基于 R2RML 的医学主题词表 RDF 转换实现*

吴思竹 修晓蕾 李艳梅 钱 庆

（中国医学科学院医学信息研究所 北京 100020）

〔摘要〕 针对医学主题词表（MeSH）数据存储情况和特点，利用 R2RML 映射语言采用一表对一表、一属性对一属性、两表之间含外键的映射等 5 种映射模式实现医学主题词表的源描述框架转换，为其他词表的 RDB2RDF 转换提供借鉴。

〔关键词〕 RDB2RDF；R2RML 映射语言；医学主题词表；RDF

〔中图分类号〕R－056 〔文献标识码〕A 〔DOI〕10.3969/j.issn.1673－6036.2019.05.015

Conversion of Medical Subject Headings（MeSH）to RDF Based on R2RML *WU Sizhu, XIU Xiaolei, LI Yanmei, QIAN Qing, Institute of Medical Information, Chinese Academy of Medical Sciences, Beijing 100020, China*

〔**Abstract**〕 The paper targets at data storage and characteristics of Medical Subject Headings（MeSH）, and makes use of R2RML mapping language to actualize the Resource Description Framework（RPF）conversion of MeSH through 5 mapping modes, such as table－to－table, attribute－to－attribute, and foreign keys between tables, so as to provide references for the RDB2RDF conversion of other subject headings.

〔**Keywords**〕 RDB2RDF; R2RML mapping language; Medical Subject Headings（MeSH）; RDF

1 引言

2001 年 Bernerslee T 等人首次提出语义网概念[1]。随着语义网的兴起，关联数据得到越来越多的发展[2]，图书馆界已意识到关联数据和语义网络是公开集合数据更好的手段，而源描述框架（Resource Description Framework，RDF）是更适合于语义网的数据模型。为将关系数据库中的数据融入到语义网中需要将其转换成 RDF，这一转换过程被称为 RDB2RDF。

作为国际上最具代表性、使用最广泛的受控医学综合性叙词表——医学主题词表（Medical Subject Headings，MeSH）[3]，一些机构尝试从不同出发点、利用不同转换技术和语义模型进行 MeSH 的 RDF 转换。2004 年 Van Assem 等人利用简单知识组织系统（Simple，SKOS）的 RDF 模型，首次将 MeSH 转换成 RDF[4]，2006 年科学共享研究人员对 Van Assem 转换模型进行稍微修改，实现 MeSH 主题词－副主题词的组配[5]，2006 年 Bio2RDF 项目采用 Web 语义技术对 MeSH 转换[6-7]及 BioPortal 对其进行尝试。这些机构的尝试仅涵盖 MeSH 功能的

〔收稿日期〕 2019－04－26

〔作者简介〕 吴思竹，博士，副研究员，发表论文 40 余篇；通讯作者：钱庆，研究员，发表论文 60 余篇。

〔基金项目〕 国家社会科学基金青年项目“基于 R2RML 的 RDB 到 RDF 的转换模式研究与实现”（项目编号：13CTQ009）。

一部分，缺少原始资源中一些细节，如主题词－副主题词的组配、上下位关系等。2009 年美国国立医学图书馆（National Library of Medicine，NLM）的医学本体研究组织（Medical Ontology Research，MOR），利用扩展样式表转换语言（Extensible Stylesheet Language Transformations，XSLT）将 MeSH 的可扩展标记语言（Extensible Markup Language，XML）表示形式转换成 RDF 且对其定期维护更新至今[8]。该版本保留 MeSH 的 3 级概念结构，实现上下位关系，具有一定的权威性和借鉴性。但 MeSH 结构太复杂，限制在 XML 版本中明确呈现 MeSH 特征的 RDF 表示方式不能很好地表达一些重要特征（包括主题词之间的层级关系），也不能实现主题词－副主题词组配。

