面向关联数据的信息检索服务研究综述*

黄永文[1]　钱　力[1,2]
[1](中国科学院国家科学图书馆　北京 100190)
[2](中国科学院大学　北京 100049)

【摘要】简要介绍图书馆领域最近开展的一些关联数据活动,提出关联数据可以应用到信息检索服务中的许多方面,如改善用户相关词检索服务、改善图书馆 OPAC 检索服务、丰富检索结果的内容类型、进行检索结果排序以及构建人员的研究网络,并从关联数据的 HTTP URI 访问定位技术、关联数据的动态查询技术、关联数据链接关系的发现技术、关联数据的整合及索引技术和展示技术 5 个方面分析基于关联数据实现信息检索服务的相关技术。
【关键词】关联数据　信息检索　图书馆服务
【分类号】G250.76

Research on Information Retrieval Service Towards Linked Data

Huang Yongwen[1]　Qian Li[1,2]
[1](National Science Library, Chinese Academy of Sciences, Beijing 100190, China)
[2](University of Chinese Academy of Sciences, Beijing 100049, China)

【Abstract】This paper introduces recent activities about linked data in the field of library, and brings forward that linked data can be applied to many aspects of information retrieval services, such as improving user's associated word retrieval services, improving the library OPAC retrieval services, enriching the content type of results, sorting the results, building personnel networks, and then analyzes five related technologies about realizing information retrieval service based on linked data from the HTTP URI access technology, the dynamic query technology, discovery technology of link relationship, integration technology, indexing techniques and display technology.
【Keywords】Linked data　Information retrieval　Library service

1　引　言

自 2006 年 Berners-Lee[1] 提出关联数据以来,关联数据得到了广泛的关注。关联数据有可能改变图书馆创造、获取、发现信息的各个方面,图书馆界在关联数据发布和应用方面正在进行尝试。2011 年以来图书馆开展了一系列非常重要的关联数据活动,W3C 图书馆关联数据孵化组发布了最终报告[2],美国国会图书馆先后发布主题标目、人名规范等 15 个词表[3],并宣布了数字化时代基于关联数据的新书目框架[4],OCLC 将主题术语的分面应用(Faceted Application of Subject Terminology, FAST)发布为关联数据[5],Kuali OLE 计划对共享知识库或者联盟知识库提供结构化和开放关联的数据服务[6],大英图书馆把书目转换成关联数据之后计划将期刊文章也发布成关联数据[7],欧洲国家图书馆的数字资源门户 Europeana 在 2012 年 2 月发布了关联数据试点(Linked Open Data Pilot)[8],开放数据类型涉及文本、图像、视频和音频。可以看出目前的关联数据源,除了包含书目数据、规范文档

收稿日期:2012-10-09
收修改稿日期:2012-11-13
*本文系国家社会科学基金资助项目"我国数字图书馆集成融汇服务方法研究"(项目编号:10BTQ004)的研究成果之一。

之外,还逐渐扩展到文章元数据、知识库、多媒体等类型,这些高质量的数据源可以作为图书馆开展用户服务的基础,特别是信息检索扩展服务。本文主要从信息检索服务出发,对国外关联数据应用的相关研究、实践以及所涉及的技术进行系统的梳理,希望对国内开展关联数据应用方面的工作起到抛砖引玉的作用。

2　基于关联数据的主要信息检索服务

W3C 图书馆关联数据孵化小组的报告中提到:当结构化数据之间的关联变得更加丰富时,用户才可能意识到发现和使用信息资源的能力提高,跨图书馆和非图书馆信息资源的导航变得更加完善[2]。关联数据可以应用到信息检索服务中的许多方面,如改善用户相关词检索服务、改善图书馆 OPAC 检索服务、丰富检索结果的内容类型、进行检索结果排序、构建人员的研究网络等。

