基于 R2RML 的 STKOS 超级科技词表 RDF 转换实现*

王　颖[1]　吴思竹[2]
[1](中国科学院文献情报中心　北京 100190)
[2](中国医学科学院医学信息研究所　北京 100020)

摘要: 【目的】实现 STKOS 超级科技词表从关系数据库到 RDF 数据的自动转换。【方法】构建 STKOS 超级科技词表语义描述模型, 针对 STKOS 超级科技词表的数据存储情况和数据特点, 分别建立将科技术语、规范概念、范畴类、来源概念和术语等从关系数据库存储字段转换到 RDF 数据集的 R2RML 映射文档, 并利用 R2RML Parser 工具执行自动批量转换。【结果】完成 STKOS 超级科技词表大规模发布数据的 RDF 转换, 生成 1.9 亿 RDF 三元组, 并存入 Virtuoso 数据库中提供 SPARQL 查询功能。【局限】R2RML 的自定义谓词不够灵活, 对于复杂数据结构需要进行预先拆分和转换。【结论】本文基于 R2RML 开展了 STKOS 超级科技词表的 RDF 转换实践, 其映射方法可为其他关系数据库或叙词表的 RDF 转换提供参考。

关键词: R2RML　STKOS 超级科技词表　RDF
分类号: TP393
DOI: 10.11925/infotech.2096-3467.2018.0423

1　引　言

叙词表、主题词表、分类表、术语表等知识组织体系一直在信息资源组织、揭示、检索、发现等方面发挥着巨大作用。随着语义网和关联数据的发展, 传统知识组织体系也纷纷定义语义化描述模型, 采用 SKOS 作为词表的描述标识, 将词表转化为轻量级本体, 或者发布成关联数据服务, 其存储方式也从关系型数据库(Relational DataBase, RDB)、XML 格式转换为 RDF、OWL 格式。例如, 美国国家癌症研究所的 NCI 叙词表早已发布其 OWL 版本[1]; 美国国立医学图书馆开始提供 MeSH 的官方 RDF 发布[2]; 联合国粮食及农业组织的 AGROVOC 多语言叙词表基于 SKOS-XL 概念框架, 以 RDF 格式存储, 已作为一个关联开放数据集(Linked Open Data, LOD)发布并与农业领域相关的 16 个多语言知识组织体系建立关联[3]。

由国家科技文献信息中心牵头组织实施的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和示范应用”, 构建了面向外文科技文献的知识组织体系(Scientific & Technological Knowledge Organization Systems, STKOS), 为我国海量外文科技文献信息的组织和利用提供支撑[4]。项目建成的超级科技词表 STKOS 融合了词表、术语表、叙词表等各种知识组织素材, 以科技术语为基本单元, 以概念为核心, 以来源词表的原有关系为依托, 是通过概念与来源词表术语进行语义关联的词网络[4]。为实现 STKOS 超级科技词表的语义化表示, 提供开放、共享、利用服务, 建立与其他知识组织体系的关联, 有必要

通讯作者: 吴思竹, ORCID: 0000-0003-4540-9910, E-mail: wu.sizhu@imicams.ac.cn。

*本文系国家科技图书文献中心“下一代国家科技创新知识服务开放系统”先期研发任务课题“STKOS 关联数据发布及开放共享关键技术研究”(项目编号: XQYF0104)和国家社会科学青年基金项目“基于 R2RML 的 RDB 到 RDF 转换模式研究与实现”(项目编号: 13CTQ009)的研究成果之一。

实现 STKOS 超级科技词表的 RDF 转换和存储。

为促进语义网的应用，万维网联盟(W3C)的 RDB2RDF 工作组制定了 R2RML 映射语言，将关系数据库中数据快捷地转换成 RDF 数据格式。本文基于 R2RML 设计实现 STKOS 超级科技词表的映射文档，将 STKOS 超级词表中的科技术语、规范概念、范畴类、来源概念和术语等进行 RDF 描述，利用 R2RML Parser 工具自动执行批量转换并将转换文件导入 Virtuoso 数据库，提供 SPARQL 查询服务，完成 STKOS 超级科技词表的 RDF 转换和利用。

2 R2RML 映射语言

R2RML 是一种将关系数据库映射为 RDF 数据集的规范语言，2012 年成为 W3C 的推荐标准。R2RML 将关系库中的一个逻辑表(LogicalTable)，如一个基本表、视图或者有效的 SQL 查询，通过三元组映射(TriplesMap)将其中的每一行映射为一系列 RDF 三元组[5]。主体映射(SubjectMap)从逻辑表中生成 RDF 三元组中的主体(Subject)，主体通常是基于表的主键列生成的 IRI[5]。谓词客体映射(PredicateObjectMap)由谓词映射(PredicateMap)和客体映射(ObjectMap)或引用客体映射(RefObjectMap)组成，引用客体映射由连接(Join)定义映射的条件[5]。默认情况下，所有 RDF 三元组都在输出数据集的默认图中，一个三元组映射也可以包括图映射(GraphMap)，将一些或全部三元组放入命名图中[5]。R2RML 的逻辑框架如图 1 所示。

