and services based on linked data and so on.

Keywords: Linked Data, Information organization, Information retrieval

(收稿日期: 2011-04-25)