大规模RDF三元组转换及存储工具比较研究*

李悦[1] 孙坦[1,2] 赵瑞雪[1,2] 李娇[1] 黄永文[1] 罗婷婷[1] 鲜国建[1,2]
(1. 中国农业科学院农业信息研究所，北京 100081；2. 农业农村部农业大数据实验室，北京 100081)

摘要：富含语义知识的数据网络是实现大数据智能的基石。资源描述框架（Resource Description Framework，RDF）是用于描述网络资源的W3C标准。大规模转换、存储管理RDF三元组是构建关联数据网络或语义知识图谱，实现数据可查找、可访问、可交互、可再用的重要路径。本文选择国际主流的10种RDF三元组转换工具，以及6种广受欢迎的RDF存储系统，从技术原理、性能特点及应用场景等多个视角进行对比分析，并总结存在问题和不足。提出未来大规模RDF三元组数据转换与存储管理需要实现的目标是实现RDF抽取、转换和加载（ETL）的流程化和集成化，并重点支撑4类典型应用需求场景，包括从非RDF数据到RDF数据的转换，不同RDF数据格式之间的双向转换，RDF三元组在数据库之间的数据迁移，以及RDF数据的动态更新和进化管理。

关键词：RDB2RDF；RDF转换；RDF存储；大数据智能；知识图谱
中图分类号：G250　　DOI：10.3772/j.issn.1673-2286.2020.11.001
引文格式：李悦，孙坦，赵瑞雪，等. 大规模RDF三元组转换及存储工具比较研究[J]. 数字图书馆论坛，2020(11)：2-12.

当前，全球已进入数据密集型科学研究第四范式、开放科学和媒体融合发展新时代。同时，我国也正加速推进人工智能、大数据中心等新型基础设施建设。而大数据作为新的生产要素，只有在使其数据化、知识化、关联化和可计算化之后，才能真正成为大数据智能时代支撑科技创新发展的新引擎。随着语义网和知识图谱的快速发展与场景式应用，数据之间的互联互通和互操作变得至关重要。目前，海量结构化、半结构化数据广泛存储在各类关系数据库系统和文件系统中。关系型数据库由于数据结构的不同而导致互操作困难、只有结构缺乏语义[1]等缺点逐渐显现出来，而以大量分散孤立的文件存在的数据资源开放共享状况则更不乐观。

RDF是W3C为促进语义网的应用而推出的资源描述规范，包括RDF抽象模型和一组RDF编码格式规范，如RDF/XML、Turtle、N-Triples等[2]。RDF的基本模型是有向标记图，图中节点表示实体或资源，边表示实体间关系或实体的属性[3]。RDF通过主语、谓语和宾语形成的三元组描述互联网资源之间的语义关系，是实现数据可查找、可访问、可交互、可再用（FAIR原则）[4]的重要路径，有利于实现数据的共享、重用和语义互操作，是构建新型大数据基础设施的基石。而与关联数据一脉相承的知识图谱，作为信息互联、知识共享时代的知识库，在智能搜索、智慧医疗、社区推荐、网络安全等领域发挥着重要作用[5]。RDF三元组是知识图谱较为简单实用的知识表示方法[3]，基于RDF构建大规模知识图谱，可充分表达各个实体、概念间的语义关系，进而为文本理解、语义搜索、智能问答等上层应用提供有力支撑。目前，越来越多的系统和应用开始利用RDF改造底层数据结构，如联合国粮农组织创建的农业领域关联数据集AGROVOC[6]用于对农业文献信息数据库AGRIS标识资源进行索引，使搜索更加有效；英国广播公司（BBC）将关联数据技术应用于媒体领域，

*本研究得到国家社会科学基金项目“科技论文全景式摘要知识图谱构建与应用研究”（编号：19BTQ061）资助。

在降低构建成本、及时更新数据和提升用户体验等方面具有明显的应用优势[7]。

在多源异构大数据治理环境下，如何从海量异构数据中抽取、转换、映射为RDF并进行存储成为亟待解决的共性问题。目前，国际上已有众多RDF转换工具及三元组存储系统，如D2RQ、Any23、Jena-TDB、Marklogic等，研究人员通常要投入较多时间和精力，针对不同应用场景选择最适合的RDF转换工具及存储系统，并解决各类工具、系统的互操作及集成问题。因此，本文将从基于关系型数据库的RDB2RDF工具、文件格式或在线的RDF转换工具和RDF-ETL工具体系等RDF生成技术，以及RDF三元组存储中间件和数据库等方面进行多视角对比研究，最后对当前RDF转换及存储存在的不足进行总结与展望，提出大规模RDF三元组数据转换与存储管理的目标是实现RDF抽取、转换和加载（ETL）的流程化和集成化，并重点支撑4类典型应用需求场景，以期为大规模RDF三元组数据治理提供参考。

1 RDF生成技术

RDF的概念自提出以来，在各个领域得到了广泛的发展与运用。越来越多的研究者、机构开始关注如何将多源异构数据转换为RDF三元组数据，并研发了多种映射工具、应用软件。本节将梳理3类RDF生成技术，介绍国际主流的10种RDF三元组转换工具，并从多个视角对其进行对比分析。

1.1 基于关系型数据库的RDB2RDF

目前，关系型数据库仍是存储管理结构化数据的主流方案，大量高质量、高可用和高价值的数据资源还存放于关系数据库。将这些数据转化为RDF三元组是释放这些数据潜在价值的重要方式。而将关系型数据库中的结构模式和数据转换成RDF三元组模型和RDF数据的过程被称为RDB2RDF。本节首先介绍W3C RDB2RDF工作组推荐的两种RDB2RDF映射语言，然后介绍4种RDB2RDF常见工具并进行对比分析。

1.1.1 映射语言

W3C的RDB2RDF工作组推荐了2种RDB2RDF映射语言Direct Mapping和R2RML，它们用来定义将关系数据库中的数据转换为RDF数据的各种流程和机制，包括URI的生成、RDF类和属性的定义、空节点的处理、数据之间的关联表达等[1]，输出的RDF数据具有一致的语义逻辑和标准化的数据格式。