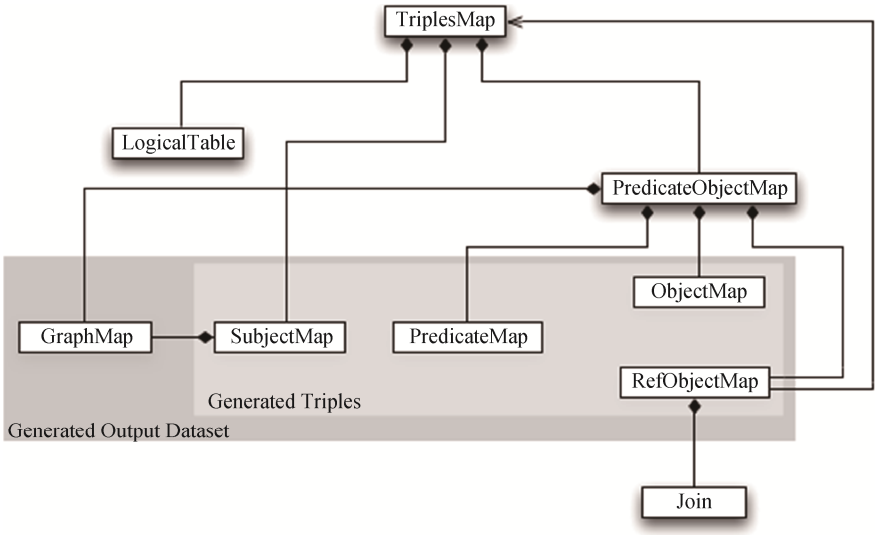

图 1 R2RML 逻辑框架[5]

3 STKOS 超级科技词表的语义描述模型

STKOS 超级科技词表由基础词库、规范概念集和范畴体系三个层次构成，三部分相互依托，构成一个有机的整体[4]。其中，规范概念指从科技术语、来源术语等途径提炼形成的科技领域概念；科技术语是用来表达规范概念的术语，是规范概念的名称；来源概念和术语指来源于经遴选、格式转换和规范后的概念和术语；范畴表用来组织概念或术语，一般采用分类表结构组织范畴类目；范畴类是范畴表中的类目[6]。

为描述 STKOS 超级科技词表各元素的语义信息和它们之间的相互关联，本文在 STKOS 超级科技词表发布元数据规范的基础上定义其语义描述模型，如图 2 所示。使用 stkos:term、stkos:concept、stkos:categoryClass、stkos:category、stkos:sourceTerm、stkos:sourceConcept、stkos:thesaurus 分别定义科技术语、规范概念、范畴类、范畴表、来源术语、来源概念、来源词表等，其中 stkos 用于定义 STKOS 数据描述使用的命名空间“http://www.nstl.gov.cn/stkos/”。科技术语、来源概念、来源术语均采用 iso25964:lexicalValue 声明英文或中文名称，使用 stkos:LinktoSourceTerm、stkos:LinkToSourceConcept、stkos:LinktoConcept、skos:inScheme 声明科技术语连接的来源术语、来源概念、规范概念和词表；规范概念通过 iso25964: hasPreferredTerm 连接优选科技术语，使用 stkos:LinkToTerm 连接非优选科技术语，

并利用 iso25964:hasPreferredDefinition、iso25964:Definition 定义优选释义和非优选释义，通过 stkos:LinktoCategoryClass 连接范畴类；范畴类采用 skos:prefLabel 和 skos:altLabel 声明优选词和非优选词，skos:broader 和 skos:narrower 声明范畴类之间的层级关系，skos:inScheme 连接范畴表。由于 STKOS 继承来源词表的原有关系和属性，仅对多来源的关系属性进行梳理和必须的纠正，关系属性为来源概念和术语所有而非科技术语或规范概念，为此本文使用 stkos:rel/*、stkos:attribute/*分别声明来源概念和术语的相关关系(如用、用和、组代、参等)、属性(如术语类型、专业代码、注释、词频等)，skos:broader 和 skos:narrower 声明来源概念之间的层级关系。此外，使用 dc:title、dc:alternative、dc:type、stkos:admin、stkos:edition 分别声明来源词表或范畴表的题名、交替题名、类型、创建者、版本等信息。

