Digital Library Forum 数字图书馆论坛
http: // www.dlf.net.cn

从关联数据的应用场景看信息服务发展的新动向*

□ 白海燕 / 中国科学技术信息研究所 北京 100038

摘要：文章分析关联数据的应用，对比传统信息服务，归纳总结信息服务发展的新动向，即：从资源组织技术来看，出现了从文献资源整合向数据融合与跨界混搭的新发展；从检索技术来看，出现了从分面浏览、高级字段检索向基于对象和关系的语义检索的新发展；从支持第三方和嵌入式服务来看，出现了从开放接口向支持开放数据、开放服务的技术新发展。

关键词：关联数据，数据融合，语义检索，开放数据，开放服务
DOI：10.3772/j.issn.1673—2286.2012.06.009

引言

关联数据是指Web上通过RDF和可参引URI，发布、共享和联接各类数据的技术方法，关联数据技术所具有的框架简洁、标准化、自助化、去中心化、低成本的特点，为构建人机理解的数据网络，提供了强有力的支持，为实现语义网远景目标奠定了坚实的基础，关联数据是语义网领域近年来的新方向和研究热点，并在短短数年间取得显著成效[1]。

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。关联数据在数据层建立了链接机制，准确描述数据的结构信息，并通过URI来确保机器能够自动链接各种数据，为信息聚合的智能化、语义化处理提供了基础。因此，关联数据的众多应用，向我们展示了信息服务发展的新动向：从资源组织技术来看，出现了从文献资源整合向数据融合与跨界混搭的新发展；从检索技术来看，出现了从分面浏览、高级字段检索向基于对象和关系的语义检索的新发展；从支持第三方和嵌入式服务技术来看，出现了从开放接口向支持开放数据、开放服务的新发展。

1 数据融合和跨界混搭

现代信息服务在经历数字化、网络化发展阶段之后，开始走上跨平台系统、跨组织机构的全球集成化发展道路。关联数据支持细颗粒度、结构化数据的跨平台、跨系统、跨机构、跨领域的灵活调用、组合与嵌套，促进资源融合和跨界混搭机制的形成，带来信息组织与信息服务范式的变革[2]。

（1）PubMed与bio2RDF的整合手段对比

PubMed①是美国国立医学图书馆的医学文献搜索引擎，近年来出现了整合科学数据的新功能。例如，查询基因“Nur77”，PubMed在文献库和基因数据库中同时进行检索，分别反馈：命中文献库459条记录，获得基因注释信息“nuclear receptor subfamily 4,group A .member 1(Homo sapiens)” 1条，以从基因库获得的相关信息如基因功能等作为扩展检索条件，命中文献记录152条。如图1所示，文献记

* 本文为十二五国家科技支撑计划项目“面向外文科技文献信息的知识组织体系建设与应用示范——基于STKOS的知识服务应用示范”课题的子课题“科技信息资源关联数据服务应用示范”的研究成果之一（编号：2011BAH10B06—06）。
① http://www.ncbi.nlm.nih.gov/PubMed/.