Direct Mapping是将关系数据库的数据结构直接映射为RDF词表，关系数据库中的一个表转换为一个RDF类（class），一个字段转换为一个RDF属性（property），RDF词表中的术语名称与关系数据库中的表名和字段名保持一致，是一种本地本体。

R2RML具有高度的可定制性和灵活性，为RDB2RDF映射定义了系统性的逻辑框架，提出“逻辑表”的概念，将关系数据库中的一个表、视图，或是一个有效的SQL查询定义为“逻辑表”，突破了关系数据库表的物理结构限制，在生成RDF数据之前可以对RDB中的数据进行计算处理、筛选、清洗和整合，为不改变数据库原有的结构而灵活地按需生成RDF数据奠定了基础。

1.1.2 典型案例

利用第三方工具是RDB2RDF一个更加灵活的选择，本节将选取4个代表性工具介绍，并从以下维度进行对比分析。①映射语言：指工具支持W3C推荐的2种RDB2RDF映射语言Direct Mapping和R2RML，从而可以使转换流程、结果更加标准化。②配置便捷性：指工具是否需要人工配置实验操作环境，不需要人工配置或需要人工简单配置实验环境的工具具有更好的使用便捷性。③支持的数据输入格式：指工具可以支持的转换数据输入格式，如关系型数据库、JSON文件等。④支持的数据输出格式：指工具支持的多种RDF输出格式（如RDF/XML、N-Triples、Turtle等）或其他类型数据输出格式。⑤是否开源：开放源代码的工具更易于使用者学习，或根据具体需求在此基础上进行修改。

（1）D2RQ。D2RQ[8]平台是一个开源的以虚拟、只读RDF视图的形式访问关系数据库的系统，它通过Web访问将关系数据库的内容映射为RDF格式，生成虚拟、只读的RDF视图，最新版本是2012年发布的v0.8.1。D2RQ平台包括D2RQ映射语言、D2RQ引擎和D2RQ Server。D2RQ映射语言是将关系数据库模式映射到RDF词汇表和OWL本体的声明性映射语言，定义了映射规则，在2012年发布的v0.8.1版本中，开始支持W3C推荐的直接映射规范。D2RQ引擎建立在Jena接口上，

使用一个可定制的D2RQ映射文件，它可以在访问关系数据库时将RDF数据的查询语言SPARQL转换为RDB的查询语言SQL，并将SQL查询结果转换为RDF三元组或者SPARQL查询结果，从而将关系型数据库中的数据映射成虚拟的RDF格式，并将查询结果传递到平台框架的更高层。D2RQ Server是HTTP服务器，提供对RDF数据的查询访问接口，供上层的RDF浏览器、SPARQL查询客户端以及HTML浏览器调用。

（2）db2triples。db2triples[9]是一个开源jar工具包，于2011年发布了第一个版本，最新版本是2019年发布的db2triples-2.2。它的主要功能是从关系数据库中提取数据并将数据加载到RDF三元组中。db2triples是基于RDF数据处理框架Sesame和Apache Maven，需配置Maven来管理和编译：先对关系数据库中的关系数据表对应的实体和关系进行分析；其次根据R2RML映射语言开发ttl语义映射文件，同时配置相应的输入输出参数；最后在db2triples的RDB2RDF用户接口文件中调用R2RML映射器实现RDF实例化。此外，db2triples还支持直接映射规范，因此db2triples同时支持W3C推荐两种映射语言。

（3）R2RML Parser。R2RML Parser[10]是一个开源的基于R2RML映射文档，可以将关系数据库内容导出为RDF图的工具，于2013年发布了第一个版本，最新版本是2016年发布的0.8。其涵盖了R2RML的大部分功能，支持将Oracle、MySql、PostgreSql数据库中的数据转换为RDF格式，同时可以使用Apache Jena TDB将转换结果导出到数据库中，与sparql to sql转换器相比，该工具RDF的导出速度更快。

（4）Triplify。Triplify[11]是一个用于Web应用程序的轻量级PHP插件，它通过将关系数据库内容转换为RDF、JSON或关联数据来揭示关系数据库中编码的语义结构。Triplify基于特定Web应用程序中关系数据库的查询来检索信息并将查询的结果转换为RDF、JSON或关联数据，从而在Web上创建大量的语义表示。Triplify对操作环境的要求较低，便于用户安装使用，它可以使Web应用程序更易于处理，并为下一代基于语义的Web搜索奠定了基础。

1.1.3 对比分析

本节采用1.1.2节中提出的5个维度对这些工具进行比较分析，以为后续学者、机构在选择和使用RDB2RDF工具时提供参考。表1从7个方面比较了上述介绍的4种常见RDB2RDF工具。

在映射语言方面，R2RML Parser和Triplify有工具内部的转换映射逻辑，均不支持W3C推荐的映射语言规范Direct Mapping和R2RML。D2RQ可以支持Direct Mapping，而db2triples遵循了W3C推荐的映射语言规范，使得映射过程和输出的RDF数据格式更加标准化。

在配置便捷性方面，这4种RDB2RDF工具均需要人工进行操作环境的安装配置，但Triplify相较于其他3种工具而言，配置过程较简单，操作便捷性更好。

在支持的数据输入格式方面，各工具支持的数据库种类、数量各不相同。db2triples和R2RML Parser均支持Oracle、MySql和PostgreSQL这3种关系型数据库。Triplify支持常见关系型数据库。D2RQ则支持大部分主流关系型数据库，包括Oracle、MySQL、SQL Server、PostgreSQL、HSQLDB、Interbase/Firebird，其在应用场景上更具灵活性。

在支持的数据输出格式方面，Triplify支持RDF、JSON、关联数据的输出格式，在输出的数据格式多样性方面表现较弱。db2triples和R2RML Parser均支持Turtle、RDF/XML、Notation3、N-Triples这4种格式作为数据输出，输出格式的种类相对丰富，其中N-Triples格式适用于在大型数据库中的存储管理。D2RQ在此基础上还支持RDF/XML-ABBREV的输出格式，因此其在输出的数据格式多样性方面表现较好。