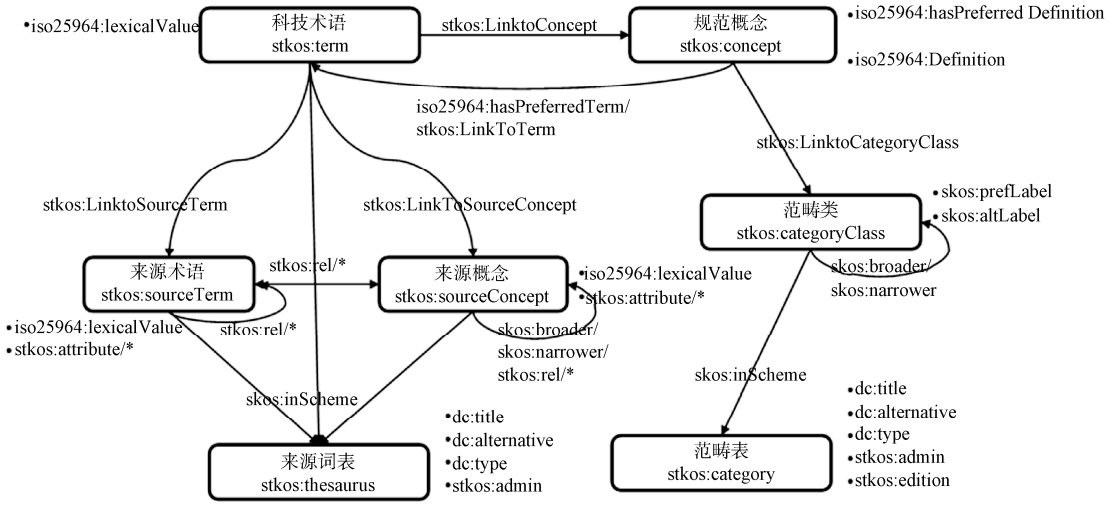

图 2　STKOS 超级科技词表语义描述模型

4　R2RML 映射文档

STKOS 超级科技词表发布数据分别存储在 MySQL 数据库的 10 个数据表中，表存储结构和字段如图 3 所示。基础术语表 stkos_atom 表存储科技术语的基本信息以及科技术语与规范概念、来源概念和术语之间的关联信息，stkos_concept 和 stkos_concept_def 存储规范概念的基本信息和释义，stkos_c2c 和 stkos_category 存储概念与范畴类的连接关系以及范畴类的基本信息和层级关系，stkos_attribute、stkos_hierarchies、stkos_relationship 存储来源概念和术语的属性、层级关系、关联关系等信息，stkos_translation 存储科技术语的译名数据，stkos_sab 存储词表的基本信息。

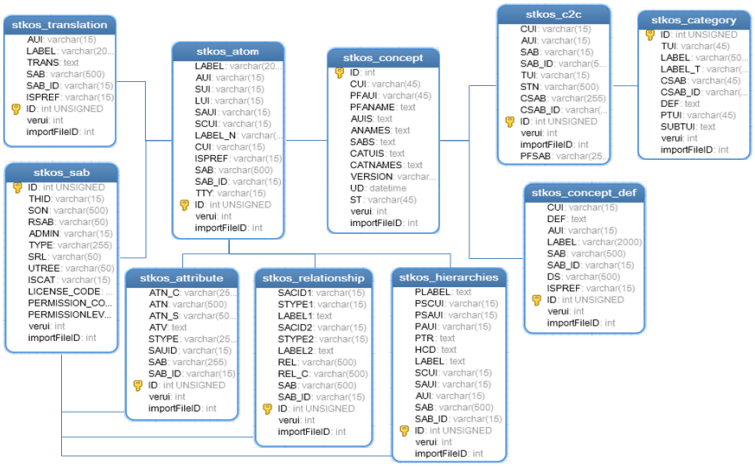

图 3　STKOS 超级科技词表关系数据表

根据 STKOS 超级科技词表的语义描述模型，分别定义 R2RML 映射文档，建立从关系数据库到 RDF 数据集的映射关系。

4.1 科技术语映射

关系数据表 stkos_atom 存储科技术语的基本信息，利用 R2RML 映射其中相关字段的代码如下所示。

```
@prefix   rr:<http://www.w3.org/ns/r2rml#>.
@prefix   stkos:<http://www.nstl.gov.cn/stkos/>.
@prefix   iso25964:<http://iso25964.org/>.
@prefix   dc:<http://purl.org/dc/elements/1.1/>.
@prefix   skos:<http://www.w3.org/2004/02/skos/core#>.

<#TriplesMap_term>
 rr:logicalTable [rr:sqlQuery """
  select AUI, LABEL, SAUI, SCUI, CUI, SAB_ID from stkos_atom
 """];
 rr:subjectMap [
 rr:template "http://www.nstl.gov.cn/stkos/term/{AUI}";
 rr:class   stkos:term;
 ];
 rr:predicateObjectMap [
 rr:predicate iso25964:lexicalValue;
 rr:objectMap [rr:column "LABEL"; rr:language "en"];
 ];
 rr:predicateObjectMap [
 rr:predicate stkos:LinkToSourceTerm;
 rr:objectMap [rr:column "http://www.nstl.gov.cn/stkos/
   sourceTerm/{SAUI}"];
 ];
 rr:predicateObjectMap [
 rr:predicate stkos:LinkToSourceConcept;
 rr:objectMap [rr:column "http://www.nstl.gov.cn/stkos/
   sourceConcept/{SCUI}"];
 ];
 rr:predicateObjectMap [
 rr:predicate stkos:LinkToConcept;
 rr:objectMap [rr:column "http://www.nstl.gov.cn/stkos/
   concept/{CUI}"];
 ];
 rr:predicateObjectMap [
 rr:predicate skos:inScheme
 rr:objectMap [rr:column "http://www.nstl.gov.cn/stkos/
    thesaurus/{SAB_ID}"];
 ].
```

