研究论文

Web 数据到 RDF 数据的框架实现*

陈　涛　张永娟　陈　恒
(中国科学院上海生命科学信息中心　上海 200031)

摘要: 【目的】构建 Web 数据到 RDF 数据(W2R)转换框架，实现 Web 数据的 RDF 结构化。【方法】采用 W2R 词表构建转换框架的底层结构，并根据设计的系统本体和 Web 页面元素组成映射文件进行数据的 RDF 结构化，同时采用 Virtuoso 数据库进行数据存储。【结果】通过对映射文件的灵活配置，在不修改任何程序代码的基础上，实现 Web 数据的 RDF 结构化、不同数据源之间数据的整合以及 RDF 数据的 Named Graph 存储及推理。【局限】系统的本体结构以期刊和文献结构为主，尚不支持其他知识领域。此外，针对 RDF 数据的持久化存储，W2R 框架目前仅支持 Virtuoso 数据库。【结论】W2R 框架实现 Web 数据的 RDF 结构化，为语义网络和关联数据的应用提供标准化数据。

关键词: 本体　语义网络　数据采集　W2R
分类号: G202

1　引　言

互联网上存在着大量非结构化数据和采用不同标准的结构化数据，关联数据是一种简单的语义网实现技术，可以将多种数据开放并连接在一起，允许用户发现、关联、描述并再利用各种数据，它使互联网迈出了向语义网(Semantic Web)进化的重要一步。

近年来，越来越多的机构、组织及政府部门都对外开放其数据，并与其他机构发布的数据关联，实现跨数据库的数据交换，范围涉及多媒体、文献出版、生命科学、地理信息等。美国、英国、巴西、新西兰等国家也逐渐将政府信息(涵盖卫生、农业、税务、教育等方面)发布成可重用的关联数据。因此，“文档的网络(the Web of Document)”向“数据的网络(the Web of Data)”转变，已经是大势所趋[1]。

RDF(Resource Description Framework)为 Web 资源描述提供了一种通用框架，即通过“资源–属性–值”的三元组形式描述 Web 上的各种资源。它以一种机器可理解的方式被表示出来，提供了 Web 数据集成的元数据解决方案，可以很方便地进行数据交换。相对于其他数据形式，RDF 数据具有易控制、易扩展、易综合以及高包容性和可交换性等特点。通过 RDF 的帮助，Web 可以实现一系列应用，如可以更有效地发现资源，提供个性化服务，分级与过滤 Web 的内容，建立信任机制，实现智能浏览和语义 Web 等[2-4]。

目前，将关系型数据库中的数据转为 RDF 数据，即 RDB2RDF 的研究较多，主要有 D2R (http://d2rq.org/d2r-server)、R2RML (http://www.w3.org/TR/r2rml)。实现方法主要是通过映射文件(Mapping File)，将关系型数据库中的表和字段依据相互之间的关系映射成 RDF 三元组数据[5-6]。RDB2RDF 的映射方式主要适用于企业和机构进行内部数据的 RDF 转换，但是在制定映射文件时，必须获取数据库的访问权限，并对数据结构具有相当程度的了解。而对于 Web 数据转换，一

通讯作者：陈涛，ORCID: 0000-0002-6609-4914，E-mail: chentao01@sibs.ac.cn。
*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)的研究成果之一。

般不太可能获取其数据结构以及访问权限，因此 RDB2RDF 不太适合 Web 数据的 RDF 结构化。

当然，也存在一些将 Web 数据转为 RDF 数据的方法，如: Apache Marmotta LDClient (http://marmotta.apache.org)和 Apache Any23 (http://any23.apache.org)。这些方法主要以提供 API 为主，需要用户从逻辑层的角度自行转换 Web 数据，不仅要求使用者具有一定的编程能力，而且还需要熟悉语义网络及关联数据的相关技术，使用门槛较高。本文则从应用层的角度出发，试图对 Web 数据的转换机制进行封装，使用者只需通过简单的模板设置就可以实现对 Web 数据的 RDF 结构化。