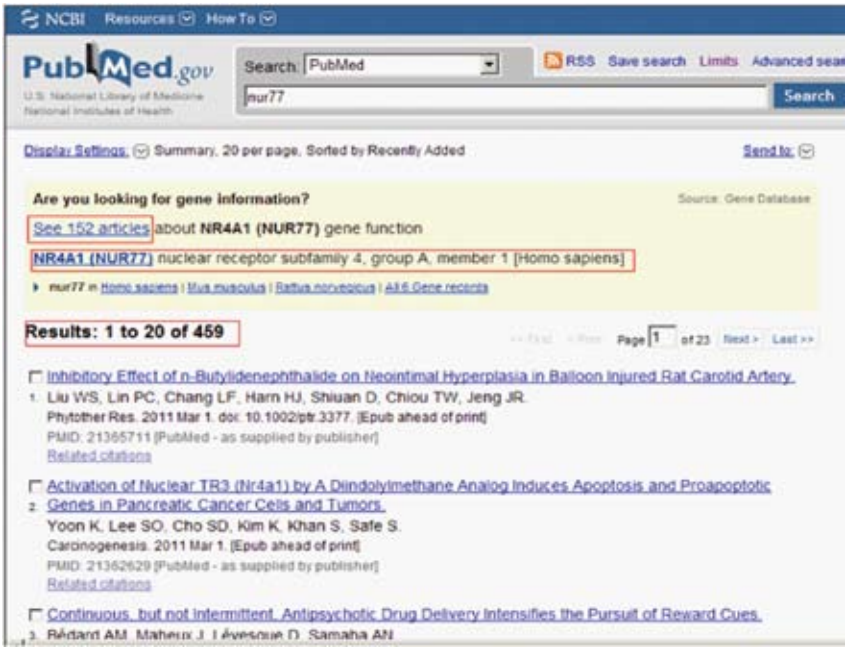

图1 PubMed文献记录与科学数据的整合

录与科学数据整合在了同一检索结果界面，并提供了指向具体信息源的引导链接，但整合的层次仅仅是资源库和文献类型，缺少基于事物本身的细颗粒度整合和相关知识、逻辑关系的整合。

Bio2RDF[②]是由Genome Canada/Genome Quebec资助的生命科学整合系统，使用关联数据技术提供生命科学数据的内在关联，以支持知识发现。其整合规模超过40多个不同类型的生命医学数据源，包括PubMed、MeSH、Gene Ontology、UniProt、UniParc等。Bio2RDF为不同的数据源提供者建立命名空间列表，使得创建标准

图2 bio2RDF的整合检索结果

的URI成为可能，并分析各个数据源，用RDF模型进行表示，同时利用RDF转换工具将各个数据源信息转换成统一的RDF格式，目前已有超过300亿的三元组[3]。如图2所示，查询蛋白“Q16665”，得到有关该蛋白在各种数据库的相关属性信息和相关对象链接。

Bio2RDF不仅实现异构数据库的整合检索，同时对检索结果在语法层面和语义层面，基于命名一致、相同属性值或领域知识模式，进行了数据融合。如图3（a）所示，在语法层基于命名一致性实现数据融合，来自不同数据源的同一对象Q16665，基于相同的命名，合并为统一视图；图3(b)为语义层基于相同属性的数据融合[4]。

（2）BIS的研究基金浏览器（BIS Research Funding Explorer）

研究基金浏览器（http://bis.clients.talis.com/）是英国商务、创新和技能部（Department of Business Innovation and Skills，BIS）、英国研究理事会（Research Councils, UK）、知识产权局（Intellectual Property Office，IPO）、技术战略委员会（Technology Strategy Board）的联合项目，通过整合多种类型的数据，基于可视化工具提供统一的科研投入与产出视图[5]。

该项目整合的数据包括学科领域数据、地理区域、1500家组织机构（学术和商业机构）数据、2000年至2018年这些机构的科研项目、基金和专利产出数据。这些不同类型的数据以关联数据的形式进行发布和融合，通过可视化工具FLASH和数据融合工具GoogleMaps构建了多种形式的数据表现，见图4，包