在是否开源方面，这4种RDB2RDF工具均开源，便于用户对源代码进行学习和修改。

这些工具可以将特定的关系型数据库直接转换成RDF数据，但存在难以与本体知识建模结果结合及映射，数据库多表间关联数据查看或导出效率低下，也难以同其他类型数据和处理流程进行融合，对于大规模海量异构数据映射以及新数据的增量映射支持较为困难[12]等问题或不足。

1.2 基于文件格式或在线的RDF转换器

除了关系型数据库，还有大量数据资源以多种文件格式存在，包含结构化或半结构化数据的文件如CSV、Excel、XML、JSON等。此外，还存在大量已经以不同RDF语法格式存在的关联数据文件，如施普林格·自然（Springer Nature）就将科技论文、作者、基金等数

表1 常见RDB2RDF工具对比表

	是否支持DM	是否支持R2RML	是否需人工配置	支持的数据库	支持的输出格式	是否开源
D2RQ	是	否	是	Oracle、MySQL、SQL Server、PostgreSQL、HSQLDB、Interbase/Firebird	Turtle、RDF/XML、RDF/XML-ABBREV、Notation3、N-Triples	是
db2triples	是	是	是	Oracle、MySQL、PostgreSQL	Turtle、RDF/XML、Notation3、N-Triples	是
R2RML Parser	否	否	是	Oracle、MySQL、PostgreSQL	Turtle、RDF/XML、Notation3、N-Triples	是
Triplify	否	否	是	常见关系数据库	RDF、JSON、关联数据	是

据转为RDF，并以JSON-LD格式发布为开放关联数据图谱SciGraph[13]。如果要将这些数据更好地融入语义Web以支撑更多智能应用，就需要支持多源数据处理、更加灵活的工具或服务将其向多种RDF格式数据进行转换。本节将选取5种RDF转换工具进行介绍，然后对其从是否需要人工配置、支持的数据类型、支持输出格式、是否开源等方面进行对比分析。

1.2.1 基于文件的RDF转换器

基于文件格式的RDF转换器通常需要人工通过命令行、编写代码等方式对源文件进行安装、配置，本节将介绍3种此类型的工具。

（1）Apache Any23。Anything To Triples（Any23）[14]提供从Web文档中提取RDF的Web服务，同时作为一个开源的命令行工具，可以提取和转换支持的数据格式，最新版本是2020年发布的2.5。Any23的逻辑模块是从Web检索原始数据，然后对收集的数据进行分析，确定数据编码和内容MIME类型，MIME类型的标识用于为之后的元数据抽取选择一个可激活的Extractor列表。一些影响Extractor运行的问题会影响Web上公开的大部分数据，Any23引入了可扩展的规则集合来检测、修复这些问题，并利用前一阶段激活的所有Extractor生成问题报告。然后Any23的模块将过滤Extractor生成的语句，以删除虚假、重复或不需要的三元组。最后使用模块提供的RDF编写器以RDF格式转换过滤的语句生成RDF。Any23支持多种主流的RDF格式作为数据的输入、输出，可灵活适用于多种RDF数据格式间的转换。

（2）ELMAR-to-GoodRelations。ELMAR-to-GoodRelations[15]是一个Python脚本，可以将ELMAR（Electronic Marketing List Information）的XML和CSV数据转换为用于语义Web的RDF/XML格式的GoodRelations电子商务数据。GoodRelations是一种电子商务的Web词汇表，可准确描述业务提供的内容，如产品及其功能、价格、商店、营业时间、付款选项等。在领域适应性方面，ELMAR-to-GoodRelations仅适用于电子商务领域的数据转换。

（3）PoolParty Extractor（PPX）。PPX[16]是一个面向企业的语义中间件平台，提供基于语义知识模型的精确文本挖掘算法来构建企业知识图谱。PPX可以创建并管理本体模型，自动分析大量文档和数据记录提取有意义的短语、命名实体类别或其他元数据并自动分类。它支持将结构化和非结构化信息源集成到一个独立的智能搜索索引中。不同的元数据模式可以映射到一个统一的知识模型，包括SKOS taxonomy和利用PoolParty叙词表管理系统创建的SKOS叙词表。PPX可与RDF图数据库集成，也可与市场上的搜索引擎集成，主要应用于内容推荐、精准语义搜索、智能文本标记、自动内容分类等场景。

1.2.2 RDF在线转换服务

在线的RDF转换服务免去了人工部署环境、配置文件花费的时间和精力，直接登录网页即可进行将数据格式向RDF三元组转换的操作，具有简单、直观、易操作的优点。但在线转换服务受限于网络环境和应用系统，因此其缺点是无法高效地转换处理大规模数据。本节将介绍2种此类型的工具。

（1）csv2rdf。csv2rdf[17]提供了REST Web服务接口，可以在线将带有表头的CSV文件转换为RDF格式。需要在线输入5种参数包括CSV文件的URL、转换后目标RDF文件的xmlbase URL、已转换属性的命名空间URL、用作主键的CSV文件中列的名称，以及输出的目标RDF文件格式，提交后即可快速完成文件转化。csv2rdf具有操作过程简单、无须部署即可完成文件转

化比较高效的优点。其支持的RDF输出格式为RDF/XML、N-Triple。

（2）RDF Translator。RDF Translator[18]是一个开源的由URI或直接文本输入触发的多格式在线转换器。它提供了从RDF/XML等RDF数据到RDFa等RDF数据或微数据的数据格式之间的转换功能。此外，它还提供了一个简单的REST API，允许输入URI后将原格式转化成目标格式，用HTML和CSS格式化转化结果，允许执行HTTP POST请求，并将数据附加到html网页中。RDF Translator支持RDFa、RDF/XML、Notation3、N-Triples、Microdata和JSON-LD格式间的相互转换。显然，RDF Translator比csv2rdf在支持的数据输入、输出格式方面更具适应性。