2 系统框架设计

2.1 W2R 转换框架

目前流行的关联数据和语义网络开源框架主要有 Jena (http://jena.apache.org)和 Sesame (http://www.penrdf.org)，本文提出的 W2R 转换框架后台采用 Jena 框架，可以集成到 Java 应用中，作为数据转换接口实现 Web 数据的 RDF 结构化[7-9]。转换框架设计如图 1 所示:

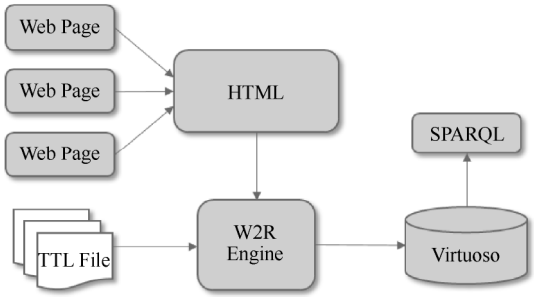

图 1 W2R 转换框架

具体转换流程如下:

(1) 通过 Get 或者 Post 方法从网络页面(Web Page)中获取 HTML 源文件。

(2) 制定系统本体，根据 Web 页面元素从本体中抽取相应属性构成抓取模板，即 TTL 文件。

(3) 数据转换方面采用自行研发的 W2R 转换引擎。该引擎根据 TTL 映射模板，从 HTML 源文件中抽取所需字段信息，并将信息转换为 RDF 三元组格式进行持久化存储。

(4) 数据的持久化存储采用 Triple Store 数据库，这里选用 Virtuoso 数据库(http://www.openlinksw.com/)，该数据库不仅支持 RDF 数据管理，同时还针对 RDF 数据提供 SPARQL 访问节点。

2.2 本体知识属性的扩展

类似于 RDB2RDF，Web 数据的 RDF 结构化同样需要映射文件，W2R 框架的映射文件主要由系统结构本体中的相关属性组成。本文的系统结构本体如图 2 所示，主要涉及 4 个类: 期刊(Journal)、文献(Article)、作者(Author)以及组织机构(Organization)，当然也可以扩展到其他类，如基金类、主题词类等。

项目设计本体时，主要是在采用已有的开放词表的基础上，扩展一些私有词表，这里采用的开放本体词表主要有 dct (dcterms)、foaf、wgs84_post、bibo 等。而 cba 词表则为扩展的私有词表，这些词表的属性暂称为知识属性，用于描述不同对象间的知识结构。系统结构本体在设计时，建议尽可能采用开放本体词表，这样可以方便不同系统之间数据的共享、关联与复用。对于私有词表，可以通过 owl:sameAs 建立与其他开放

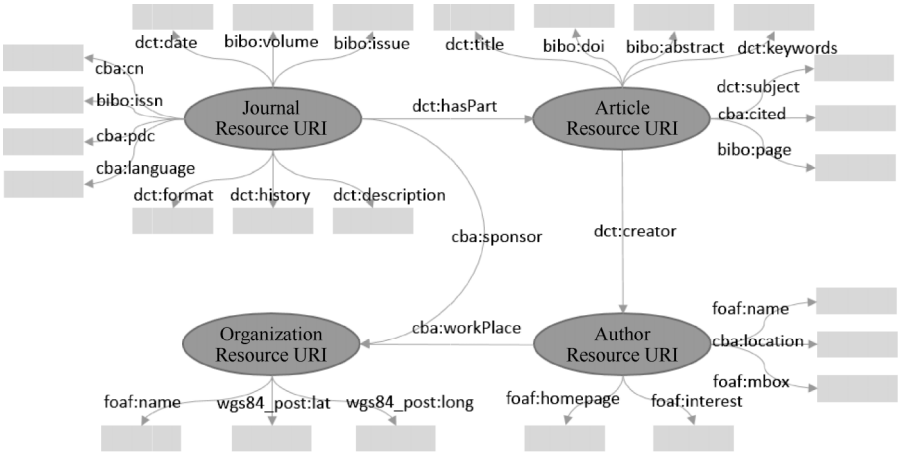

图 2 系统本体设计

本体词表属性的连接关系，实现与其他系统之间数据的关联。

2.3　系统结构属性的设计

系统的结构本体除了定义的知识属性外，还包括一些用于系统框架的底层结构，即结构属性。这些结构属性，主要是以 W2R 词表为主的底层词表。

(1) w2r:graph，目标 Graph(图)，用来进行 Named Graph 的数据存储。从页面转换后的数据将被存储到定义的 Graph 中，每个模板至少有一个 Graph。

