

Efficient RDF Schema Mapping and Triples Generation Based on ETL Tool

Jiao Li, Guojian Xian
Agricultural Information Institute of CAAS

Current methods to generate RDF(Resourse Description Framework) data

1. RDF data extraction from Relational Database (RDB)

- mainstream, RDB-to-RDF/RDB2RDF

2. other format (CSV, Excel, JSON and XML files) to RDF

The screenshot shows the Semantic Web Wiki page for 'Category:RDF Generator'. The page title is 'Category:RDF Generator' and it is a subcategory of 'Tool'. The page contains a list of tools used to extract RDF data, categorized by letter. The tools listed include: A (Alchemy, Any23, Asio), B (BMEcat2GoodRelations), C (Cvs2rdf, Cypher), D (D2RQ, Db2triples, DBpedia Spotlight, DSP Platform), E (ELMAR2GoodRelations, Extractiv), F (FOAF-o-matic, FRED), G (GoodRelations for Joomla, GoodRelations Snippet Generator, GoogleProductFeedConverter), H (H2RDF), I (Inqle, InstantRDF for Umbraco, ITM), J (Java-rdfa, JsonLD), K (KIM Platform, Krestor), L (L2RDF), M (MSemantic), N (Ntriples), O (Ontos API, Open Calais, OpenLink Virtuoso), P (PCS2OWL, PHP POWDER Processor, PoolParty Extractor, POWDER Processor, PyRdfa), Q (Q2RDF), R (R2RM, Parser), S (SCF, SKOS2OWL, SPARQL-RW, SPARQL2XQuery), T (TransOnto, Triplify), U (Ultrawrap, URIBurner), V (V2RDF), W (Wikimeta, WPEC), X (X2RDF), Y (Y2RDF), Z (Zemanta).

https://www.w3.org/2001/sw/wiki/Category:RDF_Generator

Current methods to RDB-to-RDF

- **Ontology matching:** Concepts and relations are extracted from relational schema or data by using data mining, and then mapped to a temporal established ontology or specific database schema.
- **Mapping Language:** This involves cases of low similarity between database and target RDF graph, as exemplified by R2RML, which enables users express the desired transformation by following chosen structure or vocabulary.
- **Query Engine-based:** Transformation process is based on the SPARQL query of search engines with capability in supporting large collection of concurrent queries

General Tools for RDB2RDF

Tool	Description	Input	Output Format
D2RQ	<p>a system for accessing relational databases as virtual, read-only RDF graphs. It offers RDF-based access to the content of relational databases without having to replicate it into an RDF store. Using D2RQ you can:</p> <ul style="list-style-type: none">•query a non-RDF database using SPARQL•access the content of the database as Linked Data over the Web•create custom dumps of the database in RDF formats for loading into an RDF store•access information in a non-RDF database using the Apache Jena API	Oracle MySQL PostgreSQL SQL Server HSQLDB Interbase/Firebird	RDF
Triplify	<p>a small PHP plugin for Web applications, which reveals the semantic structures encoded in relational databases by making database content available as RDF, JSON or Linked Data</p>	Relational Database	RDF JSON Linked data
R2RML Parser	<p>export relational database contents as RDF graphs, based on an R2RML mapping document. Contains an R2RML mapping document for the DSpace institutional repository solution</p>	Relational Database MySQL PostgreSQL Oracle	Turtle N-Triples RDF/XML Notations3

But, these tools can not fully included:

- support most non-RDF data formats and output formats
- offer a packaged and multifunctional RDF data process method without programming
- integrated use with the triple stores

So we tried to:

- merge RDF generation with ETL(Extract-Transform-Load)
- redevelop the prominent ETL tool to an RDF ETL framework in a semantic-based way
- provide a user-friendly, open to use and intuitive interface

Our solution for RDF generation and management

RDF ETL plugin : RDFZier

New developed plugin:

- based on Kettle (a leading open-source ETL application on the market) in an ETL environment
- RDF 4J
- support multiple mainstream non-RDF format inputs AND ETL of multi-source heterogeneous data
- offer one-stop templates without coding
- efficient paralleling process that can provide multithreaded operations
- store multiple types of outputs into a selected RDF endpoint (triple store) or file system

General View

The screenshot shows the Spoon IDE window titled "Spoon - rdf-output-test (changed)". The interface includes a menu bar (File, Edit, View, Action, Tools, Help, Neo4j), a toolbar, and a "Connect" button. On the left, a "View" pane shows a tree structure of "Transformations" with "rdf-output-test" expanded to show "Input-JournalArticle" and "Output-RDF data". The main workspace displays a "Transformation diagram" with a flow from "Input-JournalArticle" to "Output-RDF data" via an "RDFizer" component. A red bracket at the bottom left labels the left pane as "Component", and a red bracket at the bottom center labels the main workspace as "Transformation diagram".

