基于 Sesame 及 Rdfizer 扩展工具的关联数据应用平台*

张永娟 陈涛 张珅

[摘 要] 采用 MetaStudio 和 DataScraper 对网络源非结构化数据按照需要进行自动抽取和 XML 结构化，并自主开发 Rdfizer 软件包，将 XML 数据转换为 RDF 数据，进而基于自行扩展的 Sesame 框架构建关联数据发布平台，实现关联数据的存储、发布、查询、整合和推理等功能。

[关键词] Sesame 框架 RDF 关联数据 Rdfizer 信息整合
[分类号] G202
DOI:10.7536/j.issn.0252-3116.2013.16.025

1 引 言

关联数据(linked data)是语义网上发布、共享数据的一种方式，也是语义网的最佳实践[1]，随着关联数据的不断增多，关联数据应用正成为研究的重点，以强调关联数据应用为核心的知识组织体系模式也逐渐成为图书馆领域研究和应用的热点。

国外研究主要集中在应用关联数据实现知识发现、资源交换与复用以及数据融合等方面，已开发出多个较成熟的知识组织模式和成功案例。而国内研究主要围绕关联数据格式、发布方式以及国外成功案例的研究，近年也有发布的数据出现，如上海图书馆已把上海市中心图书馆名录发布为关联数据[2]；黄华军等[3]根据中文叙词表本体的需求和特点，遵循关联数据创建四原则，实现了 OTCSS 的 Linked Data 服务模块，可将中文叙词表本体发布为关联数据，并提供叙词款目信息的 9 种共享格式(均采用 RDF 标准)的下载。

本研究以《生命科学研究快报》[4](Biological Science Information Express，BSIExpress)的非结构化网络数据为研究对象，对 Sesame 框架[5]进行扩展，并开发出 Rdfizer 软件包，用来对 XML 数据进行定制解析，通过 Rdfizer 的解析，将结构化的 XML 数据转换为标准的 Triple 或 Quad 数据，利用自行扩展的关联数据系统构建 BSIExpress 关联数据应用平台，实现了关联数据的转换、发布及整合。

2 系统构建

本研究构建的 BSIExpress 关联数据应用平台基于 Sesame 框架，通过扩展使其支持关联数据的存储、转换、发布及查询。系统框架如图 1 所示：

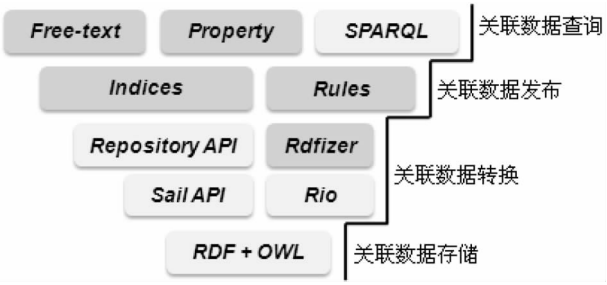

图 1 关联数据系统框架设计

底层系统为关联数据存储层，用来存储 RDF 数据以及本体 OWL。Sesame 支持数据的内存存储(memory store)、磁盘直接存储(native store)以及关系型数据库存储(RDBMS store)，同时该部分也可用来存储一些推理规则。

第二层是关联数据的核心部分，即关联数据转换层。Sail API 为底层系统 API(SPI)，用来进行 RDF 存储和推理。Rio 为 RDF 的 I/O 接口，用来解析和展现各种 RDF 的文件格式。Repository API 主要作为高一层的 API，主要提供对 RDF 数据的各种操作，如 RDF

*本文系上海市哲学社会科学规划课题青年项目“关联数据的复用与整合在图书馆知识服务体系中的应用模型构建”(项目编号：2011ETQ001)研究成果之一。
[作者简介] 张永娟，中国科学院上海生命科学信息中心馆员，硕士研究生，E-mail:zhangyj@sibs.ac.cn；陈涛，中国科学院上海生命科学信息中心工程师；张珅，中国科学院上海生命科学信息中心硕士研究生。
收稿日期:2013-07-03 修回日期:2013-07-12 本文起止页码:135-139 本文责任编辑:易飞