例如: [ ] w2r:graph bibo:Journal，表示将抓取的数据存储到 Journal Graph 中。

(2) w2r:target，指向页面中目标采集区域。

(3) w2r:mode，抓取模式，与 w2r:target 结合使用，表明在哪个目标域中进行模式抓取，这里的模式分为 solo(单例)和 repeat(循环)。其中，solo 为单例抓取，主要用于单个文献、单个期刊信息的抓取; repeat 为循环采集，主要用于某个期刊中所有文献列表的抓取。

(4) w2r:merge，用于信息合并、增量采集，可以定义多个合并因子。

例如: [ ] w2r:merge “%%ISSN%%,%%TITLE%%”，表示根据期刊“ISSN 号”和“标题”这两个因子进行不同数据源之间文献信息的合并。

(5) w2r:refer，指定引用对象，获取对象的主语。

例如: [ ] w2r:refer “%%dcterms:title%%”，表示根据标题获取资源主语。与 w2r:merge 区别在于，w2r:refer 仅获取资源主语，而 w2r:merge 获取主语后，还与资源中的信息进行合并操作。

(6) w2r:uniPattern，用于定义主语生成规则。当该属性为空时，则为空节点 BNode，RDF 数据在实际使用过程中，尽量避免使用空节点，空节点只能内部识别，不能用于数据之间的关联。

例如: [ ] w2r:uniPattern “%%uuid%%”，表示生成的主语采用 UUID 唯一码进行标识。

(7) w2r:rules，用于定义推理规则，与 w2r:graph 结合使用，表明在哪个 Graph 中应用推理规则。

例如: [ ] w2r:rules “%%rule2%%,%%rule6%%”; w2r:graph graph:Journal，表示在期刊 Graph 中采用规则 2 和规则 6 共同推理。

2.4　**W2R** 转换引擎

系统本体设计好后，可以通过 W2R 转换引擎实现 HTML 数据的 RDF 结构化。W2R 引擎接口如图 3 所示:

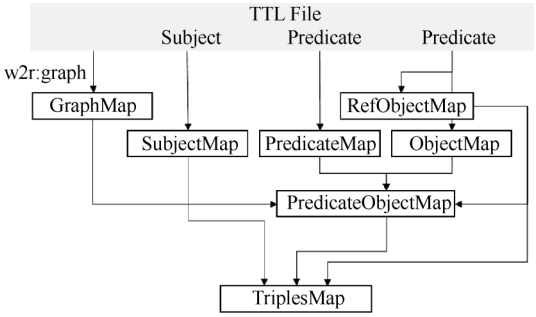

图 3　W2R 转换引擎

具体操作如下:

(1) GraphMap 的实现。每个模板可以有一个或多个 Graph，每个主语节点只能有一个 Graph。转换时，根据模板中不同主语所拥有的 w2r:graph 属性，进行 Graph 的匹配。如果该主语节点没有 w2r:graph 属性，则需要寻找上一层父节点的 Graph，即该主语节点是其他主语节点中对象属性的宾语。

(2) SubjectMap 的实现。该操作主要完成所有主语的实例化。实例化根据该主语节点中的 w2r:uniPattern 属性生成。如果某个节点中没有 w2r:uniPattern 属性，则该主语为空节点。

(3) PredicateMap 和 ObjectMap 的实现。谓语属性分为两类，数据属性和对象属性。当判断到该谓语属性是数据属性时，从模板中获取该属性对应的宾语格式，并从 HTML 中进行宾语的实例化; 当该谓语属性是对象属性时，根据引用关系从 SubjectMap 中选取实例化后的主语。PredicateMap 和实例化后的 ObjectMap 组成 PredicateObjectMap 对，并与 SubjectMap 实例化的主语一起构成三元组 Triples。