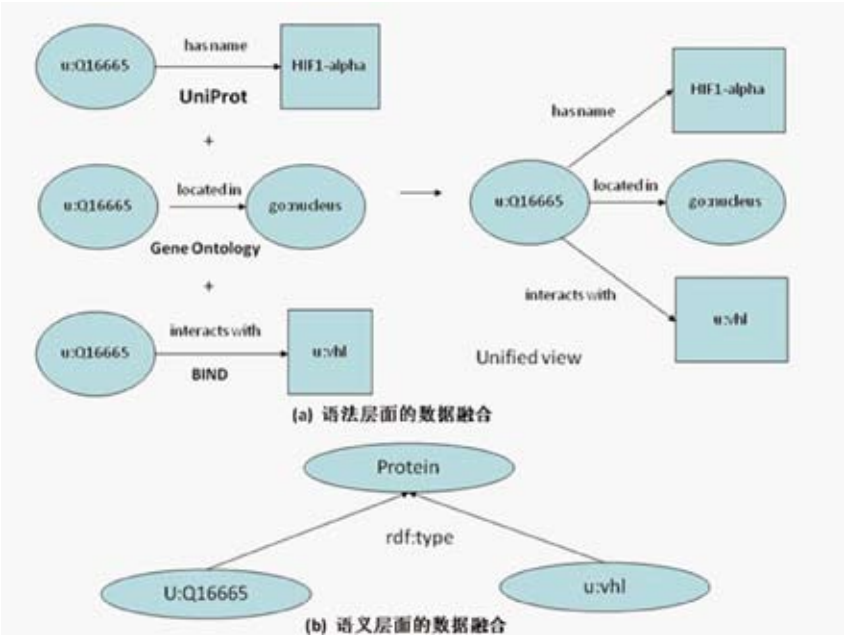

图3 bio2RDF的数据融合模式[4]

括：1）科研投入的时间与区域分布及强度：随着时间条的变化，在地图上以动画方式显示投资的地理位置，并通过颜色的出现、消失、深浅变化，显示科研投入的强度及变化；2）不同领域的科研投入与产出：通过选择某一领域或全部领域，随着时间的变化，在科研产出区域和科研投入区域，分别以不同颜色代表不同领域，显示投资的增长和相应的专利数量增长情况。该项目有助于投资者或决策管理人员，迅速直观地了解英国各地区的科研投入与产出现状与发展。该项目是英国政府关联数据运动背景下的应用，除此之外，还有多个项目采用了类似的以综合视角发布的数据融合应用。例如英国ITO（Visualising Traffic Data[③]），将英国政府网站提供的交通关联数据与车辆信息相融合，显示道路的繁忙程度；美国Facts about Transportation Enerygy[④]使用美国预算关联数据和包含政府能源消耗的关联数据源进行数据融合的应用。

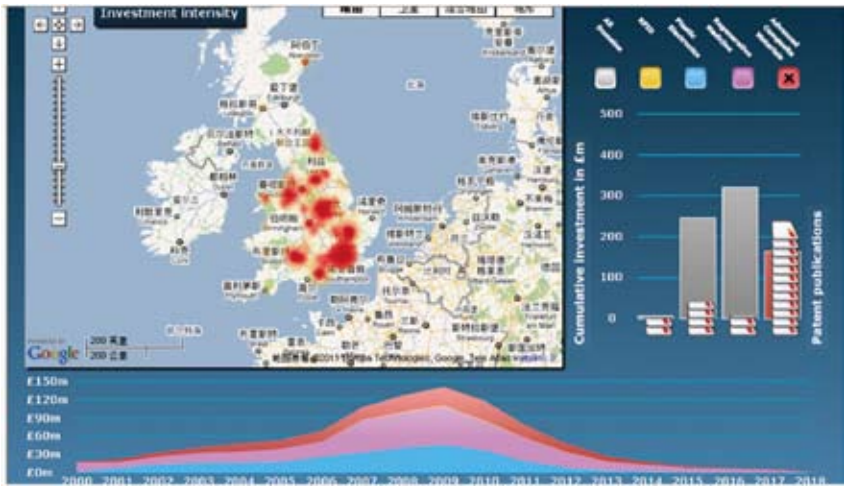

图4 BIS的研究基金浏览器

从上述应用中不难看出，传统文献资源整合出现了一些新的发展动向，具体包括：1）整合资源组织对象扩大，从单一的文献资源对象，扩大到所有使用URI标识的事物，包括信息对象和非信息对象；2）整合的颗粒化程度提高，从库、文件、记录、资源片断到数据，即以对象的概念划分资源；3）资源组织以结构化对象的属性和关系进行语法和语义层面的合并，打破了单一的结构限制，可以在数据表现层进行手段丰富的组织和服务构建，如可视化工具的应用等，并基于数据融合和数据表现实现服务增值。

2 基于对象和关系的语义检索

利用关联数据的类数据和实例数据，在检索方面可以实现基于对象和关系的语义检索，在应用上表现为动态分面检索、对象检索和关系检索。

（1）DBPedia的分面检索 Faceted Wikipedia Search

Faceted Wikipedia Search[⑤]是Wikipedia英文版的交互式检索界面，见图5，主要包括以下功能区[6]：

③ http://www.itoworld.com/static/gallerytraffic.html.
④ http://data-gov.tw.rpi.edu/demo/linked/demo-10029-fuel.html.
⑤ http://dbpedia.neofonie.de/browse/.