数据文件的导入、查询及数据的管理等。自主开发的 Rdfizer 软件包，用于抓取 XML 数据的定制解析，通过 Rdfizer 的解析，将结构化的 XML 数据转换为标准的 Triple 或者 Quad 数据，弥补了 Sesame 不具有的语义数据转换功能。

第三层是关联数据发布层。Sesame 提供 RDF 数据的管理而不提供关联数据的发布，本研究扩展出了关联数据发布这一层，可将仓库（Repository）中语义数据发布为一定格式的关联数据。关于数据推理，Sesame 只能实现 RDFs 级别上的自动推理功能，而本研究采用 OWLIM（OWLim - lite）实现 OWL 级别的推理。

第四层是关联数据查询部分。在 Sesame 支持的 SPARQL 查询的基础上，采用 Solr 搭建关联数据的全文检索索引库，扩展出全文（Free Text）检索，实现数据的快速查询，同时还实现了谓语属性（Property）检索，丰富了查询功能。

3　平台建设

BSIExpress 关联数据应用平台[6] 建设中需要解决几个主要问题：①如何将网络上的非结构化或者半结构化数据转换成结构化的数据，这也是语义网及关联数据所面临的首要问题；②如何将结构化的数据转换成 RDF/N3 - Triple 格式的语义数据，并进行存储和发布；③如何实现语义推理和 SPARQL 查询。解决方案则主要包括以下几个部分：数据提取、数据转换、关联数据的存储与模板扩展、关联数据的发布、基于 OWLIM 的推理及 SPARQL 查询等。

3.1　数据提取

本研究采用 GooSeeker 的 MetaStudio 和 DataScraper 将来自于网络的非结构化 BSIExpress 数据按照需要进行网页数据的自动抽取和结构化[7]。提取时，通过 MetaStudio 可以描述网页语义结构并自动生成网页数据抓取规则，进而利用 Web 页面信息提取工具 DataScraper 对页面信息进行连续提取，并生成 XML 格式的信息提取结果文件。提取出来的 XML 结构化数据片为：

```
<? xml version = "1.0" encoding = "UTF - 8" ? >
< extraction >
……
  < fullpath >
   <! [CDATA [ http://www. bioexpress. ac. cn/zhutis. asp? id = 11433 ]] >
  < /fullpath >
  < Detail >
  < item >
    < Info > 上传日期:2012 - 07 - 25 来源《生命科学研究快报》,共有 244 次浏览 < /Info >
    < Title > 世界首个基因治疗药物获欧洲药品管理局批准 < /Title >
    < Level0 > zhuti. asp? Y = 生物技术 < /Level0 >
    < Level1 > zhuti_1. asp? Y_1 = 基因生物技术 < /Level1 >
  < /item >
  < /Detail >
< /extraction >
```

3.2　数据转换

本研究通过自主开发的 Rdfizer 工具包将提取的 BSIExpress 的 XML 数据自动转换为 RDF 格式的语义数据，Rdfizer 工具包主要对 XML 文件结构进行解析，生成符合 BSIExpress 数据结构的语义数据，并通过 Sesame 的 Sail 接口将结构化的 RDF 数据存储到数据库中。转换时，系统支持单 XML 文件解析和 ZIP 压缩包（多 XML 文件）解析这两种数据转换方式。转换后的 N3 - Triple 数据为：

```
< http://www. bioexpress. ac. cn/zhutis. asp? id = 11433 > rdf:type bioexpress:Paper .
< http://www. bioexpress. ac. cn/zhutis. asp? id = 11433 > dc:title "世界首个基因治疗药物获欧洲药品管理局批准" .
< http://www. bioexpress. ac. cn/zhutis. asp? id = 11433 > dc:date "2012 - 07 - 25" .
< http://www. bioexpress. ac. cn/zhutis. asp? id = 11433 > dc:publisher "《生命科学研究快报》" .
< http://www. bioexpress. ac. cn/zhutis. asp? id = 11433 > bioexpress:hasCategory "生物技术" .
< http://www. bioexpress. ac. cn/zhutis. asp? id = 11433 > bioexpress:hasSubCategory "基因生物技术" .
```