2.1　利用关联数据改善用户相关词检索服务

利用词表和规范文档类的关联数据可以构建检索词自动提示和拼写建议,以及实现相关词扩展检索服务,改善现有检索系统的服务方式。目前,检索词自动提示和拼写建议主要是利用用户检索词(如搜索引擎)以及利用元数据中的关键词来实现的,与现有搜索引擎中的词汇提示功能相比,从知识组织系统中获取的词汇具有更丰富的语义联系(同义、等级、相关),质量更高[9]。利用词表和规范文档类关联数据,不仅可以实现概念浏览和语义检索,还可以增强搜索引擎的功能,实现检索词自动提示和拼写建议以及检索结果的自动归类,对检索词进行扩展,给用户提示出其他相关检索词。例如宾夕法尼亚大学将关联数据形式的 LCSH(Library of Congress Subject Headings)整合在 OPAC 服务中,提供上位词和下位词扩检服务,用户可以方便地进行扩检和缩检。

目前,有许多知识组织体系已经发布为关联数据,如 LCSH、DBpedia、YAGO、UMBEL、GeoName 等,这些数据集提供了大量概念、概念之间关系以及具体的事实,这些都可以作为自动提示和检索词扩展的基础。在美国国会图书馆发布的主题标目 LCSH 中,包括了不同形式的词(Variants)、上下位类(Broader Terms、Narrower Terms)、相关词(Related Terms)。DBpedia[10] 是 Wikipedia 关联数据的结构化的表现,提供了大量跨领域的知识库。YAGO[11] 涵盖了 Wikipedia 和 WordNet 的大部分内容,YAGO 经过人工修正,其知识组织的正确性达 95% 以上。UMBEL[12] 将 OpenCYC Ontology 中的 2 万主题概念类目及其相关关系抽取出来。

2.2　利用关联数据改善图书馆 OPAC 检索服务

目前,OPAC 仍然是用户检索和查询图书馆资源的有效方式。随着用户需求的不断改变,许多厂商也在不断改善图书馆 OPAC 的服务效果,推出下一代图书馆资源发现系统,有的厂商已经开始或者正在探索将关联数据融合到图书馆资源发现系统中。例如,Capita 公司的 Prism 系统[13] 将 MARC21 格式的记录转换成关联数据格式,通过云的软件服务模式提供接口,允许其他应用服务构建在它之上。Prism 目前主要是处理图书馆书目数据,也正在准备扩展到文摘、电子期刊文章、学位论文库,以及与当地活动联系在一起(如搜索与健康有关的讲习班);Serials Solutions 的 Intota 系统[14] 采用关联数据模型,来实现网络权威控制文档和社区编目发展;Ex Libris 公司的 Alma 系统[15] 从发布、消费和整合三方面增加对关联数据的支持,采用与关联数据原则相互兼容的方式来创建、处理数据,并在图书馆环境中使用外部的关联数据源。

采用关联数据原则之后,与 MARC 格式相比,图书馆可以采用更加灵活的方式存储数据、在 Web OPAC 上查询和展示书目数据,能够从不同的方面来展现信息,如作者、标题、出版商以及用户检索到的重要信息,改变传统以记录为中心的服务方式。用户可以自己选择浏览导航服务,以发现图书馆资源。另外,还出现了一些直接利用图书馆提供的开放书目数据,提供跨图书馆的书目检索应用,如 Lodopac 等。Lodopac[16] 是简单的开放关联数据 OPAC 应用,目的是提供标准 OPAC 界面来检索 RDF 格式的不同书目数据源,而无需知道不同数据源的数据格式和掌握 SPARQL 查询语言。

2.3　利用关联研究数据丰富检索结果的内容类型

图书馆为最终的研究用户服务,需要提供各种必要的背景资源,并尽可能与现实世界联系起来,如人员、组织、数据、地点等。由于文献和研究数据通常存储在分离的信息系统中,科研人员为了使用研究数据,需要手动解析全文以便找到可引用的研究数据。虽然科研人员尽努力获得了这些数据,但是这些研究数据的引用可能是不一致和不完整的,这在一定程度上阻