1.2.3 对比分析

本节采用1.1.2节中的4个维度对上文介绍的RDF转换器进行比较分析。表2是通过比较分析这5种RDF转换器得到的结果。

表2 常见RDF转换器对比表

工具名称	是否需人工配置	支持的数据输入格式	支持的输出格式	是否开源
Apache Any23	是	RDF/XML，Turtle，Notation3，RDFa，Microformats1，Microformats2，JSON-LD，HTML5 Microdata，CSV，都柏林、schema.org等词汇，YAML	Turtle、RDF/XML、N-Triples、JSON-LD	是
ELMAR-to-GoodRelations	是	XML、CSV	RDF/XML电子商务词汇表	是
PPX	是	不同的元数据模式	映射到统一的知识模型：SKOS taxonomy、SKOS叙词表	否
csv2rdf	否	带有表头的CSV	RDF/XML、N-Triple	否
RDF Translator	否	RDFa、RDF/XML、Notation3、N-Triples、Microdata、JSON-LD	RDFa、RDF/XML、Notation3、N-Triples、Microdata、JSON-LD	是

在配置便捷性方面，RDF在线转换服务无须人工配置，因此具有很好的配置便捷性。而基于文件格式的3种RDF转换器均需要在使用前进行文件和环境的配置，因此它们的配置便捷性较弱，使用门槛较高。

在支持的数据输入格式方面，这些转换器相比1.1.2节中介绍的RDB2RDF工具只支持关系型数据库外，还增加了支持结构化文件形式或RDF编码格式的数据类型。其中，csv2rdf仅支持CSV文件的输入，因此其支持的数据输入格式单一。ELMAR-to-GoodRelations在此基础上增加了支持XML文件的输入。PPX仅支持不同的元数据模式作为输入，因此其支持的数据输入格式也较为单一。RDF Translator支持RDFa、RDF/XML、N-Triples等主流RDF格式文件的输入，此外还支持Microdata和JSON-LD结构化文件的输入，因此其支持的数据输入格式较为丰富。Apache Any23支持的数据输入类型最为多样化，不仅支持RDF/XML、Turtle、Notation3和RDFa等RDF格式文件的输入，还可支持Microformats2、JSON-LD、HTML5 Microdata、CSV、都柏林词汇等结构化文件的输入。

在支持的输出格式方面，ELMAR-to-GoodRelations仅支持RDF/XML格式的输出，且只适用于电子商务领域，适用性较弱。PPX支持的输出格式为SKOS知识模型，因此其支持的输出格式也较为单一。csv2rdf支持RDF/XML和N-Triple两种RDF格式文件的输出，其支持的输出格式具有一定的丰富性。Apache Any23和RDF Translator支持RDF/XML、N-Triples等多种RDF格式文件和JSON-LD等结构化文件的数据输出，因此，它们支持的输出格式都具有很好的丰富性。

在是否开源方面，PPX和csv2rdf均不开源，Apache Any23、ELMAR-to-GoodRelations和RDF Translator均为开源RDF转换工具。

值得注意的是，这5种转换器支持的数据输入、输出格式有限且各不相同，这增加了工具/服务间互操作的难度，降低了研究效率。因此，未来需要研发出集成支持多种数据输入、输出格式的RDF转换平台，这样可

以免去工具/服务间的互操作，也可以满足更加多样的使用需求。

1.3 RDF-ETL工具体系Unifiedviews

RDF数据转换处理也是一项较为复杂的过程，通常由多个环节组成。上述基于关系型数据库和基于文件格式的RDF数据转换，主要是面向单一且具体的转换过程，用户难以开展较多的定制配置和交互操作。Unifiedviews[19]是一个开源的抽取-转换-加载（ETL）框架，允许用户定义、执行、监控、调试、调度和共享RDF数据处理任务，简化了关联数据发布过程的创建和维护。它提供了一个图形用户界面，数据处理任务在Unifiedviews中被表示为数据处理管道（pipeline）。每个pipeline由一个或多个数据处理单元（DPU）和这些单元之间的数据流组成，每个DPU封装处理数据的特定业务逻辑，并且可以产生对应的输出。在不同的pipeline中可以对DPU进行不同的配置。

数据单元是数据处理单元DPU使用或生成数据的容器。Unfiedviews支持3种类型的数据单元：RDF数据单元，处理RDF图；文件数据单元，处理文件和文件夹；关系数据单元，处理关系数据库中的表。有4种类型的数据处理单元DPU：Extractor是不定义任何输入数据单元的DPU，它的输入数据是由DPU的业务逻辑从外部来源获取的，如Extractor可以从远程SPARQL端点查询数据，或者从特定的URL集下载文件；Transformer是将输入转换为输出的DPU，它定义了输入和输出数据单元，如DPU将表格数据转换为RDF数据或执行SPARQL查询；Loader是定义输入数据单元但不定义任何输出数据单元的DPU，该DPU生成的输出数据不由Unifiedviews维护，而是由外部存储库进行存储；Quality Assessor是评估输入数据的质量并生成质量评估报告作为输出的DPU。此外，用户可以创建和部署自定义的DPU满足所需功能。

Unifiedviews是到目前为止，在流程和功能方面最为完整和全面的RDF数据创建与转换综合解决方案，对数据兼容性较高，支持关系型数据、RDF编码格式和文件数据向RDF数据的转换，但也存在构建复杂、人工配置困难的问题。从初步试用体验来看，该工具还是相对较为初级的原型系统，在易用性、稳定性和性能等方面与专业的数据ETL工具（如Kettle[20]）相比，还有较大提升空间。

2 大规模RDF图数据存储技术

随着大量RDF数据生成和开发应用需求日益强烈，大规模RDF的存储与管理问题受到广泛关注。为更好地管理RDF三元组数据，目前已发展出专门用于存储RDF数据的三元组数据库[21]。DB-Engines[22]是全球范围内收集和提供有关各类数据库管理系统及其排名的系统。通过分析该系统对存储RDF的数据库系统的排名发现，截至2020年12月，位于前列的分别是MarkLogic、Apache Jena、Virtuoso、GraphDB、Blazegraph、RDF4J。本节将介绍这6种三元组存储，并从以下6个维度对其进行对比分析。