首先定义 R2RML 命名空间 rr 以及 STKOS 超级科技词表数据描述使用的命名空间，如 stkos、iso25964、dc、dcterms、skos 等。使用 rr:sqlQuery 定义 SQL 语句查询 stkos_atom 表中相关的字段，将查询结果作为一个逻辑表 rr:logicalTable 进行映射。通过主体映射 rr:subjectMap 将表中 AUI 字段映射为一个科技术语类 stkos:term，并建立谓词客体映射 rr:predicateObjectMap 将科技术语的属性值和关系对象的对应字段转换为以该科技术语为主体、特定谓词和对应客体组成的一系列三元组。如定义谓词 rr:predicate 为 iso25964:lexicalValue，rr:objectMap 为 [rr:column "LABEL"; rr:language "en"]，将 LABEL 字段每一行的取值声明为对应科技术语 AUI 的名称，并通过 rr:language 定义英文语言标签。同样，通过 rr:predicateObjectMap 分别对 SAUI、SCUI、CUI、SAB_ID 字段进行映射，使用谓词 stkos:LinktoSourceTerm、stkos:LinkToSourceConcept、stkos:LinktoConcept、skos:inScheme 声明科技术语连接的来源术语、来源概念、概念和来源词表。

针对科技术语的规范译名数据，类似的建立相应利用映射文档对 stkos_translation 表中的 TRANS 字段进行数据转换，使用谓词 iso25964:lexicalValue 声明科技术语的字符串，并通过 rr:language 定义为中文语言标签。

4.2 规范概念映射

规范概念的数据存储在 stkos_concept 和 stkos_concept_def 表中，其中主要字段的映射关系如表 1 所示。直接将 stkos_concept 关系表作为 rr:logicalTable，利用 rr:subjectMap 将 CUI 字段作为三元组的主体，映射为一个规范概念 stkos:concept。在 rr:predicateObjectMap 中使用谓词 iso25964:hasPreferredTerm 定义其优选术语，将 AUI 字段映射为此规范概念的优选科技术语客体，使用 stkos:LinktoCategoryClass 连接范畴类，将 TUI 字段映射为 CUI 主体连接的范畴类客体。

针对概念释义表 stkos_concept_def，对 CUI 字段进行主体映射，利用 sqlQuery 查询语句判断 ISPREF 字段是否为"YES"，如果是，则使用 iso25964: hasPreferredDefinition 作为谓词，映射 DEF 字段为三元组的优选释义客体，否则使用 iso25964:Definition 作为谓词，映射 DEF 字段为非优选释义客体。

表 1　规范概念映射

数据表	字段	logicalTable	rr:subjectMap	rr:predicate	rr:objectMap
stkos_concept	AUI	[rr:tableName “stkos_concept”]	[rr:template “http://www.nstl.gov.cn/stkos/concept/{CUI}”; rr:class stkos:concept;]	iso25964:hasPreferredTerm	[rr:column “http://www.nstl.gov.cn/stkos/term/{AUI}”]
	TUI			stkos:LinktoCategoryClass	[rr:column “http://www.nstl.gov.cn/stkos/category/{TUI}”]
stkos_concept_def	DEF	[rr:sqlQuery “““select CUI, DEF, ISPREF from stkos_concept_def where ISPREF=’YES’ “““]	[rr:template “http://www.nstl.gov.cn/stkos/concept/{CUI}”;]	iso25964:hasPreferredDefinition	[rr:column “DEF”]
stkos_concept_def	DEF	[rr:sqlQuery “““ select CUI, DEF, ISPREF from stkos_concept_def where ISPREF=’NO’ “““]	[rr:template “http://www.nstl.gov.cn/stkos/concept/{CUI}”;]	iso25964:Definition	[rr:column “DEF”]

4.3　范畴类映射

针对范畴类表 stkos_category, 建立如表 2 所示的映射关系。将 TUI 字段作为三元组的主体, 映射为一个范畴类 stkos:catgoryClass, 使用谓词 skos:prefLabel 定义其优选词 LABEL, skos:altLabel 定义其替代词 LABEL_T, 并分别使用语言标签进行表示。此外, 使用谓词 skos:inScheme 将 CSAB_ID 转换为三元组客体, 表示范畴类连接的范畴表, 使用 skos:broader 将 PTUI 字段映射为此范畴类的上级范畴类。