碍了研究数据的有效检索和利用。

通过关联数据,可以实现文献信息和非文献信息的有机嵌入和融汇,如科技文献与科学数据、地理信息、科研项目、学术会议、基金信息、社交网络等,这些资源都是用户所需要的,特别是各类研究数据。由德国 GESIS、曼海姆大学和曼海姆大学图书馆共同承担的整合社会科学领域的研究数据和文献出版物的 InFoLiS 项目[17],目的是建立文献出版物和研究数据之间的关联,将由此产生的相互关联的数据集发布为关联数据,并提供一个综合的信息搜索服务;RPI TWC 项目组与 Elsevier 公司相互合作[18],将在 Elsevier 的 SciVerse 网站上增加对 Data. gov 数据的访问,使研究人员在科研过程中更容易发现相关的政府数据。RPI TWC 项目组从语义方面对政府数据进行了丰富,并开发了美国政府数据的检索工具,也将嵌入到 SciVerse 网站上;由英国皇家化学学会(RSC)和南安普敦大学合作[19],计划将 RSC 的免费化学结构数据库 ChemSpider 发布为关联数据。ChemSpider 包括 250 万个化学结构、属性以及相关信息,并提供与 RSC 文献之间的链接。

2.4　利用流通关联数据进行检索结果排序

用户的访问数据和流通借阅信息是判断图书的相关性、权值以及利用率的重要指标之一。目前,已经出现一些关注于流通活动数据的研究。Pfeffer 等[20]研究图书馆流通关联数据的发布和消费,认为流通数据对于信息检索和资源发现具有重要作用,可以将流通数据的统计信息作为图书重要度的判断依据,并进行了将图书借阅信息转换成 RDF 格式的实验;JISC 资助的哈德斯菲尔德大学 Library Impact Data 项目[21],采用用户的活动信息来确定利用率高的图书馆资源和利用率偏低的学科领域,以便有针对性地改善图书馆服务。

一些图书馆也积极参加到流通数据的开放关联发布活动中,为跨馆对比读者流通数据,发现不同图书馆的读者借阅趋势、评估馆藏的发展战略等提供基础。在 2008 年底,英国的哈德斯菲尔德大学在开放数据共享许可协议下共享了跨越 13 年的读者流通数据和读者荐购数据,为其他图书馆提供关于读者借阅情况的匿名信息集的下载[22];LibraryCloud 是开放的多图书馆数据服务,目前参加的图书馆有:哈佛大学、旧金山公共图书馆等。LibraryCloud 集成不同图书馆的元数据,并将其发布为关联数据,目的是使图书馆资源进入到 Web 生态系统中。目前,LibraryCloud 只包括书目和馆藏数据,将来会增加流通、评价和评论等信息[23]。

2.5　利用关联数据构建人员的研究网络

充分利用关联数据源中的关联关系,实现研究人员、所属机构、研究成果、参加会议、科研项目等多类型内容的整合和集成,构建科研人员合作和研究网络,可以在信息检索过程中辅助发现科研人员之间的共同研究领域,揭示潜在合作关系以及帮助科学研究的协作,增强人员之间的研究交流、展示研究成果等。

这方面的研究和实践主要有:VIVO、RKB、Archives Hub Linking Lives、CAF-SIAL 等。康奈尔大学图书馆将 VIVO 数据以关联数据的形式发布[24],涵盖康奈尔大学、佛罗里达大学、印第安纳大学等 8 家高校的科研人员数据,内容包括人员的名字、头衔、研究领域、所属机构、活动和产出、研究资助、谈话、课程等,构建了国际性的科学家网络,帮助发现跨学科、跨地域的科学家之间的协作。RKB Explorer[25]目的是将不同来源的数据进行标准统一化的展示,提供语义化服务,支持人员、公开文献、组织机构、项目等的搜索,采用整合的方式显示与之相关的人员、文章、活动、成果、项目、课程等信息。Archives Hub Linking Lives[26]探索关联数据展示方式,在人物和事件之间建立连接,形成完整的个人简历。