The screenshot shows the "Table input" dialog box. It has a "Step name" field set to "Input-JournalArticle" and a "Connection" dropdown set to "sqlserver-217-AgriNSTL". Below these are buttons for "Edit...", "New...", and "Wizard...", and a "Get SQL select statement..." button. The main area contains an SQL query:

```
SELECT  
  . paper_id  
  . journal_id  
  . identifier  
  . title  
  . alternative  
  . keywords  
  . abstract  
  . classification  
  . start_page  
  . end_page  
  . total_page_number  
  . local_doi  
  . paper_type  
  . issue_id  
  . guid  
  . doi  
  . isSCI  
  . journal_name  
  . issn  
  . "year"  
  . volume  
  . issue  
  . publisher  
FROM article
```

A red bracket at the bottom right labels the dialog as "Input detail".

query the chosen field information with SQL language

Format supported

Input:

- Relational database (MySQL, SqlServer), NoSQL, Data Stream/Text file (csv, Excel, json, XML)...

Output format:

- Turtle, JSON-LD, N-triples, RDF/XML, NQuads, TriG, RDF/JSON, TriX, RDF Binary

Parameters defined in RDFZier

The screenshot shows the 'RDFZier Output' dialog box with the following configuration:

- Step name: Output-RDF data
- NameSpaces: Mapping Setting, Dataset Metadata, Output
- Subject URI: http://linked.aginfra.cn/scikg/journal_article/{sid}
- Class Types: <http://linked.aginfra.cn/onts/scikg#journalArticle>
- uniqueKey: paper_id

#	Stream Field	Predicates	Object URIs	Multi-Values Seperator	DataType	Lang Tag
1	title	dc:title				zh-CN
2	LANGUAGE	dc:language				
3	journal_name	skos:label				
4	year	dc:year				
5	paper_id					
6	doi	bibo:doi	http://doi.org/{oid}			
7	alternative	dcterms:alternative				
8	issue	bibo:issue				
9	volume	bibo:volume				
10	start_page	bibo:pageStart				
11	end_page	bibo:pageEnd				
12	abstract	dcterms:abstract				
13	abstract_alternative	scikg:abstractAlternative				
14	keywords	prism:keyword				
15	classification	dc:subject				
16	journal_id	schema:isPartOf	http://linked.aginfra.cn/scikg/journal/{oid}			

Buttons: Get Fields, OK, Cancel, Help

Parameter		Description
Namespace	Prefix	collections of names identified by URI references
	Namespace	different prefixes depending on the required namespaces
Mapping Setting	Subject URI	HTTPURI template for the Subject/Resource, a placeholder {sid} would be used and replaced by UniqueKey
	Class Types	the classes to which the resource belongs, supporting multi-class types(split by semicolon), such as skos:Concepts; foaf:Person
	UniqueKey	the unique and stable primary key of resource, part of the Subject URI
	Fields Mapping Parameters	a list of field map from selected data source to target RDF schema, including the input Stream Field, Predicates, Object URIs, Multi-Values Seperator, Data Type, Lang Tag
Dataset Metadata	Meta Subject URI	URI pattern of generated dataset
	Meta Class Types	the classes to which the resource belongs
	Parameters	a list of descriptions of generated dataset, including PropertyType, Predicates, Object Values, DataType, Lang Tag
Output Setting	File system setting	option for file system storage, including Filename and RDF format
	RDF store setting	option for RDF store, including triple store name, server URL, Repository ID, Username (if any), Password, Graph URI

Output setting

The screenshot shows the 'RDFizer Output' dialog box with the 'Output' tab selected. The 'Step name' is 'Output-RDF data'. Under 'File Setting', 'Save To File' is unchecked, 'Filename' is 'H:\rdf-output-test\casdd-0.rdf', and 'RDF format' is 'RDF/XML'. Under 'Store Setting', 'Save To Store' is checked, 'Triple Store' is 'Virtuoso', 'Server URL' is 'jdbc:virtuoso://10.200.32.162:1111', 'Database/RepositoryID/NameSpace' is empty, 'Username' is 'dba', 'Password' is 'dba', and 'Graph URI' is 'http://knowledgcenter.com/Agri'. Buttons for 'Help', 'OK', and 'Cancel' are at the bottom.