● 文本查询区：用户可以通过该区域输入自由文本进行查询；

● 分面导航区：最相关的分面条件在该区域显示，用户可以通过选择或输入属性值确定分面的过滤条件；

● 过滤区：显示选中的分面条件和检索词，用户可以通过点击相应的删除按钮，去掉某些条件；

● 检索结果区：匹配命中的Wikipedia文章，包括题名、作者和文章中的图像；

Faceted Wikipedia Search能够基于查询对象的属性特征，提供有针对性的分面浏览和导航。与传统分面浏览提取通用元数据作为分面条件不同，Faceted Wikipedia Search采用的是结构化对象的属性提取方法。例如查询“stream”，命中结果的分面条件是“stream”这一对象所具有的属性特征，即 “country ”、“River mouth”、“is in district”、“latitude”、“longitude”等；而查询“person”，则分面条件为“born in”“born in year”“genre”等。依据特定对象的属性，提取分面条件，具有很好的灵活性和动态性。

图5 DBpedia的分面检索界面

Faceted Wikipedia Search还提供分面条件的查询交互，能够针对对象的多重属性特征，实现例如“哪些河流流入莱茵河且长度超过50公里？”“中国2000年前建造的超过50层的摩天大楼有哪些？”一类的复杂查询。例如查询需求为“Rivers that flow into the Rhine and are longer than 50 kilometers”，其查询表达则为http://dbpedia.neofonie.de/browse/rdf-type:River/riverMouth:Rhine/length~:50000~/?fc=30，即指定查询对象的类型（rdf-type）为“River”这一实体，并对其属性进行限定：{rdf-type=River, river.property=rivermouth river, mouth.value=Rhine}。与字段匹配不同，这是一个三元组匹配，这种匹配是一种图模式匹配，即三段结构（主语、谓词、宾语）的匹配。图模式之间可以进行组配，如果

P1={ rdf-type=River river.property=rivermouth rivermouth.value=Rhine }

P2={ rdf-type=River and river.property=length and length.value=50000}

则P1和P2可以使用Union、And、Opt和Filter来进行递归定义，即通过这些操作符，将简单图模式组合成复杂图模式[7]，从而支持针对事物的复杂属性特征的复杂检索和问答式检索，需求表达更加灵活、直观，检索结果更加精准。

（2）DBpedia的关系检索

实际上，对象的属性特征可分为数据属性（data property或attribute）和对象属性（object property）。数据属性是对象自身的属性特征，其值为具体某一文本或数值；对象属性是对象与其他对象的关系，其值为某一对象。上述案例是针对数据属性的检索，而针对对象属性的检索则可称为关系检索。

如图6所示的为DBpedia的关系检索构造器。当前查询需求为“查找球衣号为11号，效力于某一俱乐部，该俱乐部体育场拥有4万以上座席，且该球员出生在一个人口100万以上的国家”。该查询需求涉及多个对象：球员、俱乐部、国家等，每个对象均具有特定的属性，并且对象之间也存在某种关联关系。DBpedia的关联查询构建器，提供了用户描述和表达这些对象、属性和关系的方式，即以三元组和多三元组组合来表达需求，而DBpedia的数据也同样以三元组的方式描述和表达这些事物，二者进行图模式匹配，可以得到精准、直接的命中结果[8]。