①三元组规模：指三元组数据库支持存储的三元组数据规模，是选择三元组数据库对三元组数据进行存储、管理需要考虑的指标。②是否支持数据分布存储：三元组数据库支持数据的分布式存储可以提高系统的可靠性、可用性和存取效率，还有易于扩展的优点[23]，适用于对大规模RDF数据的存储管理。③存储的数据类型：指三元组数据库支持的存储RDF数据的语法格式。④查询语言：指三元组数据库支持的对数据和结构进行查询、处理的查询语言。⑤事务的ACID特性：指支持事务的数据库满足的原子性、一致性、隔离性、持久性。⑥是否开源：开放源代码的RDF数据库更易于使用者对此学习、改进。

2.1 三元组转换存储中间件

（1）Apache Jena。Apache Jena[24]是一个免费的开源Java框架，用于构建语义网和关联数据，于2000年发布了第一个版本，最新版本是2020年发布的3.15。该框架遵循了W3C标准，由各种相互作用的API组成，主要功能包括以RDF/XML、Turtle形式读写RDF、RDF存储、SPARQL查询处理、管理RDFS和OWL本体并进行推理等。其中，TDB是可对RDF基于内存或硬盘进行存储和查询的Jena组件，由Node表、索引、前缀表三部分构成，支持所有的Jena API。通过Jena中件间，可以与Virtuoso等数据库进行RDF读写和查询操作，也可基于SPARQL1.1协议进行各类操作。

（2）Eclipse RDF4J。Eclipse RDF4J[25]是一个用于处理RDF数据的开源Java框架，于2004年发布了第一个版本，最新版本是2020年发布的3.2.0。它的功能包括RDF数据的解析、存储、推理和查询，主要适用于中

小型数据集。RDF4J支持大部分主流的RDF文件格式，包括RDF/XML、Turtle、N-Triples、N-Quads、JSON-LD、TriG和TriX，提供内存和本机两种RDF存储机制，支持SPARQL1.1查询和更新语言。它提供的API可以访问MarkLogic、Virtuoso等RDF三元组数据库。

2.2 三元组存储数据库

（1）MarkLogic。MarkLogic[26]是一个具有JSON、XML、RDF等数据集成、存储、管理和搜索的企业级NoSQL商业数据库，于2001年发布了第一个版本，最新版本是2020年发布的10.0-4，主要客户有荷兰银行（ABN Amro）、索尼（SONY）、空中客车公司（AIRBUS）等。

MarkLogic是一个多模态数据库，它可以存储和查询文档、图形数据、关系型数据中的数据，使用三元组在文档之间建立关联。不需要预先定义模式，数据可以按照原本的格式在MarkLogic中进行存储，因此无须进行复杂的ETL过程，适用于数据集成；它内置了搜索引擎，可减少为标准查询构建、配置索引所需的时间和精力；它可以在商用硬件的集群中水平扩展到数百个节点、PB级数据、数十亿个文档，每秒可处理数万笔事务，集群随着数据或访问需求的增长或收缩而水平扩展，并提供自动故障转移、复制和备份服务。此外，MarkLogic具有多文档ACID事务处理能力且可以在云端运行，即在大规模事务处理应用程序中也可保证数据一致性。MarkLogic为存储和查询多个数据模型提供了一个统一的平台，具有灵活性、敏捷性、数据规模弹性可伸缩等优势。

（2）Virtuoso。Virtuoso[27]是一个多模RDBMS，用于管理以关系型表和RDF属性图表示的数据，于1998年发布了第一个版本，最新企业版本是2020年发布的8.3，而最新的开源社区版本是2018年发布的7.2。主要客户有零售行业的Gruppo Arena、金融行业的Eastern Corporate Federal Credit Union等。它具有数据虚拟化功能，其内置的推理功能使机器学习、数据治理成为数据虚拟化功能的一部分，从而可以为多个业务领域构建知识图谱。在Virtuoso中创建的知识图谱表现为关联数据语义网，其中基于属性的访问控制（ABAC）为跨企业和混合云扩展的智能数据访问策略提供了基础。

Virtuoso虽然是支持多种数据模型的混合数据库管理系统，但其基础源自开发了多年的传统关系型数据库管理系统，因此具备较为完善的事务管理、并发控制和完整性机制[18]。它把三元组直接存储在数据库表中，为了提高存储效率和查询效率，定义了额外的数据类型和表结构[28]，同时应用增强的映射规则，如可以使用W3C的R2RML等标准将RDBMS中表示的关系转换为细粒度的RDF语句，也可将SPARQL扩展为SQL语句使用。

Virtuoso提供了通过其网页客户端上传单个较小文件加载的功能，而在服务器后台可实现批量多个大文件的加载。以服务器后台为例，笔者加载了用Jena中间件转换的1 300万期刊论文及4 600多万论文作者核心数据项，共计136个RDF文件，容量40.8GB，共计有2.97亿RDF三元组，耗时2小时，平均38 485个三元组/秒；而加载Dbpedia数据共147个文件，容量700GB，共计有30亿RDF三元组，耗时16小时，平均51 574个三元组/秒。进行Sparql查询实验，从近40亿三元组中，查询单个实体对象耗时18.45毫秒。

（3）GraphDB。GraphDB[29]是基于OWL标准开发的企业级RDF三元组库，符合W3C标准，支持SPARQL1.1，于2000年发布了第一个版本，最新版本是2020年发布的9.4，主要客户案例有英国议会的数据服务、英国广播公司（BBC）2010年世界杯网站等。GraphDB有免费版、标准版、企业版3个版本，免费版支持多个图数据库的创建管理，可以解析CSV、XLS、JSON、XML等格式的结构化数据并生成RDF数据进行存储，还可基于RDF数据进行语义推理，其使用内置的基于规则的“前向链”（forward-chaining）推理机，由显式（explicit）知识经过推理得到导出（inferred）知识，对这些导出知识进行优化存储但仅限2个并发查询，导出的知识会在知识库更新后相应地同步更新[18]，同时它提供MongoDB集成，适用于大规模元数据管理。企业版在此基础上扩展了并发查询处理，允许查询吞吐量与集群节点的数量成比例地扩展，并可与Solr、Elasticsearch集成进行全文搜索。从版本8开始，GraphDB与RDF4J框架完全兼容。GraphDB实现了RDF4J的存储与推理层（SAIL），可以使用RDF4J的RDF模型、解析器和查询引擎直接访问GraphDB。