3.3　关联数据的存储与模板扩展

本文采用关系型数据库（Mysql）存储 RDF 数据，然后发布到扩展的 Solr 索引库中，由此需要扩展 Sesame 的数据模板文件，实现 Solr 与 Sesame 的整合。该模板中，扩展了属性 rep:repositoryCore "{% Repository core% }"，该属性用来将仓库节点（Repository）关联到具体的 Solr core。对于不同的仓库，可以关联到不同的 Solr core，也可以关联到同一个 Solr core。不同的 Solr core，可以对数据进行隔离查询，单个仓库的查询及维护不影响其他仓

库的数据;当关联到同一个 Solr core 时,可以对不同的仓库数据进行统一查询,从而组成 Data Hub。可在 SYSTEM 仓库中查看到创建好的新仓库的信息。

3.4 关联数据的发布

关联数据的发布工具主要有 D2R、LMF(Linked Media Framework)、Virtuoso、Pubby 等[8]。关联数据是实现数据 Web 的关键技术,构建关联数据系统应符合以下几个原则:①使用 URI 作为任何事物的标识名称;②使用 HTTP URI,使任何人可以查找到该名称;③当有人访问名称时,以 RDF 形式提供有用的信息;④尽可能提供链接去指向其他 URI,以使人们发现更多的相关信息。

根据关联数据发布原则,本文将《快报》地址映射为三部分信息(页面):用于文献标识的 resource 页面、便于人工阅读的 page 页面和供机器阅读的 data 页面。如:《快报》原文地址为"http://www.bioexpress.ac.cn/zhutis.asp? id=11432",转为关联数据后,新系统的文献资源地址为/resource/article/011432,该地址用于标记系统中文献的唯一性;文献信息地址/page/article/011432,用于显示文献具体数据;RDF 数据地址/data/article/011432 则用于显示文献 RDF/XML 或者 N3 数据结构,该地址的查询需要通过语义浏览器(如 Tobular)查看。

发布关联数据时,可以根据特定的数据条件进行发布,如根据具体文献(特定的主语)发布关联数据,或根据某类文献(特定的谓语或宾语)发布关联数据。最终将语义数据发布到 Solr core 中,供查询使用。当切换发布地址时,可以重新发布关联数据。

3.5 基于 OWLIM 的推理

《快报》的期刊信息和文献本身的信息分开存储,发布成关联数据后,进行文献信息查看时,需要将期刊本身的信息推理到文献信息中,比如,期刊为 2012 年 1 月第 1 期(总 203 期),则该期中的所有文献都隐含该期刊的年、月、期等信息。要实现这样的信息过渡,需要用到推理库。

本研究中使用 OWLIM 实现多种类型的 OWL 的推理,它可以作为 Sesame 框架扩展的存储引擎,能够进行快速推理。根据 OWLIM 的 Rule 文件推理结构,将规则定义的 Rules 部分、推理前提条件 Premise 部分以及推理结果 Corollary 部分按需要写成 N-Triple 格式:

Id:bio_volume_paper

volume <rdf:type> <bioexpress:Volume>

volume <bioexpress:hasPaper> paper

volume <dcterms:issued> year

volume <swrc:volume> issue

volume <swrc:month> month

volume <swrc:number> number

paper <dcterms:issued> year

paper <swrc:volume> issue

paper <swrc:month> month

paper <swrc:number> number

推理前提条件中定义:期刊 volume 的类型为 bioexpress:Volume,期刊含有文献 paper(bioexpress:hasPaper);期刊 volume 有 4 个数据属性:dcterms:issued(发表年);swrc:volume(总期数);swrc:month(发表月份);swrc:number(上半月|下半月)。根据期刊信息可以推断出文献具有同期刊一样的数据属性。图 2、图 3 分别显示了资源(http://202.127.22.42:8080/sesame/repositories/test/page/article/011433)是否应用推理所得的查询结果;同时用户可以浏览文献信息的 RDF(N-Triples)结构(见图 4)。