在本例中，用户可以在查询构

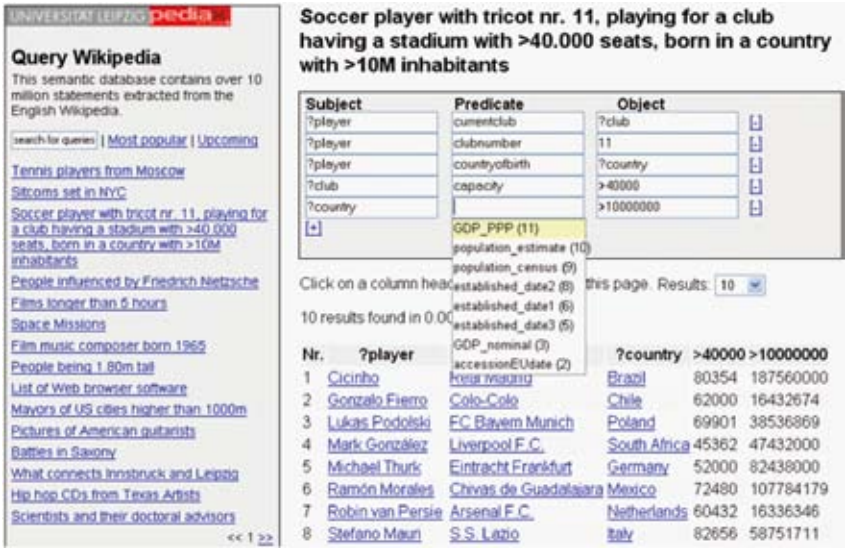

图6 DBpedia的关系查询构造器[8]

造器的subject一栏中选择或输入对象，而在predicate一栏中选择相应的对象或数据属性，并相应地在object一栏中，输入具体属性值值或关系所联结的对象，同时为对象指定相应的属性和值。通过图中的输入值可见，含有具体文本值的为数据属性，而含有”?”的代表某一对象。可见，通过不同对象之间的关系匹配，对象属性匹配，及两者的多重组合，能够实现更为智能和复杂的检索需求。

图7 参引DBpedia的应用⑦

DBpedia的对象检索和关系检索，是语义检索的具体场景应用，支持用户表达复杂的查询需求，精确定位并给出准确答案，代表了信息检索发展的新方向。

3 基于开放数据和开放服务的嵌入式应用

关联数据本身是W3C倡导的“开放数据运动”的产物，具有数据接口简单易用、数据格式标准化、结构化的特点，开放服务目前主要的应用场景包括数据的嵌入式应用和数据的语义服务。

（1）数据嵌入式应用

关联数据对嵌入式应用的支持方式主要包括URI参引、SPARQL查询和DUMP下载的本地应用[8]。

以DBpedia的URI参引和URI查询应用为例，DBpedia是wikipedia的关联数据集，词条信息不断增加和更新。用户或其他服务可以通过参引“柏林”的URI，将其全部信息嵌入自身的网页或服务中，并保持与数据的同步更新。如图7所示,Fluidops Information Workbench⑥在其服务中，参引了DBpedia的URI，并按其自身需求，对数据表现形式进行变化，提供了多种视图形式。

为了支持这种URI参引的数据嵌入式服务，DBpedia还提供了相应的URI查找服务-DBpedia Lookup，即通过用户提供的关键词查找相关的URI，实现方式包括关键词查询（Keyword search）和词

⑥ http://iwb.fluidops.com/resource/.
⑦ http://iwb.fluidops.com/resource/Berlin.

前缀查询（Prefix search）。例如，查询“Berlin”的URI，关键词查询格式为：

http://lookup.dbpedia.org/api/search.asmx/KeywordSearch?QueryClass=place&QueryString=berlin

前缀查询格式与其类似，同时具有输入词的自动完成功能，可只提供输入词的部分词首，如：

http://lookup.dbpedia.org/api/search.asmx/PrefixSearch?QueryClass=&MaxHits=5&QueryString=berl

查询结果格式为XML文件，内容包括URI，标签、简短的描述，类型、分类和在Wikipedia的内部链接引用次数。[9]

SPARQL查询方面，DBpedia数据集本身提供SPARQL的查询端http://DBpedia.org/sparql，使用OpenLink Virtuoso作为后台数据库引擎，同时用户还可以使用其他工具查询DBpedia，如Leipzig query builder⑧, OpenLink Interactive SPARQL Query Builder (iSPARQL)⑨,SNORQL query explorer⑩t和其他SPARQL敏感的客户端，通过构造和表达检索需求，获得相应的数据，嵌入自身的应用。