2 R2RML 映射语言

2003 年 W3C 发布了调查报告"Mapping Semantic Web Data with RDBMSes"[9]，其中分析进行 RDB2RDF 需要解决的主要问题及现有解决方案和相关工具，指出 RDB2RDF 的关键是定义一种映射语言。2012 年 W3C 推荐 Direct Mapping 和 R2RML 两种映射语言。Direct Mapping 为直接映射，将关系数据库表结构和数据直接输出为 RDF 图，RDF 图中用于表示类和谓词的术语与关系数据库中的表名和字段名保持一致，完全是关系数据库数据结构的反映，且不可更改[10]；而 R2RML 则有高度的可定制性。R2RML 映射是指从关系库中检索数据的逻辑表（LogicalTable），如基本表、视图或有效结构化查询语言（Structured Query Language，SQL）查询，然后使用三元组映射（TriplesMap）将每个逻辑表映射为 RDF[11]。三元组映射是指主语映射（SubjectMap）、谓语映射（PredicateMap）和宾语映射（RefObjectMap）或引用客体映射（RefObjectMap）。R2RML 逻辑框架，见图 1。

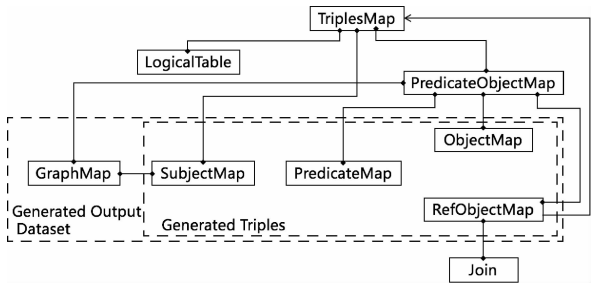

图 1 R2RML 逻辑框架[5]

3 医学主题词表（MeSH）

3.1 结构

MeSH 具有一套完整的数据体系结构，包括主题词表、副主题词表、增补概念词表、树形结构表等部分，采用 3 级概念结构模式进行组织，第 1 层由主题词、副主题词和补充概念记录（SCR）组成，第 2 层由念的同义术语组成，第 3 层由术语组成。3 级结构可以将 MeSH 的语义关系明确清晰地标识出来，描述信息也从原来的主题词级别细分到主题词、概念、术语级别。这种结构有利于计算机理解和处理，支持从多个维度组织和查询生物医学信息资源，有助于整体提升 MeSH 的易用性及性能。

3.2 解析

从 NLH 官网下载 XML 格式的 MeSH 词表（2017 版），通过解析 MeSH 的 XML 获得主题词表、副主题词表、补充概念记录表、概念表等多个词表，导入数据库中。为减少数据库冗余度和 RDB2RDF 的时间消耗及避免空白节点，对 MeSH 数据库的原始表格重新进行数据建模和适当的拆分合并，最终共建立 21 个表格。统计解析后的 MeSH 词表，共有主题词 28 472 个，副主题词 80 个，主题词术语 115 845 个，与官网数据相一致[12]。MeSH 各个词表数据关系，见图 2。其中 descriptor1 表存储的是 descriptor 基本属性中一对一的属性，descriptor2 表中存储的

是 descriptor 基本属性中一对多或多对多的属性；同理表 concept1、concept2、scrs1 和 scrs2。这样有利于减少节点冗余度和时间消耗。