表 2　范畴类映射

数据表	字段	logicalTable	rr:subjectMap	rr:predicate	rr:objectMap
stkos_category	LABEL	[rr:tableName “stkos_category”]	[rr:template “http://www.nstl.gov.cn/stkos/categoryClass/{TUI}”; rr:class stkos:categoryClass;]	skos:prefLabel	[rr:column “LABEL”; rr:language “zh”]
	LABEL_T			skos:altLabel	[rr:column “LABEL_T”; rr:language “en”]
	CSAB_ID			skos:inScheme	[rr:column “http://www.nstl.gov.cn/stkos/thesaurus/{CSAB_ID}”]
	PTUI			skos:broader	[rr:column “http://www.nstl.gov.cn/stkos/categoryClass/{PTUI}”]

4.4　来源概念和术语映射

STKOS 通过对来源概念和术语进行形式化汇总、整理、规范、去重、分类等处理建设基础科技词库, stkos_atom 表的 LABEL 既存储科技术语的基本信息, 也存储来源概念和术语的名称和来源词表信息, 因此来源概念和术语映射首先对 stkos_atom 表中的 SCUI、SAUI、LABEL、SAB_ID 等字段进行映射, 部分映射示例如表 3 所示。

STKOS 加工过程中对来源属性进行分析和处理, 归类为包括语义类型、术语类型、注释、创建时间、专业代码、词频、词性等在内的 42 种属性, 并分别定义建议的属性中英文名称。如果分别定义 42 种属性的来源名称和中英文建议名称的谓词, 会增加 RDF 转换的负担, 并难于控制数据质量。为此, 本文将语义描述模型中的(来源概念或术语, stkos:attribute/*, 属性值)三元组表述方式转换为(来源概念或术语, hasAttribute, 属性 ID), (属性 ID, stkos: ATN_C, 建议中文属性名称), (属性 ID, stkos: ATN, 建议英文属性名称), (属性 ID, stkos: ATN_S, 来源属性名称), (属性 ID, stkos: ATV, 属性值)等多个三元组, 将 stkos:attribute 作为一个类, 通过其 rdf graph 描述复杂的属性信息。来源术语属性映射示例如表 3 所示, 同理可建立来源概念属性映射文档。

在转换来源概念的层级关系表 stkos_hierarchies 时, 直接对 SCUI 和 PSCUI 列映射, 并通过 skos:broader 进行连接。除层级关系外, 来源概念和术语自身或相互之间存在多种类型的相关关系, STKOS 将这些关系梳理为用、用和、组代、参、广义、窄义、其

他关系 7 种具体关系类型。R2RML 未提供对数据表字段进行谓词映射 rr:predicateMap 的明确示例，而 R2RML Parser 工具仅支持常量谓词映射，所以枚举这 7 种关系类型分别执行映射，同时根据 STYPE1 和 STYPE2 的类型分别为来源概念或来源关系对表进行拆分。如表 3 所示，定义 SQL 语句查询表中 STYPE1 和 STYPE2 为来源术语，REL 为‘用’的行，由此分批映射来源关系表 stkos_relationship 中的数据。

表 3　来源概念和术语映射

数据表	字段	logicalTable	rr:subjectMap	rr:predicate	rr:objectMap
stkos_atom	LABEL	[rr:sqlQuery “““ select LABEL, SAUI, SCUI, SAB_ID from stkos_atom“““]	[rr:template “http://www.nstl.gov.cn/stkos/term/{SAUI}”; rr:class stkos: sourceTerm;]	iso25964:lexicalValue	[rr:column “LABEL”; rr:language “en”]
	SAB_ID			skos:inScheme	[rr:column “http://www.nstl.gov.cn/stkos/thesaurus/{SAB_ID}”]
stkos_attribute	ID	[rr:sqlQuery “““select ID, SAUID, STYPE, SAB_ID from stkos_attribute where STYPE=’TERM “““];	[rr:template “http://www.nstl.gov.cn/stkos/sourceTerm/{SAUID}”;]	stkos: hasAttribute	[rr:column “http://www.nstl.gov.cn/stkos/attribute_{ID}”]
	ATN_C	[rr:sqlQuery “““ select ID, ATN_C, AN, ATN_S, ATV from stkos_attribute “““]	[rr:template “http://www.nstl.gov.cn/stkos/attribute_{ID}”; rr:class stkos: attribute;]	stkos: ATN_C	[rr:column “ATN_C”]
	ATN			stkos: ATN	[rr:column “ATN”]
	ATN_S			stkos: ATN_S	[rr:column “ATN_S”]
	ATV			stkos: ATV	[rr:column “ATV”]
stkos_hierarchies	PSCUI	[rr:sqlQuery “““ select ID, SCUI, PSCUI from stkos_hierarchies “““]	[rr:template “http://www.nstl.gov.cn/stkos/sourceConcept/{SCUI}”; rr:class stkos: sourceConcept;];	skos: broader	[rr:column “http://www.nstl.gov.cn/stkos/sourceConcept/{PSCUI}”]
stkos_relationship	SACID2	[rr:sqlQuery “““ select SACID1, STYPE1, SACID2, STYPE2, REL from stkos_relationship where STYPE1=’ TERM’ and STYPE2=’ TERM’ and REL=’用’ “““];	[rr:template “http://www.nstl.gov.cn/stkos/sourceTerm/{SACID1}”;];	stkos: rel/used	[rr:column “http://www.nstl.gov.cn/stkos/sourceTerm/{SACID2}”]