(4) TriplesMap 的实现。在步骤(3)中，已经完成 Triples 的转换，本步骤则完成数据的持久化存储。当采集完一篇页面数据后，可以将转换后的所有 RDF 数据放置到内存模型 MemoryModel 中，待所有的数据都采集结束，统一写入数据库中，这样可以减少数据库的频繁读写操作。

3　采集模板制定

系统结构设计和实现后，可以通过对本体属性元素的随意组合，定义所需目标页面元素的抓取模板。

页面元素的分析，主要通过 Jsoup (http://jsoup.org/)实现，Jsoup 是一款 Java 的 HTML 解析器，可直接解析某个 URL 地址、HTML 文本内容。它提供了一套非常省力的 API，可通过 DOM、CSS 以及类似于 jQuery 的操作方法读取和操作数据。对于不同的数据源，可以定义不同的采集模板，这里的模板主要采用 TTL 数据格式，以期刊采集为例，模板由以下块信息段组成:

(1) @prefix 段定义了期刊采集中使用的词表前缀。

```
@prefix dcterms:    <http://purl.org/dc/terms/> .
@prefix w2r:        <http://www.cba.ac.cn/w2r/ontology/1.0/> .
@prefix :           <http://www.cba.ac.cn/ontology/1.0/> .
@prefix host:       <http://www.cba.ac.cn/> .
@prefix graph:      <http://www.cba.ac.cn/graph/> .
@prefix rdfs:       <http://www.w3.org/2000/01/rdf-schema#> .
@prefix rdf:        <http://www.w3.org/1999/02/22-rdf-syntax-ns#>.
@prefix foaf:       <http://xmlns.com/foaf/0.1/> .
@prefix bibo:       <http://purl.org/ontology/bibo/> .
@prefix skos:       <http://www.w3.org/2004/02/skos/core#> .
```

(2) host:template 段定义抓取模板自身的信息段，该段信息不写入数据库。

```
host:template rdf:type :Template;
    rdfs:comment “期刊模板”;
    w2r:target “div.left”;
    w2r:mode “solo”.
```

该段指明文件类型为“期刊模板”，除此以外，系统还支持“文献模板”。模板的制定和选择主要根据实际的 Web 页面结构进行设计，如想获取专家学者的个人信息，可以指定“专家模板”进行采集。该段还指明了抓取的目标域，即从 HTML 的 div.left 节点抓取信息。抓取模式为“单例 solo”。w2r:target 和 w2r:mode 为 W2R 框架的底层结构属性。

(3) host:journal 段定义期刊信息段，产生的 RDF 数据将存入目标 Graph 中，即属性 w2r:graph 所对应的宾语值。

```
host:journal rdf:type bibo:Journal;
    :date “yyyy-MM-dd”;
    :image “div.pic > a > img”;
    dcterms:title “div#tdInfo > p:eq(1) > a:eq(0)”;
    :sponsor host:Organization;
    :period “div#tdInfo > p:eq(1) > a:eq(2)”;
    :location “div#tdInfo > p:eq(1) > a:eq(3)”;
    :language “div#tdInfo > p:eq(1) > a:eq(4)”;
    :size “div#tdInfo > p:eq(1) > a:eq(5)”;
    bibo:issn “div#tdInfo > p:eq(1) > a:eq(6)”;
    :cn “div#tdInfo > p:eq(1) > a:eq(7)”;
    :pdc “div#tdInfo > p:eq(1) > a:eq(8)”;
    skos:historyNote “div#tdInfo > p:eq(1) > a:eq(11)”;
    skos:note “div#tdInfo > p:eq(1) > a:eq(12)”;
    w2r:uniPattern “%%uuid%%”;
    w2r:graph graph:Journal.
```

该段指明需要抓取的期刊信息：期刊名(:title)、主办单位(:sponsor)、周期(:period)、语种(:language)、开本(:size)、ISSN(bibo:issn)、CN(:cn)、邮发代号(:pdc)等信息。其中，:sponsor 为对象属性，用于连接期刊和主办单位，w2r:uniPattern 和 w2r:graph 为 W2R 框架的底层结构属性。