☐ Include inferred statements

Property	Value
owl:sameAs	<http://www.bioexpress.ac.cn/zhutis.asp?id=11433>
dc:publisher	"生命科学研究快报"
dc:title	"世界首个基因治疗药物获欧洲药品管理局批准"
rdf:type	bioexpress:Paper
bioexpress:hasCategory	"生物技术"
dc:date	"2012-07-25"
bioexpress:hasSubCategory	"基因生物技术"

图 2 未选中推理选框的查询结果

☑ Include inferred statements

Property	Value
owl:sameAs	<http://www.bioexpress.ac.cn/zhutis.asp?id=11433>
dc:publisher	"生命科学研究快报"
dc:title	"世界首个基因治疗药物获欧洲药品管理局批准"
swrc:month	"7"
swrc:number	"2"
rdf:type	bioexpress:Paper
rdf:type	rdfs:Resource
bioexpress:hasCategory	"生物技术"
dcterms:issued	"2012"
dc:date	"2012-07-25"
bioexpress:hasSubCategory	"基因生物技术"
swrc:volume	"235"

图 3 选中推理选框的查询结果

3.6 SPARQL 查询

系统平台提供多种查询方式:①支持单词和多词的全文检索(free text),如可用"基因 趋势"查询所有含有"基因"和"趋势"的文章;②支持按照 RDF 数据的

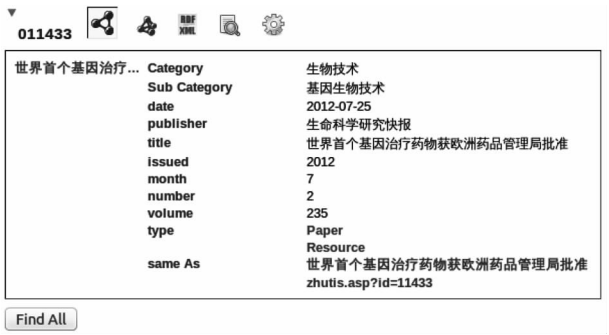

图 4　文献信息的 RDF(N－Triples)结构浏览

Property 进行查询,如可使用“swrc:volume”查询某一期的所有文章;③支持 SPARQL 查询以及利用 Service 进行的联合查询。

支持 SPARQL 查询以及利用 Service 进行的联合(Federated)查询。如使用如下 SPARQL 语句可以实现本地数据与 Pubmed 数据的联合查询:

```
PREFIXdcterms: <http://purl.org/dc/terms/>
PREFIX bioexpress: <http://202.127.22.42:8080/sesame/ontology/bioexpress.htm#>
PREFIX dc: <http://purl.org/dc/elements/1.1/>
PREFIX swrc: <http://swrc.ontoware.org/ontology#>
SELECT ? title
WHERE{
{
    ? sswrc:volume 232;
bioexpress:hasPaper ? paper .
    ? paperdc:title ? title .
} UNION {
    SERVICE <http://pubmed.bio2rdf.org/sparql> {
        ? articledcterms:title ? title.
    }
}
} LIMIT18
```

查询结果将包括本地数据库中 232 期的 14 篇文章和 Pubmed 中的 4 篇文章,见图 5。

4　结　语

本研究构建了 BSIExpress 关联数据应用平台,实现了网络源非结构化及半结构化的数据自动抽取和 XML 结构化,同时可将结构化的 XML 数据转换为标准的 RDF/N－Triples 格式,并进行分布式存储,进而按照关联数据发布的“四项基本原则”实现了关联数据的发布,平台还构建了“《快报》本体”,实现了基于主体的主题分