除上述映射外，针对 stkos_sab 表建立映射对范畴表和来源词表的基本信息进行转换，使用 dc:title、dc:alternative、dc:type、stkos:admin、stkos:edition 分别声明来源词表或范畴表的题名、交替题名、类型、创建者、版本等信息。此外，部分双向关系仅执行了单向转换，反向关系可通过离线添加或在查询时使用正向关系的 SPARQL 查询实现，如 stkos:narrower 可通过 stkos:broader 反向查询实现。规范概念 stkos:LinkToTerm 连接的非优选科技术语可通过查询反向查询 stkos:LinkToSourceConcept 该规范概念的科技术语并排除 iso25964:hasPreferredTerm 的优选术语实现。

5　系统实现

5.1　数据转换

通过调研和比较，本文选用第三方工具 R2RML Parser 实现 STKOS 超级科技词表数据的自动转换。R2RML Parser 工具[7]覆盖了大部分 R2RML 功能，支持 PostgreSQL、MySQL 和 Oracle 等关系型数据库，允许导出 rdf/xml、n-triple、turtle、n3 等多种 RDF 文件，也可以导入 Jena 的 TDB 数据库中[8]，功能满足 STKOS 数据转换的需求。使用 R2RML Parser 转换 STKOS 超级科技词表的具体操作包括以下几部分。

(1) 下载 R2RML Parser 工具存储在本地。

(2) 新建映射文件 stkos-mapping.rdf，将上述映射文档存储在这个文件中。

(3) 修改 r2rml.properties 配置文件，几个主要参数如下所示，包括制定映射文件 mapping.file 及其文件类型，设置默认命名空间，配置源数据库 db.url、db.login、db.password、db.driver 等连接参数，设置输出文件名和格式等。

```
mapping.file=stkos-mapping.rdf
mapping.file.type=TURTLE
default.namespace=http://www.nstl.gov.cn/stkos/
db.url=jdbc:mysql://127.0.0.1:3306/stkos
db.login=root
db.password=****
db.driver=com.mysql.jdbc.Driver
jena.storeOutputModelUsingTdb=false
jena.destinationFileName=dump.rdf
jena.destinationFileSyntax=N3
```

(4) 执行 r2rml-parser.bat, 输出 dump.rdf 文件, 其中 RDF 片段如下所示。

```
@prefix  rr:<http://www.w3.org/ns/r2rml#>.
@prefix  stkos:<http://www.nstl.gov.cn/stkos/>.
@prefix  iso25964:<http://iso25964.org/>.
@prefix  dc:<http://purl.org/dc/elements/1.1/>.
@prefix  skos:<http://www.w3.org/2004/02/skos/core#>.
@prefix  dcterms:<http://purl.org/dc/terms/>.

<http://www.nstl.gov.cn/stkos/term/A022949414>
 a  stkos:term;
 iso25964:lexicalValue "Repurified Water"@en;
 stkos:LinkToSourceTerm<http://www.nstl.gov.cn/stkos/sourceTerm/66402131>;
 stkos:LinkToSourceConcept <http://www.nstl.gov.cn/stkos/sourceConcept/48480654>;
 stkos:LinkToConcept <http://www.nstl.gov.cn/stkos/concept/C018858504>;
 skos:inScheme <http://www.nstl.gov.cn/stkos/thesaurus/46504810>.
 <http://www.nstl.gov.cn/stkos/concept/C018858504>
 a stkos:concept;
 iso25964:hasPreferredTerm <http://www.nstl.gov.cn/stkos/term/A023185104>;
 stkos:LinktoCategoryClass <http://www.nstl.gov.cn/stkos/ categoryClass/A022241761>
 iso25964:hasPreferredDefinition "Denotes reclaimed or recycled wastewater that is treated far beyond the most stringent standards current in force and then remixed with fresh water to augment existing water supplies....".
 <http://www.nstl.gov.cn/stkos/categoryClass/A022241761>
 a stkos:categoryClass;
 skos:inScheme <http://www.nstl.gov.cn/stkos/thesaurus/45710659>;
 skos:prefLabel "水资源调查与水利规划"@zh;
 skos:altLabel "Water resource investigation and planning"@en;
 skos:broader <http://www.nstl.gov.cn/stkos/categoryClass/A022241757>.
 <http://www.nstl.gov.cn/stkos/thesaurus/45710659>
 a stkos:category;
 dc:title "STKOS Category";
 dctems:alternative "STKOS_C";
 dc:type "主表的范畴表";
 stkos:admin "医科".
 <http://www.nstl.gov.cn/stkos/sourceTerm/66402131>
 a sourceTerm;
 iso25964:lexicalValue "Repurified Water"@en;
 skos:inScheme <http://www.nstl.gov.cn/stkos/thesaurus/46504810>.
 < http://www.nstl.gov.cn/stkos/sourceConcept/48480654>
 a sourceConcept;
 iso25964:lexicalValue "Repurified Water"@en;
 skos:broader <http://www.nstl.gov.cn/stkos/sourceConcept/46504810>;
 skos:inScheme <http://www.nstl.gov.cn/stkos/thesaurus/46504810>.
```