3　实现基于关联数据的信息检索服务的主要技术

3.1　关联数据的 HTTP URI 访问定位技术

所有使用 HTTP URI 和 RDF 描述的关联数据集就像一个无限关联的 RDF 图,用户和应用系统通过 HTTP URI 链接可以实现无缝浏览,关联数据对图书馆用户的价值源于这种基本的导航原则[2]。HTTP URI 主要有 303 URI 和 Hash URI 两种策略,这两种策略可以确保现实世界的对象和描述对象的文件不被混淆,人和机器可以获取到各自可以阅读的表达格式(如 HTML、RDF)[27]。

在 303 URI 的策略中,服务器以 HTTP 303 代码和描述现实世界对象的 URI 来响应客户端。然后,客户端参见这个新的 URI 来获取描述对象的 Web 文档。由于 303 URI 策略存在一个问题就是需要客户端发出两次 HTTP 请求,才能获得对象的描述信息。为了避

免这个问题，Hash URI 策略应运而生。当客户端检索 Hash URI 时，HTTP 协议需要在向服务器端请求 URI 之前，先剥离“#”之后的部分。这两种策略各有优缺点，303 URI 通常用于资源描述非常大的数据集合中（如 DBpedia 中的概念描述），Hash URI 通常使用在 RDF 词表的术语标识上，因为 RDF 词表的定义文件一般都比较小，方便客户端应用一次获取到整个词表。当采用 RDFa 将 RDF 嵌入到 HTML 页面时，通常采用 Hash URI 方式，以避免与 HTML 页面中的资源 URI 混淆。也可以将 303 URI 和 Hash URI 两种策略相互结合起来。例如，采用 http://domain/resource#this 的 URI 定义方式，可以灵活配置返回的资源描述，同时避免多次发送 HTTP 请求。

3.2　关联数据的动态查询技术

面对关联数据的多样性、区域性以及规模性等特点，在动态查询方法的实施中采取不同的技术方案。目前主要有三种查询技术：面向已知的单一资源采用 SPARQL 查询、面向已知的多个资源采用跨数据集的统一查询、以及面向未知资源采用基于多维索引与语义索引的搜索引擎技术。

（1）利用 SPARQL 标准语言查询。SPARQL 是解析 RDF 数据的通用标准语言，发布 SPARQL Endpoint 提供 Web 服务，通过识别 URI 唯一标识，来查询并发现关联数据，初步实现关联数据服务的实际应用。SPARQL 查询语言主要是基于图模式匹配[28]，综合利用基本图模式、组合图模式、可选择图模式、联合图模式、RDF 数据集图模式、值约束条件 6 种模式中的一种，通过匹配、分组、连接、合并、过滤等形式，根据用户定义有效地返回映射结果集。

（2）跨数据集的统一查询。面对海量的 RDF 数据，实现跨数据集的关联发现，在效率以及排序方面是一个挑战。文献[29]针对上述的问题，提出了三种查询方法，同时进行了评估，认为基于中央仓储的查询方法效果最好。基于中央仓储的查询方法以完整的 RDF 中心仓储数据向终端用户提供服务，消除了网络之间的消耗；而基于多仓储的联合查询与基于中央仓储的查询方法相比，存在从 RDF 数据集中载入数据到统一查询层的中间步骤，需要一定的额外消耗，而且在统一模型中进行更新也是一个难题；基于多 SPARQL 端点的联合查询方法是从多个 SPARQL 端点中查询并获取数据，通过 SPARQL 获取比从原始接口中获取存在一定的局限性。

（3）基于多维索引与语义索引的查询。SPARQL 端点服务是目前比较常见的关联数据查询方式，不过该方式存在响应速度慢的问题，因此有许多学者通过建立 RDF 索引来提高查询速度。RDF 数据是由结构化的知识单元组成，可以对每一类知识单元及关系建立索引，同时结合知识组织体系，构建多维索引机制。SIREn[30]是一个基于 Lucene 索引的高效语义信息检索引擎，实现查询处理、实时更新、全文检索以及使用 Shard 实现分布式索引等关键技术，并支持语义索引与语义查询。