(4) host:organization 段定义期刊主办单位信息段。

```
host:organization rdf:type foaf:Organization;
    foaf:name “div#tdInfo > p:eq(1) > a:eq(1)”.
```

该段简单指明了期刊主办单位的基本信息(主办单位)。需要注意的是本信息段没有单独制定数据需要存储的目标 Graph，即不含 w2r:graph 属性，因此主办单位信息将与期刊信息存储在同一 Graph 中。当然也可以为本信息段单独指定 w2r:graph 属性，如 graph:Organization，将主办单位信息存储到组织机构 Graph 中。

4 转换效果评估

本文提出的 W2R 转换框架可以根据需要构建多个基于 RDF 数据的应用系统。图 4 为构建的文献数据库期刊管理系统，其中的期刊为 Web 数据 RDF 格式化的显示结果，可以在“模板管理”菜单中制定不同的抓取模板以及数据的采集和整合规则。抓取后的期刊 RDF 三元组数据存储到 Virtuoso 数据库中，页面显示可以根据需要重新组合。以“癌症进展”为例，数据库中的 RDF 三元组如表 1 所示。

图 4 文献数据库期刊管理系统

对期刊的 Web 数据进行 RDF 结构化后，可以进行 RDF 三元组数据的发布。关联数据的发布方式有很多种[10-11]，如采用 Fuseki 发布 TDB 中的数据[12]；采用

表 1　Web 数据的 RDF 结构化三元组

Predicate	Object
rdf:type	Bibo:Journal
skos:historyNote	历史沿革：现用刊名：癌症进展 创刊时间: 2003
cba:language	中文
cba:period	周期：双月
cba:sponsor	_:-a1a6548:140c3d017ec:-7fbe
dcterms:title	癌症进展
dcterms:title	“Oncology Progress”@en
cba:date	2014-08-28
cba:cn	CN:11-4971/R
bibo:issn	1672-1535

Pubby (http://wifo5-03.informatik.uni-mannheim.de/pubby)构建关联数据的 SPARQL 端点；采用 RDFa 语义标识 HTML 数据等[13]。本文采用 Virtuoso 进行 RDF 三元组数据的存储，因此可以直接利用 Virtuoso 自带的 SPARQL 端点，无需进行其他配置。然而，RDF 数据的发布只是关联数据和语义网应用的第一步，如何将转换后的 RDF 数据与其他开放数据建立关联，是目前框架欠缺的部分。此外，关联数据的发现机制也正在研发之中，如通过关联机制，可以和 PubMed (http://pubmed.bio2rdf.org/sparql)进行文献关联；可以根据期刊的主办单位从 Dbpedia (http://dbpedia.org/sparql)中获取更多单位信息；也可以根据地域从 GeoNames (http://www.geonames.org/)中获取更多地域信息。

目前，系统的数据采集及 RDF 结构化只是在小范围内进行准确性验证，后期将进入大数据量的并发采集，并对采集和转化的效率进行量化评估，以便适用于更多的应用场景。

5　结　语

RDF 数据是语义网络和关联数据的数据基础，将 Web 数据转换为 RDF 数据，是进行数据整合及信息交互的前提。本文提出的 W2R 转换框架，具有一定的通用性，用户可以根据 Web 数据的实际需求，自行定义系统结构本体，并可制定多样的采集模板，对 Web 数据进行 RDF 的格式化、整合、推理等操作。尽管如此，除了正在研究的关联发现机制外，还需要在以下三个方面进行改进，以适应更多的应用场景:

(1) 目前框架仅支持 Virtuoso 数据库，后期将扩展更多数据库的存储;

(2) 完善本体结构，加入更多的初始模板，适应更多应用场合;

(3) 开发页面插件，使用可视化的方式进行页面元素的选择和 RDF 的转换。

参考文献：

[1] Linked Data: Evolving the Web into a Global Data Space [EB/OL]. [2014-05-17]. http://linkeddatabook.com/editions/1.0/.