经过分批次的批量转换, 利用 R2RML Parser 实现了 STKOS 超级科技词表 61.5 万规范概念、232.1 万科技术语、1.2 万范畴类等发布数据的 RDF 转换, 生成 1.9 亿 RDF 三元组。

5.2　SPARQL 查询

通过 R2RML Parser 工具自动执行批量转换操作, 可以将 RDF 三元组结果集直接存储在 TDB 数据库中, 也可以导出 RDF 文件, 再将其导入其他数据库中进行灵活应用。例如通过 Virtuoso 数据库自带的 Quad Store Upload 功能将导出的 dump.rdf 生成命名图 http://www.nstl.gov.cn/stkos, 导入 Virtuoso 数据库进行存储和管理, 如图 4 所示。

利用 Virtuoso 的 SPARQL 查询功能, 查询一个指定科技术语 <http://www.nstl.gov.cn/stkos/term/A022949414>, 返回结果如图 5 所示。

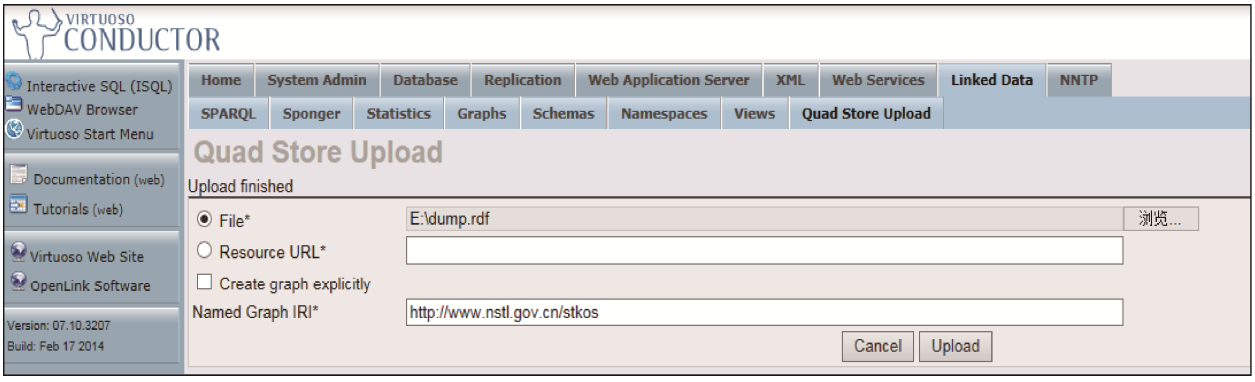

图 4　RDF 文件导入 Virtuoso 数据库[9]