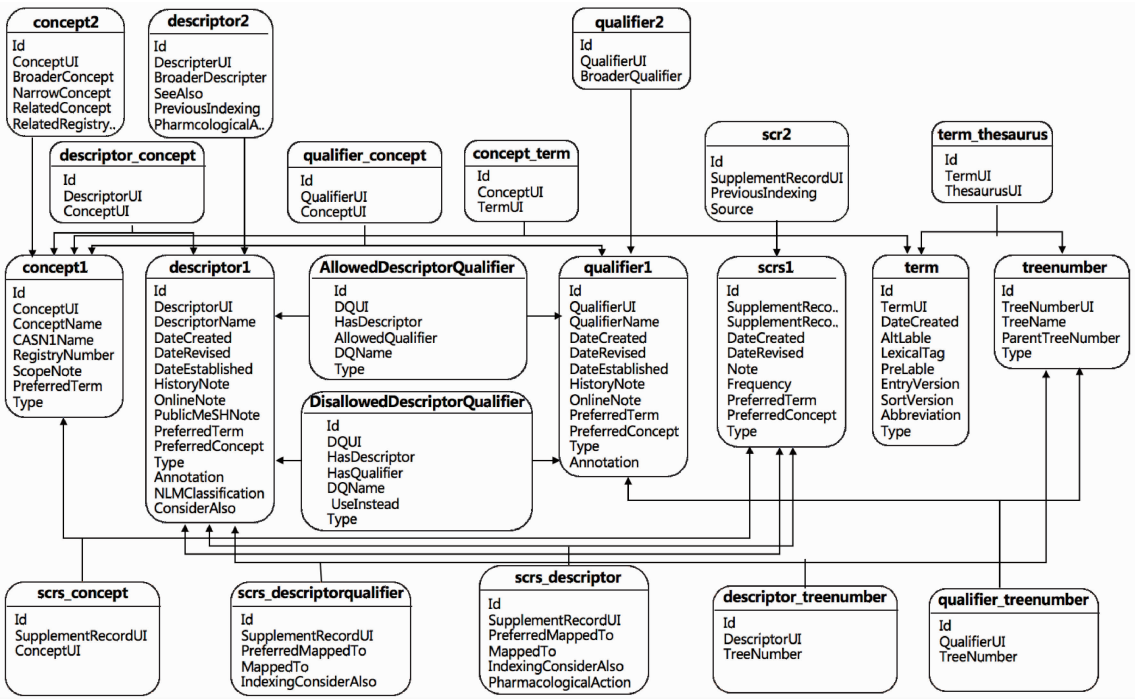

图 2 MeSH 各词表数据关系

4 R2RML 映射文档

4.1 步骤

本实验在将 MeSH 从 RDB 格式转换 RDF 的过程中，借鉴 NLM 在利用 XSLT 将 MeSH 的 XML 表示形式转换成 RDF 过程中对 MeSH 类和谓语的规定，通过 R2RML 映射规则将其转换成 RDF。根据 MeSH 体系结构规定 16 个类，各类之间的关系，见图 3。

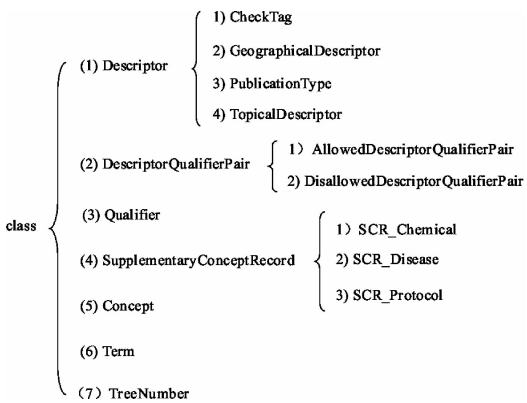

图 3 MeSH 各类之间的关系

根据 MeSH 的 XML 表示形式解析情况规定 abbreviation、allowableQualifier、altLabel、annotation、broaderConcept、broaderDescriptor 等 47 个谓语。相较于 NLM 对 MeSH 谓语的规定，本实验少了两个谓语 meshv：active 和 meshv：lastActiveYear。因查找 MeSH 的 XML 格式及 XSD 格式均未查到这两个属性。根据 MeSH 各词表在关系数据库中的存储情况及上述规定的各个类和谓语，分别定义 R2RML 映射文档，将存储在关系数据库表中的数据转换为 RDF 数据。书写映射文档，首先需定义 R2RML 的命名空间 rr 以及 MeSH 医学主题词表数据描述使用的命名空间 ex 及 RDF 语法模式的命名空间 rdf、rdfs 等。在映射文档中需要定义一系列 RDF 术语，如语言标签 rr：language、数据类型 rr：datatype 等。根据 R2RML 映射模式将数据库中 MeSH 词表的所有映射文档分为 4 类。

4.2 一表对一表映射

即主题词表、副主题词表、补充概念记录词

表、概念表等的基本属性表，其无需与其他表进行关联，现以 term 基本属性表为例。模式 1：为表创建三元组映射，指定表名对应 rr：logicalTable 的值。在 TriplesMap 中使用 rr：template 创建 rr：subjectMap 来定义每一行的 URI template，具有映射文档如下：

```
<TriplesMap_term>
    rr:logicalTable [ rr:tableName "term" ];
    rr:subjectMap [ rr:template "http://imicams.ac.cn/mesh/{TermUI}";rr:class ex:Term];
    rr:predicateObjectMap [rr:predicateex:abbreviation; rr:objectMap[rr:column "Abbreviation";rr:language "en"];];
    rr:predicateObjectMap [rr:predicateex:altLabe; rr:objectMap[rr:column "AltLabel";rr:language "en"];];
```