GraphDB提供了通过其网页客户端上传单个较小文件加载的功能，也可通过网络端加载事先已上传服务器特定目录的多个大文件，也提供了丰富的解析参数设置功能。笔者基于上述同样的期刊论文RDF数据，用

GraphDB免费版进行后台数据加载实验，加载2.97亿RDF三元组，耗时约6小时，而进行SPARQL单实例查询实验，响应时间大约为0.1秒。

（4）Blazegraph。Blazegraph[30]是一个基于RDF三元组库的图数据库系统，支持SPARQL1.1系列规范，于2006年发布了第一个版本，最新版本是2020年发布的2.1.6，主要客户有《财富》500强的EMC、Autodesk等。Blazegraph支持多租户架构，允许在多用户的环境下使用相同的系统且保证用户间数据的隔离。它可以部署为嵌入式数据库或独立服务器或高可用性集群或水平分片横向扩展的模式，在生命科学领域得到广泛应用。

2.3 对比分析

表3采用第2节中的6个维度比较了上述介绍的6种RDF存储体系。

表3 RDF三元组存储对比表

	三元组规模	数据分布存储	存储的数据类型	查询语言	事务的ACID特性	是否开源
Jena TDB	亿级	否	RDF/XML、Turtle	SPARQL	是	是
RDF4J	亿级	否	RDF/XML、Turtle、N-Triples、N-Quads、JSON-LD、TriG、TriX	SPARQL	是	是
MarkLogic	PB级	是	JSON、XML、RDF、地理位置、文本	JavaScript、XQuery、SPARQL、SQL	是	部分开源，有免费非商业版
Virtuoso	百亿级	是	XML、RDF、关系数据、文本	SPARQL、SQL	是	是
GraphDB	百亿级	否	CSV、XLS、JSON、XML、RDF	SPARQL、SeRQL	是	部分开源，有免费非商业版
Blazegraph	百亿级	是	RDF	SPARQL	是	是

在三元组规模方面，Jena TDB和RDF4J支持亿级的三元组数据存储，具有一定的存储规模。Virtuoso、GraphDB和Blazegraph均支持百亿级的数据存储，具有的存储规模更大。Marklogic则支持PB级的数据存储，可以更好地对大规模RDF数据进行存储应用。

在数据分布存储方面，Jena TDB、RDF4J和GraphDB采用传统的系统结构，不支持数据的分布式存储，因此这三者在处理大规模RDF数据时会面临性能受限的问题。MarkLogic、Virtuoso和Blazegraph均支持数据的分布式存储，在处理大规模RDF数据时有较好的性能和存取效率。

在存储的RDF数据类型方面，Marklogic、Virtuoso、GraphDB和Blazegraph均支持RDF数据和结构化文件的存储。Jena TDB支持RDF/XML和Turtle两种RDF文件格式的存储，具有一定的存储类型多样性。RDF4J支持RDF/XML、Turtle、N-Triples、N-Quads数据的存储，因此其支持存储的RDF数据类型较为丰富。

在查询语言方面，它们均支持SPARQL标准化查询语言。

在事务的ACID特性方面，它们均支持事务的ACID特性从而可以保障数据的正确性。

在是否开源方面，它们均有开源版本，便于使用者对源代码进行学习、改进。

值得注意的是，上述RDF数据库系统支持存储的数据类型各不相同，导致用户可能需要针对不同的数据类型使用不同的数据库系统进行存储管理，这增加了操作的难度。因此，未来的数据库系统需要支持更加多元化的存储格式和不同数据库间的数据迁移功能，这可以更便于异源异构数据的整合和统一管理。

3 大规模RDF一体化治理场景设计

目前，大规模RDF数据转换及三元组数据存储的理论、方法、技术、系统处于快速发展和逐渐完善的阶段，学界和工业界对RDF数据管理的研发投入正在不断增加。对比分析发现，上述多个工具系统在支撑海量RDF数据转换与管理方面还存在诸多不足，如RDF三元组转换工具的安装、映射配置较为复杂，对使用用户计算机技术能力的要求较高；转换工具很少支持直接与数据库进行对接，支持的数据库或文件格式各有侧重、种类较为单一、缺乏兼容性，导致用户可能需要同时使用多个平台、分多个步骤完成RDF数据的转换、存

储。三元组存储方面，多数RDF存储系统支撑的数据规模和体量有限，导致在并发处理大规模海量数据时性能下降或无法支撑。

ETL技术是将源数据抽取（extract）、转换（transform）和加载（load）至目标端这一过程的描述。ETL从源数据中抽取所需数据，按照预先设定好的映射规则将抽取的数据进行转换、清洗、加工得到统一的数据格式，最后将转换好的数据按照规则增量式或全部导入目标数据库中[31]，是在大数据治理环境下对多源异构数据进行集成管理的重要手段。大规模RDF数据的转换、存储需要充分利用ETL技术实现处理过程的流程化管理，提升处理效率。总体而言，大部分工具系统还无法有效融合大数据治理的全生命周期，如数据的采集、转换、加载、存储等，因此还需要建立更普遍适用的RDF-ETL工作流程和数据转换引擎。Unifiedviews虽已朝着RDF的ETL整体流程努力，但目前的功能、性能处于初级和起步阶段，还无法满足大模型数据治理需求。

针对以上不足，笔者认为还需要从工程级、零编码、动态配置、百亿级、高并行和用户友好等方面，进一步突破大规模RDF三元组动态抽取、映射转换和加载存储等共性技术，将RDF数据转换与存储管理进一步融入整个大数据治理技术体系中来，形成流程化和集成化的RDF-ETL解决方案，并重点解决和支持4类RDF-ETL的应用场景。