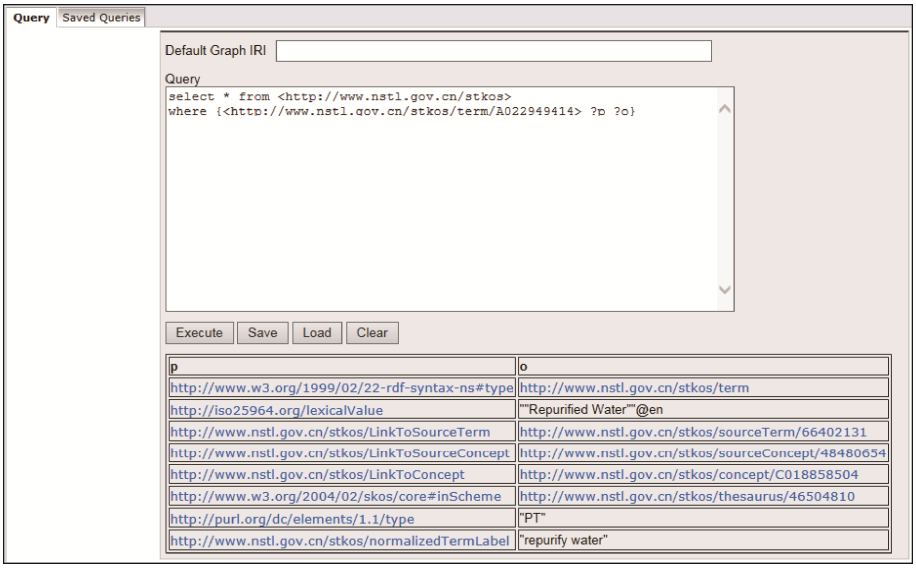

图 5　SPARQL 查询示例

6　结　语

本文基于 R2RML 映射语言，利用 R2RML Parser 工具实现 STKOS 超级科技词表的 RDF 转换。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，提高了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。而 R2RML 的语法规则也存在一定局限性，如自定义谓词不够灵活，需要在建立映射文档前对关系数据表执行查询操作或对表格进行拆分以满足映射要求，如果要对更为复杂的数据结构进行转换，需要采用更灵活的方法。在未来的工作中，将借鉴 R2RML 语言的映射逻辑，直接定义映射规则，通过 JDBC 进行数据传输和导入，并探索可扩展的通用映射模型，开发便捷的转换工具。

参考文献:

[1] NCI Enterprise Vocabulary Services[EB/OL]. [2015-03-24]. http://evs.nci.nih.gov/.

[2] Medical Subject Headings (MeSH) RDF Linked Data (beta) [EB/OL]. [2017-04-24] https://id.nlm.nih.gov/mesh/.

[3] AGROVOC Multilingual Agricultural Thesaurus [EB/OL]. [2017-04-24]. http://aims.fao.org/vest-registry/vocabularies/agrovoc-multilin_gual-agricultural-thesaurus.

[4] 孙坦，刘峥. 面向外文科技文献信息的知识组织体系建设思路[J]. 图书与情报，2013(1): 2-7. (Sun Tan, Liu Zheng. Methodology Framework of Knowledge Organization System for Scientific & Technological Literature[J]. Library and Information，2013(1): 2-7.)

[5] Das S, Sundara S, Cyganiak R. R2RML: RDB to RDF Mapping Language[EB/OL]. [2015-03-12]. http://www.w3.org/TR/r2rml/.

[6] 中国科学院文献情报中心. STKOS 超级科技词表发布元数据规范[R]. 北京：中国科学院文献情报中心，2012: 6-7. (National Science Library, Chinese Academy of Sciences. Publishing Metadata Specification for Scientific & Technological Literature[R]. Beijing: National Science Library, Chinese Academy of Sciences, 2012: 6-7.)

[7] Konstantinou N, Spanos D E, Kouis D, et al. Mitrou N: An Approach for the Incremental Export of Relational Databases into RDF Graphs[J]. International Journal on Artificial Intelligence Tools, 2015, 24(2): 1-15.

[8] Konstantinou N, Kouis D, Mitrou N. Incremental Export of Relational Database Contents into RDF Graphs[C]// Proceedings of the 4th International Conference on Web Intelligence, Mining and Semantics, Thessaloniki, Greece. 2014.

[9] OpenLink Virtuoso (Version: 07.10.3207) [CP/OL]. [2014-02-17]. https://virtuoso.openlinksw.com/.

作者贡献声明:

王颖：设计研究方案，数据清洗与实验，起草论文；
吴思竹：提出研究思路，工具测试与选型，论文修改与修订。

利益冲突声明:

本文研究中使用了 OpenLink 公司的 Virtuoso 数据库免费版本以及开源工具 R2RML Parser。所有作者声明不存在利益冲突关系。

支撑数据:

支撑数据由作者自存储, E-mail: wangying@mail.las.ac.cn。
[1] 王颖. stkos-mapping.rdf. 映射文件.
[2] 王颖. r2rml.properties. 配置文件.
[2] 王颖. dump.rdf. 转换生成的 RDF 文件.

收稿日期: 2018-04-20
收修改稿日期: 2018-07-13

Converting STKOS Metathesaurus to RDF Triples with R2RML

Wang Ying[1]　Wu Sizhu[2]
[1](National Science Library, Chinese Academy of Sciences, Beijing 100190, China)
[2](Institute of Medical Information, Chinese Academy of Medical Sciences, Beijing 100020, China)

Abstract: [Objective] This paper aims to convert STKOS Metathesaurus from records of relational database to RDF triples. **[Methods]** First, we defined the semantic schema of the STKOS based on their storage features and data characteristics. Then, we mapped the scientific terms, standard concepts, categories, as well as source concepts and terms with the help of R2RML. Finally, we converted the documents stored in relational database to RDF datasets with the R2RML parser. **[Results]** The proposed method could process STKOS metathesaurus automatically and generated 190 million RDF triples. All new records were stored in the Virtuoso database and could be queried with SPARQL. **[Limitations]** Predicates in the R2RML lacks flexibily, therefore, more complex data sets need to be splited and transformed first. **[Conclusions]** The proposed model shed light on future research on converting other relational database records or thesaurus to RDF datasets.

Keywords: R2RML　STKOS Metathesaurus　RDF