（2）语义数据应用

各领域关联数据集的结构化数据实际上是由本体、实例和组织模型构成的知识库，因此，能够基于自身的知识数据优势，通过开放化的工具，提供语义化服务。语义标注是目前关联数据集提供的典型服务。

以路透社的OPEN CALAIS为例，在其关联数据基础上，通过自然语言处理技术，提供了语义标注服务。例如，用户输入一段自由文本，“President Obama called Wednesday onCongress toextend atax break forstudents included inlast year's economic stimulus package, arguing that the policy provides more generous assistance”，提交后获得相应的主题（topic）、社会标签（Social Tags）、实体（Entity）等标注结果。如图8所示，该文本的主题为“politics”，主要实体有“organization=Congress”，“Person=Obama”，“Position=President”。

DBpedia也提供了类似的工具DBpedia Spotlight[11]，能对自由文本中所涉及到的DBpedia概念进行自动标注，为非结构化信息资源通过DBpedia关联到关联数据云（Linked Data Cloud）提供了实现基础。该工具可执行命名实体的抽取，包括实体检测和命名消岐，也可以在其他信息抽取任务中构建自己的命名实体识别解决方案。

DBpedia Spotlight的具体功能包括：实体标注、实体消岐和最佳推荐，其功能模块包括：1）web application模块：提供HTML/Javascript界面，允许用户在WEB浏览器中输入或粘贴文本，获得文本的标注结果；2）Web Service模块：提供RESTful/SOAPWeb API，实现文本的标注和消岐功能；3）Annotation Java/Scala API模块，提供执行标注和消岐的基本推理方法；4）Indexing Java/Scala API模块：执行标注和消岐算法所必需的数据处理；5）评价模块：测试消岐效果、记录结果日志，使用这些

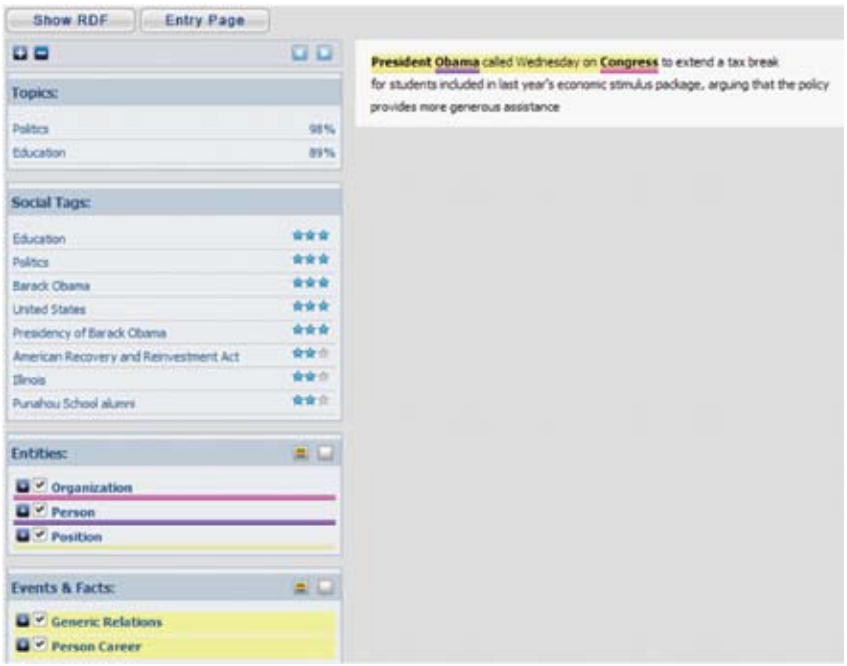

图8 CALIAS的语义标注结果

⑧ http://querybuilder.dbpedia.org.
⑨ http://dbpedia.org/isparql.
⑩ http://DBpedia.org/snorql.
http://dbpedia.org/spotlight

评测结果用于该工具进一步的功能训练。[10]