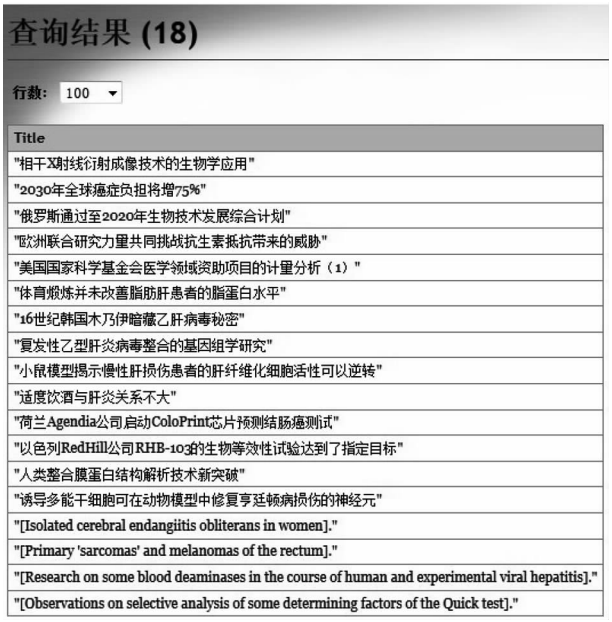

图 5　与 Pubmed 等外部关联数据的整合

类,同时也尝试了基于 OWLIM 的语义推理,并支持 SPARQL 查询及多元数据联合查询。

关联数据的发布是基础,如何更好地应用关联数据才是最终目的,实现关联数据的复用、整合,并进一步实现语义推理和知识挖掘是下一步亟需解决的问题。本研究对网络源非结构化数据发布为关联数据(WEB2RDF)做了较深入探索,构建了具有关联数据转换、发布、查询功能的应用平台,并对关联数据在语义层面的操作应用方面做了尝试,打通了一条通向关联数据应用的小路,对关联数据在知识组织和资源发现等领域的应用有一定的推动作用。

参考文献:

[1] Berners-Lee T. Linked data [EB/OL]. [2013－06－30]. http://www.w3.org/DesignIssues/LinkedData.html.

[2] 上海市中心图书馆名录 [EB/OL]. [2013－06－30]. http://data.libnet.sh.cn:8080/.

[3] 黄华军,曾新红,林伟明. OTCSS 关联数据服务的研究与实现[J]. 现代图书情报技术,2012(7/8):40－47.

[4] 生命科学研究快报 [EB/OL]. [2013－06－30]. http://www.bioexpress.ac.cn/.

[5] Pedrinaci C, Bernaras A, Smithers T, et al. A framework for ontology reuse and persistence integrating UML and sesame [C]. //Proceedings of the 10th Conference of the Spanish Association for Artificial Intelligence, San Sebastian, 2003.

[6] 生命科学研究快报(BSIExpress)关联数据应用平台(test) [EB/OL]. [2013－06－30]. http://202.127.22.42:8080/sesame/repositories/test/bioexpress.

[7] GooSeeker [EB/OL]. [2011－10－08]. http://www.gooseeker.com/.

[8] 夏翠娟,刘炜,赵亮,等. 关联数据发布技术及其实现——以 Drupal 为例[J]. 中国图书馆学报,2012,38(1):49-57.

A Linked Data Application Platform Based on the Sesame and Customized-Rdfizer

Zhang Yongjuan Chen Tao Zhang Shen
Shanghai Institutes for Biological Sciences/Shanghai Information Center for Life Sciences,
Chinese Academy of Sciences, Shanghai 200031

[Abstract] The paper introduces how to convert and publish unconstructed data from web source to linked data. MetaStudio and DataScraper are used to capture raw data and convert them to XML format, which can be switched to RDF/XML conveniently. Customized-Rdfizer package is the most important component of our platform which completes the data conversion. The terminated RDF data can be stored in Sesame framework.

[Keywords] sesame framework RDF linked data Rdfizer resources integrity