（1）从非RDF数据（如Excel、CSV、XML、JSON文件、关系数据库等）到RDF数据的转换。需要为用户提供可视化、可交互的友好操作界面和流程，支持普通用户在不学习编程或较少地掌握相关技术前提下，简便、高效地实现来源数据格式向RDF主、谓、宾三元组的映射配置。

（2）不同RDF语法格式之间的双向转换。RDF目前有Turtle、RDF/XML、N-Triples、NQuads、JSON-LD、RDF/JSON、TriG、TriX和RDF Binary等主流格式。由于不同的RDF数据格式具有不同的特性，用户在面对不同的需求和应用场景时，需要在这些格式之间进行任意互转互换。

（3）RDF三元组在不同数据库中数据迁移。如将Virtuoso、GraphDB或SPARQL查询终端等来源的RDF数据进行导出、转换，或在不同数据库之间进行RDF数据的迁移。这可以降低跨项目、跨领域的迁移成本，便于大规模RDF的数据组织和管理，从而实现对海量、异源、异构数据的深度整合。

（4）RDF数据的增量更新和进化管理。需要基于SPARQL Update1.1协议，按需要开展整体或局部数据的增量更新和进化维护，对存储在图数据库中的RDF数据进行动态管理，包括从图数据中新增、删除、修改RDF三元组实例数据（Graph Update），以图为单元整体拷贝、删除、移动RDF数据（Graph Management）等。

当前，主流的ETL工具有Informatica、Datastage、Kettle、Sqoop和SSIS等。其中，Kettle因其是基于Java和插件开发的开源工具，平台自身提供开放接口，开发者按照规范实现接口就能进行插件开发，是当前最受欢迎的开源ETL解决方案。因此，为充分发挥Kettle在大数据治理生态体系中的技术成熟度和稳定性，以及被社区广泛使用和支持的优势，在实现RDF-ETL数据治理一体化方面，笔者放弃了类似Unifiedviews重新搭建完整技术体系的方案，而是选择基于Kettle生态体系及其插件开发接口，集成RDF语义中间件Apache Jena、Eclipse RDF4J接口，Virtuoso、GraphDB等图数据库驱动，以及SPARQL Query和SPARQL Update标准协议，开发支撑上述4种场景的RDF插件，从而以最小代价将大规模RDF三元组的数据治理有机地融合到现有常规大数据治理环境和ETL技术流程之中。

4 总结与展望

随着人工智能、大数据技术的普及与发展，面向通用及行业的智能化解决方案应运而生。RDF通过主语（Subject）、谓语（Predicate）和宾语（Object）三元组描述互联网资源之间的语义关系，揭示资源中蕴含的知识，构建知识间的关联，是构建语义知识图谱、实现认知智能的重要基础。本文介绍了现有的大规模RDF生成及存储管理的技术方法，选择了国际主流的10种RDF三元组转换工具，以及6种广受欢迎的RDF存储系统，从技术原理、性能特点及应用场景等多个视角进行了对比分析，同时总结了各类工具系统在支撑海量RDF数据转换与管理方面存在的不足，并从4个方面对大规模RDF一体化治理场景进行了设计：从非RDF数据到RDF数据的转换；不同RDF数据格式之间的双向转换；RDF三元组在数据库之间的数据迁移；RDF数据的动态更新和进化管理。目前笔者所在团队正在围绕此目标，基于知名的企业级开源ETL工具Kettle，研发

支撑上述4类典型场景的RDF转换处理插件，将在下一阶段完成相关技术的实现与工程化应用研究。

参考文献

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

[2] W3CRDFWorking Group. RDF 1.1 Concepts and Abstract Syntax [EB/OL]. [2020-10-18]. http://www.w3.org/TR/2014/REC-rdf11-concepts-20140225.

[3] 知识图谱发展报告（2018）第一章知识表示与建模 [C] //中国中文信息学会，语言与知识计算专委会. 知识图谱发展报告（2018）. 中国中文信息学会，2018：11.

[4] WILKINSON M D，DUMONTIER M，AALBERSBERG I J J，et al. The FAIR Guiding Principles for scientific data management and stewardship [J]. Scientific data，2016，3（1）：1-9.

[5] 王勇超，罗胜文，杨英宝，等. 知识图谱可视化综述 [J]. 计算机辅助设计与图形学学报，2019，31（10）：1666-1676.

[6] 联合国粮食及农业组织. AGROVOC [EB/OL]. [2020-10-18]. http://www.fao.org/agrovoc/zh.

[7] 贾君枝，梅玥. BBC关联数据实现研究 [J]. 数字图书馆论坛，2020（9）：18-23.

[8] D2RQ [EB/OL]. [2020-10-18]. http://d2rq.org.

[9] Antidot. db2triples [EB/OL]. [2020-10-18]. https://github.com/antidot/db2triples.

[10] Hellenic Academic Libraries Link/National Technical University of Athens，National Documentation Centre/National Hellenic ResearchFoundation. r2rml-parser [EB/OL]. [2020-10-18]. http://github.com/nkons/r2rml-parser.

[11] University of Leipzig. Triplify [EB/OL]. [2020-10-18]. http://triplify.org.

[12] 王昊奋，丁军，胡芳槐，等. 大规模企业级知识图谱实践综述 [J]. 计算机工程，2020，46（7）：1-13.

[13] Springer Nature. SN SciGraph research repository [EB/OL]. [2020-10-24]. https://sn-scigraph.figshare.com/.

[14] Apache Any23 [EB/OL]. [2020-11-08]. http://any23.apache.org/.

[15] Universität der Bundeswehr München. ELMAR-to-GoodRelations [EB/OL]. [2020-10-18]. http://code.google.com/p/elmar-to-goodrelations/.

[16] PoolParty Extractor（PPX）[EB/OL]. [2020-10-18]. http://poolparty.biz/.

[17] csv2rdf Web Service [EB/OL]. [2020-10-08]. https://data-gov.tw.rpi.edu//ws/csv2rdf.html.