这是一个表映射的直接映射。对于表的本体类，用户可以在主语映射 rr：subjectMap 中利用 rr：class 为映射表设置相应本体类（如 < TriplesMap_term > ），用户也可以在谓语对象映射（rr：predicateObjectMap）中选择指定表的特定本体类，如 descriptor 基本属性表的映射：

```
<TriplesMap_descriptor>
    rr:logicalTable [ rr:tableName "descriptor1" ];
    rr:subjectMap [ rr:template "http://imicams.ac.cn/mesh/{DescriptorUI}"];
    rr:predicateObjectMap [rr:predicaterdf:type; rr:objectMap[ rr:template
"http://imicams.ac.cn/mesh/vocab#/{Type}"];]
```

4.3 一属性对一属性的映射

各 MeSH 词表中无需与其他列进行关联的各列的映射，如 DateCreated、DateEstablished、DateRevised、HistoryNote 等，现以 treenumber 表为例。模式 2：给定一个 TriplesMap，为该属性创建 rr：predicateObjectMap，对于本体属性和 rr：objectMap 属性只有一个 rr：predicate，具有映射文档如下：

```
<TriplesMap_treenumber>
    rr:logicalTable [ rr:tableName "treenumber" ];
    rr:subjectMap [ rr:template "http://imicams.ac.cn/mesh/{TreeNumberUI}";rr:class ex:TreeNumber];
    rr:predicateObjectMap [rr:predicateex:parentTreeNumber;
    rr:objectMap[rr:template"http://imicams.ac.cn/mesh/{ParentTreeNumber}"];];
```

这是属性映射的一个直接映射，可为各个属性自动生成一个唯一的本体属性。

4.4 两表之间含外键映射

MeSH 中含外键的两表之间的映射，如主题词表、副主题表与概念表、术语表等之间的映射，主题词表、副主题表均含有外键 PreferredConcept、PreferredTerm，现以 concept 表与 term 表之间的映射为例。模式 3：给定两张表，一张表视为子类，另一张表则是父类。为每个表创建一个 TriplesMap。给子类 TripleMap 创建一个 rr：predicateObjectMap，除 rr：predicate 外还有一个 rr：objectMap，其具有一个 rr：parentTripleMap 和一个 rr：joinCondition。rr：parentTripleMap 将指向父类 TripleMap，rr：joinCondition 将有一个 rr：child 和 rr：parent，分别表示子表和父表中的连接属性。具有映射文档如下：

```
<TriplesMap_concept>
    rr:logicalTable [ rr:tableName "concept1" ];
    rr:subjectMap [ rr:template "http://imicams.ac.cn/mesh/{ConceptUI}";rr:class ex:Concept];
    rr:predicateObjectMap [rr:predicateex:identifier; rr:objectMap[rr:column "ConceptUI"];];
    rr:predicateObjectMap [rr:predicate ex:preferredTerm;
      rr:objectMap[a rr:RefObjectMap;
        rr:parentTriplesMap <Table_Map_term>;
        rr:joinCondition [rr:child "PreferredTerm";rr:parent "TermUI";];];].
<TriplesMap_term>
    rr:logicalTable [ rr:tableName "term" ];
    rr:subjectMap [ rr:template "http://imicams.ac.cn/mesh/{TermUI}"; rr:class ex:Term].
```

4.5 含外键的多表之间映射

模式 4：创建一个具有 R2RML 视图的 TriplesMap，该视图由 rr：logicalTable 组成，rr：logicalTable 包含一个 rr：sqlQuery，其包含一个表示连接的 SQL 查询，其 R2RML 映射模式如下：

```
<Table_Map_descriptor_concept>
    rr:logicalTable [ rr:sqlQuery """
        SELECT dc.DescriptorUI AS DescriptorUIdc,d1.DescriptorUI AS DescriptorUI1,dc.Concept AS
Concept,c.ConceptUI AS ConceptUIFROM descriptor1 d1,descriptor_concept dc,concept1 c   WHERE
d1.DescriptorUI=dc.DescriptorUI   AND dc.Concept = c.ConceptUI """ ];
    rr:subjectMap [ rr:template "http://imicams.ac.cn/mesh/{DescriptorUI1}"];
    rr:predicateObjectMap [rr:predicateex:concept; rr:objectMap[ rr:template
"http://imicams.ac.cn/mesh/{ConceptUI}"];].
```

如果是两个表之间的映射可以使用模式 4，但多个表之间的映射则必须使用模式 4 的 SQL 查询。模式 4 与模式 3 不同的是：模式 3 用户需要添加额外的三元组实例来映射连接两表的属性；而模式 4 用户是通过修改 SQL 查询来增加连接属性。