3.3　关联数据链接关系的发现技术

从技术上来说，外部的 RDF 链接是一个 RDF 三元组，目前主要有三种 RDF 链接关系类型[27]：

（1）关系链接（Relationship Links）指向其他数据源中相关的事物。例如，关系链接可以指向人的背景信息（如出生地、居住地等）。

（2）身份/标识链接（Identity Links）指向其他数据源用于识别同一个真实世界的物体或抽象概念的 URI 别名，通过身份/标识链接能够从其他数据源获取实体的进一步说明。

（3）词表链接（Vocabulary Links）是从数据指向表示数据的词表术语定义，也可以从这些定义指向其他词表的相关术语的定义。

图书馆可以从日期、地理信息、人物、机构等方面与网络上其他类型的数据资源建立关联。目前，不同数据集间的链接关系和关联关系的发现主要有三种方法：手动方法、半自动的方法和自动方法。手动方法主要是通过关联数据云查找相应的数据集，发现和得到相应关联的 URI，或者通过 Sindice、Uriqr 等查询相关资源的 URI；半自动的方法主要借助信息分析技术（如内容分析技术、路径分析技术等），只需少量的反馈性参与即可，目前基于半自动方式关联发现框架是主流；自动方法主要依赖于机器分析和学习，根据元数据和信息资源本身的特性，发现两个数据集之间的潜在关联关系，主要采用基于特定的命名模式的算法、基于属性的关联算法和基于统计的方法。目前，实现这种链接关系和关联关系发现的工具主要有：Silk、RKB-CRS、LD-mapper、KnoFuss、ODD-linker 等。

3.4　关联数据的整合及索引技术

关联数据整合可以分多种层次，本文从两个层面进行整合方案的研究：基于元数据整合和基于知识组织体系（Knowledge Organization System，KOS）整合；同时借助“基于 RDF 图模式建立语义索引”技术，解决快速获取、交互以及导航等方面的场景应用。

（1）基于元数据的关联数据整合。主要是对当前异构复杂的数据资源进行管理与利用，最终实现结构化、计算化和知识化的目标。针对博物馆、档案馆与图书馆为代表的多种资源关联数据的整合，郑燃等[31]提出为了实现数字资源的互操作与数据共享，将各种资源数据集以关联数据形式进行发布，通过中间的元数据项建立各个关联数据的映射转换，最终达到多种关联数据的整合。

（2）基于知识组织体系的关联数据整合。知识组织体系是对知识组织的各类规范和方法的统称，是获取、利用知识的重要手段[32]，将术语、概念及关系进行结构化存储，形成语义网实现的知识基础。而关联数据是语义网实现的重要技术手段，对已经发布的异地关联数据资源进行有效整合，是最终实现语义网的一个关键突破，并为 Web3.0 语义检索的应用奠定良好的数据基础。基于知识组织体系的关联数据整合主要过程如下：首先对分散的关联数据进行获取、抽取与噪音处理；识别多个数据源的知识实体，利用知识组织体系进行相关或者相同实体（概念）的归一化合并；对多个数据源的关联数据进行关键知识链的提取；最后将各个知识实体和抽取的知识链以 RDF 标准描述框架进行结构化保存，其中保存方式可以是临时性缓存保存，也可以借助语义存储工具进行永久性的物理设备保存，从而提供全面、规范的知识语义服务。

（3）RDF 语义索引技术。RDF 数据存储与管理对于提供语义服务至关重要，从粒度划分上有多个层次，如全局 RDF 图、RDF 文件、命名图（可以看作全局 RDF 图的一个子图）以及 RDF 三元组等多种划分方法。YARS2[33]作为爱尔兰国立大学研发的分布式 RDF 数据管理存储工具，其语义索引机制是将 RDF 三元组数据中的 Subject（S）、Predicate（P）、Object（O）和 Context（C）创建 6 个不同的索引，并把这些索引通过 Hash 函数映射到各个存储上，实现快速定位查询。Virtuoso[34]是一款支持多协议的数据库服务，主要用来整合存储 RDF 语义描述知识，以基于图的模型以及两种索引模式即 G，S，P，O 和 O，G，P，S，实现数据存储与索引，支持 ODBC 和 JDBC 两种访问获取方式提供对外服务接口，将 RDF 三元组数据中的 Graph（G）、Subject（S）、Predicate（P）、Object（O）创建 6 个不同的索引，同时支持语义推理。