[18] Universität der Bundeswehr München. RDF Translator [EB/OL]. [2020-10-18]. http://rdf-translator.appspot.com/.

[19] KNAP T，HANEČÁK P，KLÍMEK J，et al. UnifiedViews：an ETL tool for RDF data management [J]. Semantic Web，2018，9（5）：661-676.

[20] Pentaho group. Data Integration-Kettle [EB/OL]. [2020-10-02]. https://community.hitachivantara.com/s/article/data-integration-kettle.

[21] 王鑫，邹磊，王朝坤，等. 知识图谱数据管理研究综述 [J]. 软件学报，2019，30（7）：2139-2174.

[22] DB-Engines Ranking of RDF Stores [EB/OL]. [2020-12-06]. https://db-engines.com/en/ranking/rdf+store.

[23] 胡文波，徐造林. 分布式存储方案的设计与研究 [J]. 计算机技术与发展，2010，20（4）：65-68.

[24] Apache Jena [EB/OL]. [2020-11-01]. https://jena.apache.org/.

[25] Eclipse RDF4J [EB/OL]. [2020-11-01]. https://rdf4j.org/.

[26] MarkLogic Corp. MarkLogic [EB/OL]. [2020-11-01]. www.marklogic.com.

[27] OpenLink Software. OpenLink Virtuoso [EB/OL]. [2020-11-01]. https://virtuoso.openlinksw.com/.

[28] 肖佳，肖诗斌，王洪俊. 海量RDF数据存储查询研究 [J]. 北京信息科技大学学报（自然科学版），2017，32（3）：63-69.

[29] Ontotext. GraphDB [EB/OL]. [2020-11-08]. http://graphdb.ontotext.com/.

[30] Blazegraph by Systap，LLC. Blazegraph [EB/OL]. [2020-11-03]. https://www.blazegraph.com/.

[31] 张怡敏，卜佳，李杨梅，等. ETL技术在船舶制造海量异构数据处理中的应用 [J]. 造船技术，2020（5）：77-82.

作者简介

李悦，女，1996年生，硕士研究生，研究方向：知识组织、知识图谱。

孙坦，男，1970年生，博士，研究员，研究方向：数字信息描述与组织。

赵瑞雪，女，1968年生，博士，研究员，研究方向：信息管理与信息系统、数字图书馆、知识组织与知识服务。

李娇，女，1989年生，博士研究生，助理研究员，研究方向：知识组织、知识图谱。

黄永文，女，1975年生，博士，副研究馆员，研究方向：知识组织与知识服务。

罗婷婷，女，1985年生，硕士，助理研究员，研究方向：知识组织、大数据融汇治理、信息管理与信息系统。

鲜国建，男，1982年生，博士，研究员，通信作者，研究方向：大数据融汇治理、知识组织、知识图谱，E-mail：xianguojian@caas.cn。

A Comparative Study of Large-scale RDF Triple Conversion and Storage Tools

LI Yue[1] SUN Tan[1,2] ZHAO RuiXue[1,2] LI Jiao[1] HUANG YongWen[1] LUO TingTing[1] XIAN GuoJian[1,2]
(1. Agricultural Information Institute of CAAS, Beijing 100081; 2. Key Laboratory of Agricultural Big Data, Ministry of Agriculture and Rural Affairs, Beijing 100081)

Abstract: Data network rich in semantic knowledge is the cornerstone of realizing big data intelligence. Resource Description Framework (RDF) is the W3C standard for describing web resources. Large-scale conversion and storage management of RDF triples is an important path for building a linked data network or semantic knowledge graph and realizing data Findable, Accessible, Interoperable and Reusable (FAIR principle). In this paper, ten international mainstream RDF conversion tools and six popular RDF triple storage systems are selected, and a comparative analysis is made from the perspectives of technical principles, performance characteristics and application scenarios, and briefly summarize the existing problems and shortcomings. It is proposed that the goal of large-scale RDF triple data conversion and storage management is to realize the flow, integration and integration of RDF Extract-Transform-Load (ETL), and to focus on supporting four typical application requirements scenarios, including: conversion from non-RDF data to RDF data; bidirectional conversion between different RDF data formats; data migration of RDF triples between databases; dynamic update and evolutionary management of RDF data.

Keywords: RDB2RDF; RDF Conversion; RDF Triple Store; Big Data Intelligence; Knowledge Graph

（收稿日期：2020-11-04）

书 讯

《汉语主题词表》

《汉语主题词表》自1980年问世以后，经1991年进行自然科学版修订，在我国图书情报界发挥了应有作用，曾经获得国家科学技术进步二等奖。为适应网络环境下知识组织与数据处理的需要，由中国科学技术信息研究所主持，并联合全国图书情报界相关机构，自2009年开始进行重新编制工作，拟分为工程技术卷、自然科学卷、生命科学卷、社会科学卷四大部分逐步完成。目前工程技术卷和自然科学卷已出版。

《汉语主题词表（工程技术卷）》共收录优选词19.6万条，非优选词16.4万条，等同率0.84，在体系结构、词汇术语、词间关系等方面进行了改进创新。《汉语主题词表（自然科学卷）》共收录专业术语12.4万条，包含数学、物理学、化学、天文学、测绘学、地球物理学、大气科学、地质学、海洋学、自然地理学等学科领域，收词系统、完整，语义关系丰富、严谨，每条词汇都有相应的学科分类号表现其专业属性，并与同义英文术语对应。同时，建立《汉语主题词表》网络服务系统，提供术语查询、文本主题分析、知识树辅助构建等服务。《汉语主题词表》可用于汉语文本分词、主题标引、语义关联、学科分类、知识导航和数据挖掘，是文本信息处理及检索系统开发人员不可或缺的工具。

《汉语主题词表（工程技术卷）》已于2014年由科学技术文献出版社出版，分为13个分册，总定价3 880元。《汉语主题词表（自然科学卷）》已于2018年由科学技术文献出版社出版，分为5个分册，总定价1 247元。两卷均可分册购买。