4.6 链接多个表之间的映射

模式 5：为多对多表创建 TriplesMap。指定 rr：logicalTable，其值对应于多对多表的表名。在 TriplesMap 中使用 rr：template 创建一个 rr：subjectMap，以定义多对多关系中一个表的 URI template。创建一个 rr：predicateObjectMap 实例，其具有本体属性的 rr：predicate。最后使用 rr：template 创建一个 rr：objectMap 来定义多对多关系的另一个表的 URI template，具体如下：

```
<TriplesMap_descriptor2_descriptor1>
    rr:logicalTable [ rr:sqlQuery """
        SELECT d1.DescriptorUI AS DescriptorUI1,d2.DescriptorUI AS DescriptorUI2,d2.PharmacologicalAction AS
PharmacologicalAction,d2.PreviousIndexing AS PreviousIndexing,d2.SeeAlso AS SeeAlso,d2.BroaderDescriptor AS
BroaderDescriptor      FROM descriptor1 d1,descriptor2 d2 WHERE d1.DescriptorUI=d2.DescriptorUI """ ];
    rr:subjectMap [ rr:template "http://imicams.ac.cn/mesh/{DescriptorUI1}"];
    rr:predicateObjectMap [rr:predicateex:broaderDescriptor; rr:objectMap[rr:template
"http://imicams.ac.cn/mesh/{BroaderDescriptor}"];];
    rr:predicateObjectMap [rr:predicateex: allowedQualifier; rr:objectMap[rr:template
"http://imicams.ac.cn/mesh/{ AllowedQualifier}"];];
```

5 系统实现

在进行数据转换的过程中通过调研和比较选择的编程语言是 Java，数据库服务器使用 MySQL，转换工具选择第 3 方工具 DB2Triples。DB2Triples[13] 是由 Antidot 公司开发的用于从关系型数据库中抽取数据并将其转换为 RDF 三元组存储的开源工具。其同时支持 R2RML 和 Direct Mapping 两种映射语言标准。DB2Triples 支持数据实体化的映射实现方式，但不提供数据查询方式。实体化后的 RDF 图可以 RDF/XML、N3、N - Triples 或 Turtle 格式进行序列化。在 Direct Mapping 模式下可选择使用来自 SPARQL 文件的查询以转换 RDF 图。根据上述书写好的 R2RML 映射规则，基于 MySQL 数据库和 Java 编程语言，利用 DB2Triples 实现 RDB2RDF 的转换。现以 Descriptor 表为例，简单介绍其数据转换过程：新建 eclipse 的工作区 workplace 和 java project；将 DB2Triples 工具包导入 eclipse。代码如下：

```
public class Descriptor1 {
    public static void main(String[] args) {
        Descriptor1 mydescriptor1=new Descriptor1();
        myd  escriptor1. Descriptor1 ();
    }
    public Descriptor1() {
    }
    public void Descriptor1 (){
        String argstr="-b "+DB_NAME+" -l "+DB_URL+" -m r2rml  -o "
                +outputfile+" -p "+DB_PASSWD+" -r "+mappingfile+" -t NTRIPLES -u "+DB_USER+" -f";
        String[] args=argstr.split(" ");
        Db2triples mytri=new Db2triples();
        Db2triples.mainrun(args);
    }
    }
```

将上述映射文档 descriptor1. ttl 存储在 java project 中；运行 descriptor1. java 代码，输出 RDF 文件 descriptor1. n3。其 RDF 片段如下：

```
<http://imicams.ac.cn/mesh/D000005> <http://www.w3.org/1999/02/22-rdf-syntax-ns#type>
<http://imicams.ac.cn/mesh/vocab#/TopicalDescriptor> .
<http://imicams.ac.cn/mesh/D000005> <http://imicams.ac.cn/mesh/vocab#dateEstablished> "1966-01-
01"^^<http://www.w3.org/2001/XMLSchema#date> .
<http://imicams.ac.cn/mesh/D000005> <http://imicams.ac.cn/mesh/vocab#annotation> "GEN: prefer specifics; abdom
muscles = ABDOMINAL MUSCLES but RECTUS ABDOMINIS is available; abdominal pain = ABDOMINAL PAIN; abrupt dis
requiring e  merg surg = ABDOMEN, ACUTE"@en .
<http://imicams.ac.cn/mesh/D000005> <http://imicams.ac.cn/mesh/vocab#dateCreated> "1999-01-
01"^^<http://www.w3.org/2001/XMLSchema#date> .
<http://imicams.ac.cn/mesh/D000005> <http://imicams.ac.cn/mesh/vocab#identifier> "D000005" .
```