3.5　关联数据的展示技术

目前，虽然关键词检索仍然是 Web 检索的主流，但是它以文档为单元的展示方式已经不能满足用户的需求。随着 RDF 及语义标识技术的逐步成熟，语义检索势必成为未来的检索主流。RDF 数据在展示技术方面具有其自身的特征，不能以传统的方法来展示信息，需要采用新的方式来呈现 RDF 数据之间的关系。

VisualDataWeb[35]组织针对 RDF 数据的可视化进行探索与发现，从语义关系的发现、语义对象的发展趋势、语义对象的分面浏览以及语义对象的分层揭示等几种展示方式，为用户提供了一个揭示关联数据的知识单元与潜在的知识关系的算法与工具软件平台。其中 RelFinder、SemLens、gFacet 以及 tFacet 四个关联数据的展示技术平台，采用 ORVI[36]（Object Mapping-Relationship search-Visualization-Interactive Exploration）作为关联数据可视化的框架，如图 1 所示：

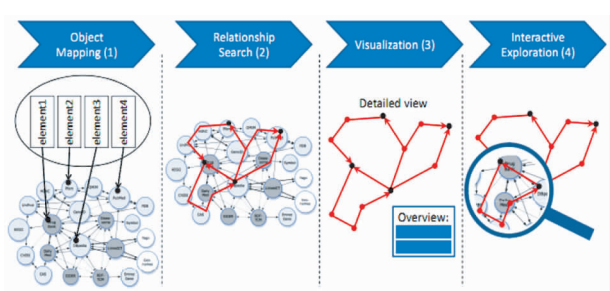

图 1　ORVI 语义可视化知识发现框架[36]

图 1 展示数据对象是以 RDF 数据为基础，以标准 SPARQL 语言作为数据获取的主要方法，同时从多个角度考虑可视化界面的设计，让用户真正参与到可视化的过程中，从而实现将关联数据的发现过程与发现结果进行合理、直观的展示。SemDis[37]项目研究内容之一是从关联数据的路径发现与揭示关系，主要从关系路径的重要度排序方面进行展示，可以在一定程度上清晰地揭示出关系链的强弱关系。

4　结　语

关联数据的优势在于可以促进结构化信息的协作式产生和重复性使用,这是一种使图书馆、档案馆、博物馆数据、政府数据和研究数据等相互结合的有效方式,图书馆可以通过统一、标准、去中心化的关联数据整合机制,将资源整合到更大的信息空间中,将原生资源和增值资源嵌入用户的信息环境,帮助用户发现通过其他方式无法找到的相关资源。不过,基于关联数据的信息检索速度也受关联数据源的稳定性、SPARQL 查询效率等的影响。另外,目前高质量的关联数据源比较少且数据类型较为单一,图书馆开放的也多为书目级别的关联数据,而对于用户更有用的数据是文章、引文、馆藏信息等数据。值得欣喜的是,Nature 出版集团[38]已经将从 1869 年至今的超过 45 万篇文章的元数据开放为关联数据,Open Citations 项目[39]也在积极探索与出版商合作,获取并发布参考文献的关联数据。随着出版领域、图书馆领域等更多高质量、多类型的关联数据发布以及关联数据查询技术的不断优化,在此之上构建的关联数据检索应用服务可以更加有效地丰富用户体验。

参考文献:

[1] Berners-Lee T. Linked Data[EB/OL]. [2012 - 09 - 20]. http://www.w3.org/DesignIssues/LinkedData.htm.

[2] Baker T, Bermès E, Coyle K, et al. Library Linked Data Incubator Group Final Report[R/OL]. [2012 - 03 - 20]. http://www.w3.org/2005/Incubator/lld/XGR - lld - 20111025/.

[3] Dataset Descriptions[EB/OL]. [2012 - 05 - 10]. http://id.loc.gov/descriptions/.