[2] 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(5): 1-7. (Huang Yongwen. Research on Linked Data-driven Library Applications [J]. New Technology of Library and Information Service, 2010(5): 1-7.)

[3] 刘炜，夏翠娟，张春景. 大数据与关联数据：正在到来的数据技术革命[J]. 现代图书情报技术，2013(4): 2-9. (Liu Wei, Xia Cuijuan, Zhang Chunjing. Big Data and Linked Data: The Emerging Data Technology for the Future of Librarianship [J]. New Technology of Library and Information Service, 2013(4): 2-9.)

[4] 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(11): 1-9. (Shen Zhihong, Zhang Xiaolin. Linked Data and Its Applications: An Overview [J]. New Technology of Library and Information Service, 2010(11): 1-9.)

[5] R2RML: RDB to RDF Mapping Language [EB/OL]. [2014-07-27]. http://www.w3.org/TR/2012/REC-r2rml-20120927/.

[6] Unbehauen J, Stadler C, Auer S. Accessing Relational Data on the Web with SparqlMap [C]. In: Proceedings of the 2nd Joint International Conference, Nara, Japan. 2013: 65-80.

[7] Chen T, Zhang Y, Zhang S, et al. Building Semantic Information Search Platform with Extended Sesame Framework [C]. In: Proceeding of the 8th International Conference on Semantic Systems. New York, USA: ACM, 2012: 193-196.

[8] Antoniou G, van Harmelen F. A Semantic Web Primer [M]. The 2nd Edition. London: The MIT Press, 2008.

[9] 张永娟，陈涛，张珅. 基于 Sesame 及 Rdfizer 扩展工具的关联数据应用平台[J]. 图书情报工作, 2013, 57(16): 135-139. (Zhang Yongjuan, Chen Tao, Zhang Shen. A Linked Data Application Platform Based on the Sesame and Customized-Rdfizer [J]. Library and Information Service, 2013, 57(16): 135-139.)

[10] Bizer C. How to Publish Linked Data on the Web [EB/OL]. [2014-05-10]. http://wifo5-03.informatik.uni-mannheim.de/bizer/pub/LinkedDataTutorial/ .

[11] 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. (Xia Cuijuan, Liu Wei, Zhao Liang, et al. The Current Technologies and Tools for Linked Data: A Case of Drupal [J]. Journal of Library Science in China, 2012, 38(1): 49-57.)

[12] Fuseki: Serving RDF Data over HTTP [EB/OL]. [2014-06-08]. http://jena.apache.org/documentation/serving_data/index.html.

[13] RDFa 1.1 Primer-Second Edition [EB/OL]. [2014-04-13]. http://www.w3.org/TR/xhtml-rdfa-primer/.

作者贡献声明：

陈涛: 提出研究思路, 设计研究方案及框架, 论文起草及最终版本修订;

张永娟: 验证方案, 数据采集, 进行实验;

陈恒: 系统本体结构分析及构建。

收稿日期: 2014-08-06

收修改稿日期: 2014-09-20

Implementation of the Framework for Converting Web-data to RDF (W2R)

Chen Tao Zhang Yongjuan Chen Heng

(Shanghai Information Center for Life Sciences, Chinese Academy of Sciences, Shanghai 200031, China)

Abstract: [**Objective**] The article aims at building W2R framework for converting Web data to RDF format. [**Methods**] Build the bottom infrastructure of the framework with W2R vocabulary, and convert Web data to RDF format with mapping file which is consisted of system Ontology and Web page elements extracted in XPath syntax. Furthermore, use Virtuoso database as the persistent storage of RDF data. [**Results**] With the W2R framework, it is convenient for converting Web data to RDF format, merging data in different resources, storing them in named graphs and implementing simple inferences without changing any source code. [**Limitations**] The system Ontology is made up of public namespaces that describe the bibliographies currently. RDF data is only stored in Virtuoso database. [**Conclusions**] Through the W2R framework, this paper provides a new way of generating the standardized RDF data for semantic network and linked data applications.

Keywords: Ontology Semantic network Data acquisition Web data to RDF data