6 结语

本文基于 R2RML 映射语言，利用 R2RML Tool 工具实现医学主题词表的 RDF 转换。无论关系数据库表结构如何都可以进行映射转换。关系数据库相较于 XML 格式的数据存储，其数据组织更加灵活。但正因如此原始的关系数据模型与 RDF 数据模型往往会存在不完全匹配的情况[13]。因此为达到更优的映射效果，不仅需要构建好的映射规则，在应用 R2RML 映射规则进行 RDB2RDF 数据转换之前不要受当前关系数据库表结构的局限，根据所映射的 RDF 实体之间的关系适当对原有关系表进行拆分和合并，重新进行数据建模，做好关系数据库的设计，构建好数据库的主外键关联关系，以通过 R2RML 标准映射规则的编写达到相对合适的转换要求，保证转换效果。

R2RML 提供一种将关系数据库中数据结构映射为 RDF 数据模型的便捷方法，提高不同工具平台之间的互操作性，有利于促进 RDF 数据以及关联数据的产生和更广泛的应用。R2RML 映射规则还存在一定的不可移植性以及非健壮性[14]，如当关系数据库模式发生变化时 R2RML 映射文档基本需要重新映射和修改。

参考文献

1 Bernerslee T, Hendler J, Lassila O. The Semantic Web [J]. Scientific American, 2001, 284 (5): 34 - 43.

2 Williams A. The Growth of Linked Data [EB/OL]. [2018 - 12 - 12]. http://readwrite.com/2011/01/18/the - concept - of - linked - data.

3 高岚. 网络医学信息资源检索 [M]. 北京：化学工业出版社，2005：5.

4 Assem M, Menken M R, Schreiber G, et al. A Method for Converting Thesauri to RDF/OWL [C]. Berlin: Semantic Web - iswc: Third International Semantic Web Conference, 2014: 17 - 31.

5 Kanda J, Kaynar L, Kanda Y, et al. Pre - engraftment Syndrome After Myeloablative Dual Umbilical Cord Blood Transplantation: risk factors and response to treatment [J]. Bone marrow transplantation, 2013, 48 (7): 926 - 931.

6 Belleau F, Nolin M A, Tourigny N, et al. Bio2RDF: towards a mashup to build bioinformatics knowledge systems [J]. Journal of Biomedical Informatics, 2008, 41 (5): 706 - 716.

7 Callahan A, José Cruz - Toledo, Dumontier M. Ontology - based Querying with Bio2RDF's Linked Open Data [J]. Journal of Biomedical Semantics, 2013 (4): 4 - 17.

8 Bushman B, Anderson D, Fu G. Transforming the Medical Subject Headings into Linked Data: creating the qauthorized version of MeSH in RDF [J]. Journal of Library Metadata, 2015, 15 (3 - 4): 157 - 176.

9 Beckett D, Grant J. Mapping Semantic Web Data with RDBMSes [EB/OL]. [2017 - 01 - 23]. http://www.w3.org/2001/sw/Europe/reports/scalable_rdbms_mapping_re - port/#sec - mapping.

10 夏翠娟. RDB2RDF 标准及应用研究 [J]. 现代图书情报技术，2013 (4)：10 - 17.

11 Das S, Sundara S, CYGANIAK R. R2RML: RDB to RDF mapping language [EB/OL]. [2018 - 12 - 12] http://www.w3.org/TR/r2rml/.

12 NLM. What's New for 2017 MeSH [EB/OL]. [2018 - 12 - 03]. https://www.nlm.nih.gov/pubs/techbull/nd16/nd16_mesh.html.

13 沈志宏，刘筱敏，郭学兵，等. 关联数据发布流程与关键问题研究——以科技文献、科学数据发布为例 [J]. 中国图书馆学报，2013，39 (2)：53 - 62.

14 王思丽，祝忠明，姚晓娜. 机构知识库语义知识获取方法分析及实验研究 [J]. 现代图书情报技术，2014，30 (4)：7 - 13.