[4] A Bibliographic Framework for the Digital Age[EB/OL]. [2012 - 05 - 10]. http://www.loc.gov/marc/transition/news/framework - 103111.html.

[5] OCLC Releases FAST (Faceted Application of Subject Terminology) as Linked Data[EB/OL]. [2012 - 05 - 10]. http://www.oclc.org/news/releases/2011/201171.htm.

[6] Update on Kuali OLE: Our First Year[EB/OL]. [2012 - 05 - 10]. http://www.slideshare.net/sarnoa/kuali - update - v4 - mw.

[7] Linked Open BNB[EB/OL]. [2012 - 05 - 10]. http://www.bl.uk/bibliographic.

[8] Linked Open Data Pilot[EB/OL]. [2012 - 05 - 20]. http://pro.europeana.eu.

[9] 王军,程煜华. 基于传统知识组织资源的本体自动构建[J]. 情报学报,2009,28(5):651 - 657. (Wang Jun, Cheng Yihua. An Automatic Approach to Ontology Building by Integrating Traditional Knowledge Organization Resources[J]. *Journal of the China Society for Scientific and Technical Information*, 2009, 28(5): 651 - 657.)

[10] DBpedia[EB/OL]. [2012 - 09 - 20]. http://dbpedia.org/About.

[11] YAGO[EB/OL]. [2012 - 09 - 20]. http://en.wikipedia.org/wiki/YAGO_(database).

[12] UMBEL [EB/OL]. [2012 - 09 - 20]. https://github.com/structureddynamics/umbel.

[13] Capita Build Linked Data Libraries[EB/OL]. [2012 - 05 - 20]. http://consulting.talis.com/case-study.

[14] Wilson K. Introducing the Next Generation of Library Management Systems[J]. *Serials Review*, 2012, 38(2): 324 - 349.

[15] Schomburg S, Krabo U, Harper C, et al. Linked Data and Ex Libris Products[EB/OL]. [2012 - 03 - 20]. http://igelu.org/conferences/haifa - 2011/archive-of-presentations.

[16] Lodopac: Simple Linked Open Data OPAC[EB/OL]. [2012 - 05 - 20]. http://www.aurochs.org/lodopac/lodopac.php.

[17] GESIS: Integration of Research Data and Literature in the Social Sciences[EB/OL]. [2012 - 03 - 20]. http://www.gesis.org/.

[18] U. S. Government Dataset Search Opens Data.gov to Scientists [EB/OL]. [2012 - 05 - 20]. https://www.data.gov/communities/node/116/view/215.

[19] RSC Publishing and Southampton Drive the Chemical Semantic Web[EB/OL]. [2012 - 05 - 20]. http://blogs.rsc.org/technical/2011/05/16/.

[20] Pfeffer M, Eckert k. Publishing and Consuming Library Loan Information as Linked Open Data[EB/OL]. [2012 - 05 - 20]. http://jakoblog.de/tag/library/.

[21] Library Impact Data Project[EB/OL]. [2012 - 05 - 20]. http://library.hud.ac.uk/blogs/projects/lidp/.

[22] Pattern D. Free Book Usage Data from the University of Huddersfield[EB/OL]. [2012 - 05 - 12]. http://www.daveyp.com/blog/archives/528.

[23] LibraryCloud [EB/OL]. [2012 - 05 - 20]. http://www.librarycloud.org.

[24] VIVO [EB/OL]. [2012 - 05 - 20]. http://vivoweb.org/.

[25] RKB Explorer[EB/OL]. [2012 - 05 - 20]. http://www.rkbexplorer.com.

[26] Archives Hub Linking Lives[EB/OL]. [2012 - 05 - 20]. http://archiveshub.ac.uk.

[27] Heath T, Bizer C. Linked Data: Evolving the Web into a Global Data Space [M]. 1st Edition. Morgan & Claypool Publishers, 2011.

[28] 肖竹军. 基于 SPARQL 的 RDF 数据节点间关系路径检索[J]. *微型机与应用*,2011,30(9):50 - 53. (Xiao Zhujun. Relationship Path Search Between RDF Data Nodes Based on SPARQL [J]. *Microcomputer & Its Applications*,2011,30(9):50 - 53).