对于上例中的自由文本，该工具的WEB使用格式如下：

http://spotlight.dbpedia.org/rest/annotate?text=President Michelle Obama called Thursday on Congress to extend a tax break for students included in last year's economic stimulus package, arguing that the policy provides more generous assistance.&confidence=0.2&support=20

其运行结果如下图所示，标注出众多命名实体，并提供指向DBpedia的链接。

图9 DBpedia Spotlight的标注结果

文献资源服务的发展方向之一，是面向最终用户和代理，提供嵌入科研过程和构建科研情境的服务。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，而更重要的是将经过有效组织、深度序化和知识化整序后的资源，以更细的颗粒程度，更结构化的数据关系和具有语义的知识，动态时实，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。

结语

关联数据技术的应用，对于实现资源组织的充分结构化和深度序化，对于提升信息资源整体的知识化组织程度具有重要作用；基于关联数据构建信息服务，有利于实现语义数据发布、数据融合与复用和智能开放获取，有利于构建支持扩展、整合和混搭的新型信息服务平台，更高效、更智能地满足各类信息需求。十二五期间，国家科技图书文献中心将在科技知识组织体系（STKOS）及相关工具的基础上，通过多资源集基于STKOS、基于属性特征、基于对象同一性等多种关联模式，向用户提供多个分布式异构数据源的整合和关联访问，将来自不同数据源的同一个信息对象，按STKOS中的概念进行整合，返回给用户关于该关联对象的所有相关信息的统一视图。

参考文献

[1] 刘炜.关联数据:概念、技术及应用展望[J].大学图书馆学报,2011(2).
[2] 李春旺,肖伟.集成融汇:概念、模式与应用[J].现代图书情报技术,2007(12).
[3] DUMONTIE M.Bio2RDF: A biological knowledge base for the Semantic Web[EB/OL].[2012-03-16].http://www.slideshare.net/micheldumontier/bio2rdf-a-biological-knowledge-base-for-the-semantic-web-1603939.
[4] BELLEAU F, NOLIN M A, et al. Bio2RDF:Towards a Mashup to Build Bioinformatics Knowledge Systems[J]. Journal of Biomedical Informatics, 2008, 41(5): 706-719.
[5] LOGAN A.Exploiting U.K. Government Linked Data.[EB/OL].[2012-03-16]. http://www.abdn.ac.uk/~csc323/teaching/abdn.only/CS5915/information/CS5577-Logan A J.pdf.
[6] HAHN R, BIZER C, SAHNWALDT C, et al.Faceted Wikipedia Search [C/OL]// 13th International Conference onBusiness Information Systems (BIS 2010), Berlin, Germany, May2010.[2012-03-16].http://www4.wiwiss.fu-berlin.de/bizer/pub/hahn-etal-faceted-wikipedia-search-BIS2010.pdf.
[7] 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.
[8] AUER S,BIZER C, el al. DBpedia: A Nucleus for a Web of Open Data[EB/OL].[2012-03-16].http://it.haitianyuan.com/listen/b/wp-content/uploads/2010/09/DBpedia-A-nucleus-for-a-web-of-open-data.pdf.
[9] DBpedia Lookup Service[EB/OL].[2012-03-16].http://wiki.dbpedia.org/Lookup?v=1e13.
[10] DBpedia Spotlight – User's Manual[EB/OL].[2012-03-16]. http://wiki.dbpedia.org/spotlight/usersmanual.

作者简介

白海燕，中国科学技术信息研究所信息技术支持中心，研究馆员。E-mail: bhy@istic.ac.cn

New Trends of Information Services from the View of Application of Linked Data

Bai Haiyan / Institute of Scientific and Technical Information of China, Beijing, 100038

Abstract:This paper concludes new trends of informationservices based on analysis of application oflinked data in contrasting to traditional information services. These new trends include new development from document resources integration to data fusion and mashup in field of information organization, new development from facet navigation and advanced search to semantic retrieval based on object search and relationship search in field of information retrieval and new development from open interfaces to open data, open services in field of supporting the third-party and embeddedservices.

Keywords:Linked data, Data fusion, Semantic retrieval, Open data, Open services

（收稿日期：2012-03-17）