[29] Haase P, Mathäβ T, Ziller M. An Evaluation of Approaches to Federated Query Processing over Linked Data[C]. In:*Proceedings of I-SEMANTICS*,2010.

[30] SIREn[EB/OL]. [2012 - 05 - 20]. http://siren. sindice. com/.

[31] 郑燃,唐义,戴艳清. 基于关联数据的图书馆、档案馆和博物馆数字资源整合研究[J]. *图书与情报*,2012(1):71 - 76. (Zheng Ran,Tang Yi,Dai Yanqing. Digital Resources Convergence of Libraries, Archives and Museums Based on Linked Data Applications [J]. *Library and Information*,2012(1):71 - 76.)

[32] 贺德方. 国内外知识组织体系的研究进展及应对策略[J]. *情报学报*, 2010,29(6):963 - 972. (He Defang. Research Advances in Knowledge Organization Systems at Home and Abroad and China's Coping Strategies[J]. *Journal of the China Society for Scientific and Technical Information*, 2010,29(6): 963 - 972.)

[33] Harth A, Umbrich J, Hogan A, et al. YARS2:A Federated Repository for Querying Graph Structured Data from the Web[C]. In: *Proceedings of the* 6*th International Semantic Web Conference*, 2*nd Asian Semantic Web Conference*. 2007: 211 - 224.

[34] Erling O, Mikhailov I. RDF Support in the Virtuoso DBMS[C]. In: *Proceedings of the* 1*st Conference on Social Semantic Web* (*CSSW*). 2007:59 - 68.

[35] Studies in Computational Intelligence[EB/OL]. [2012 - 05 - 20]. http://www. visualdataweb. org/.

[36] Heim P, Lohmann S, Stegemann T. Interactive Relationship Discovery via the Semantic Web[C]. In: *Proceedings of the* 7*th International Conference on the Semantic Web*: *Research and Applications*. 2010:303 - 317.

[37] Semantic Discovery: Discovering Complex Relationships in Semantic Web[EB/OL]. [2012 - 05 - 20]. http://lsdis. cs. uga. edu/projects/semdis/.

[38] Nature Publishing Group Releases Linked Data Platform[EB/OL]. [2012 - 05 - 20]. http://www. nature. com/press_releases/linkeddata. html.

[39] Open Citations Project[EB/OL]. [2012 - 05 - 20]. http://opencitations. net/.

(作者 E - mail:huangyw@ mail. las. ac. cn)

IBM 收购大数据公司 StoredIQ

IBM 近日宣布,将收购大数据公司 StoredIQ,这是一家总部设在美国德克萨斯州奥斯汀市的一家私人信息管理公司。StoredIQ 专门分析和管理非结构化的企业数据。IBM 试图将其软件纳入 IBM 整体的大数据战略,其大数据战略特别注重治理(诉讼和监管合规)、安全数据处理和降低存储成本等方面。StoredIQ 将成为 IBM 的信息生命周期管理(Information Lifecycle Governance)套件的一部分。

IBM 希望帮助企业找出短期数据和长期数据中哪些数据是重要的,这样很容易找到有价值的数据(并且能够更容易消除不需要的数据)。除此之外,这也将有助于降低存储这些信息所有必要的硬件成本。

StoredIQ 之所以会引起 IBM 的兴趣,是因为:StoredIQ 的软件能够为不同的和分散的电子邮件以及文件共享和协作网站提供可扩展的分析和治理。其中包括发现、分析、监控、保留、收集、重复项删除和处理数据的能力。此外,StoredIQ 可以快速分析大量非结构化数据,并自动处理符合监管要求的文件和电子邮件。

StoredIQ 到目前为止大约拥有 120 家客户,涵盖金融服务、医疗保健、政府和制造等行业。此项交易预计于 2013 年第一季度完成。

(编译自:http://www. zdnet. com/ibm - to - acquire - storediq - in - big - data - play - 7000009028/)

(本刊讯)