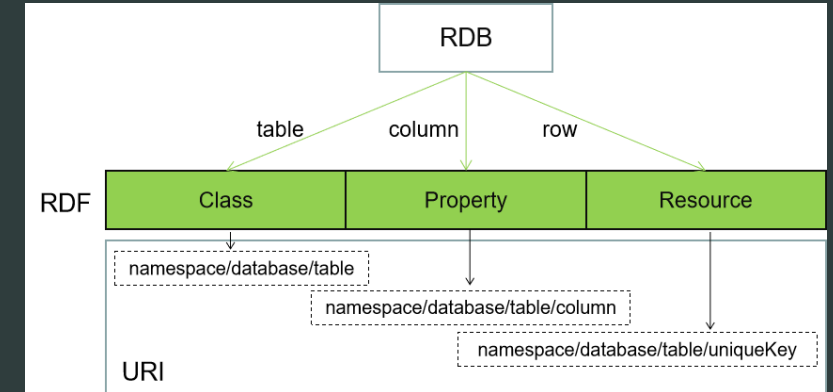
Save to File: local system

Save to Store:

- virtuoso
- GraphDB
- Blazegraph
- MarkLogic

Example of use

- one-stop RDF generation from RDB
- direct mapping
- field mapping rules or a semantic schema is must



doc_id	paper_id	journal_id	identifier	timestamp	setSpec	title
JJ0210397746	H.19406373	N2008EPST0009826	19D088E2-C07E-8F38-B7E	2011-07-12T15:25:30Z	eng	Age-Dating of Slope Failures on The Sigsbee Escarpment
JJ021965784	11054729	N2008EPST0008789	11054729	2011-10-26T20:47:25Z	eng	Survey of the chemical composition of 57
JJ0237424618	J20160923000637	N2007EPST0002797	J20160923000637	2016-10-30T18:56:35Z	eng	Synthesis and Characterization of Estolide
JJ0213842061	H.21531511	N2008EPST0014315	30217BA0-7C8A-831C-DC	2011-07-14T22:33:32Z	eng	The Origin and Impact of CPG New-Produ
JJ024919661	H.13001619	N2008EPST0001794	A07E8DF5-E206-676B-A5	2011-07-13T07:28:10Z	eng	AGP buys direct to meet export rules
JJ026484948	H.22246122	N2008EPST0004944	4BCE065A-6A2C-B06B-25	2011-07-14T19:58:31Z	eng	How materialism affects environmental be
JJ0240876314	J20180427021374	N2008EPST0000879	J20180427021374	2018-04-27T19:05:39Z	eng	Temperature extremes in the Argentina ce
fe946c99ee426	J20200617004291	N2013EPST0000116	J20200617004291	2020-07-28T12:50:03Z	eng	S,N-Codoped oil-soluble fluorescent carb
JJ0225392226	J20120330010674	N2008EPST0007451	J20120330010674	2012-03-30T18:44:50Z	eng	Risk analysis and its link with standards o
JJ0224813517	J20120117000767	N2008EPST0000136	J20120117000767	2012-03-28T16:04:36Z	eng	Molecular insights into miRNA processing
JJ025553750	H.13667707	N2007EPST0002317	286CD477-FDA2-3AA6-A8	2011-07-12T10:06:52Z	eng	Correlated motions in native proteins fro
JJ0235764356	J20160116011392	N2007EPST0000937	J20160116011392	2016-01-22T19:33:52Z	eng	Partial cross-enhancement in models for c
JJ0214565821	H.18680236	N2008EPST0012585	D88DC482-D6C7-0550-E0	2011-07-12T10:29:00Z	eng	Fiabilisation des index: UMOTEST PEAUFH
JJ025602681	H.22324799	N2007EPST0000935	912F890D-8B52-B65F-777	2011-07-14T19:06:20Z	eng	The sea of uncertainty surrounding ductal
JJ0218802247	4823889	N2008EPST0004833	4823889	2011-11-12T00:59:17Z	eng	Fort Rucker
JJ029696389	H.20372101	N2007EPST0003245	2C64A524-DBDB-7219-A5	2011-07-12T14:32:29Z	eng	Beyond the Benzene Dimer:An Investigati
JJ0218695544	4689968	N2007EPST0003028	4689968	2011-11-11T23:30:08Z	eng	Financial Surveillance
JJ0240393814	J20180119028581	N2007EPST0001730	J20180119028581	2018-01-19T19:01:20Z	eng	Validation of a Single-Gene Next-Generat
JJ0212613309	H.12135856	N2007EPST0002845	0B49CD41-455F-1AEA-0D	2011-07-11T20:11:06Z	eng	Ballistic magnetoresistance in small-size c
JJ023507084	H.13261966	N2007EPST0002800	3D8D9B7A-97DE-10B8-1C	2011-07-12T01:36:44Z	eng	Molecular phylogeny of the green algal o
JJ0217843951	4254030	N2008EPST0008080	4254030	2011-03-05T14:37:02Z	eng	Relations between abnormal transmission



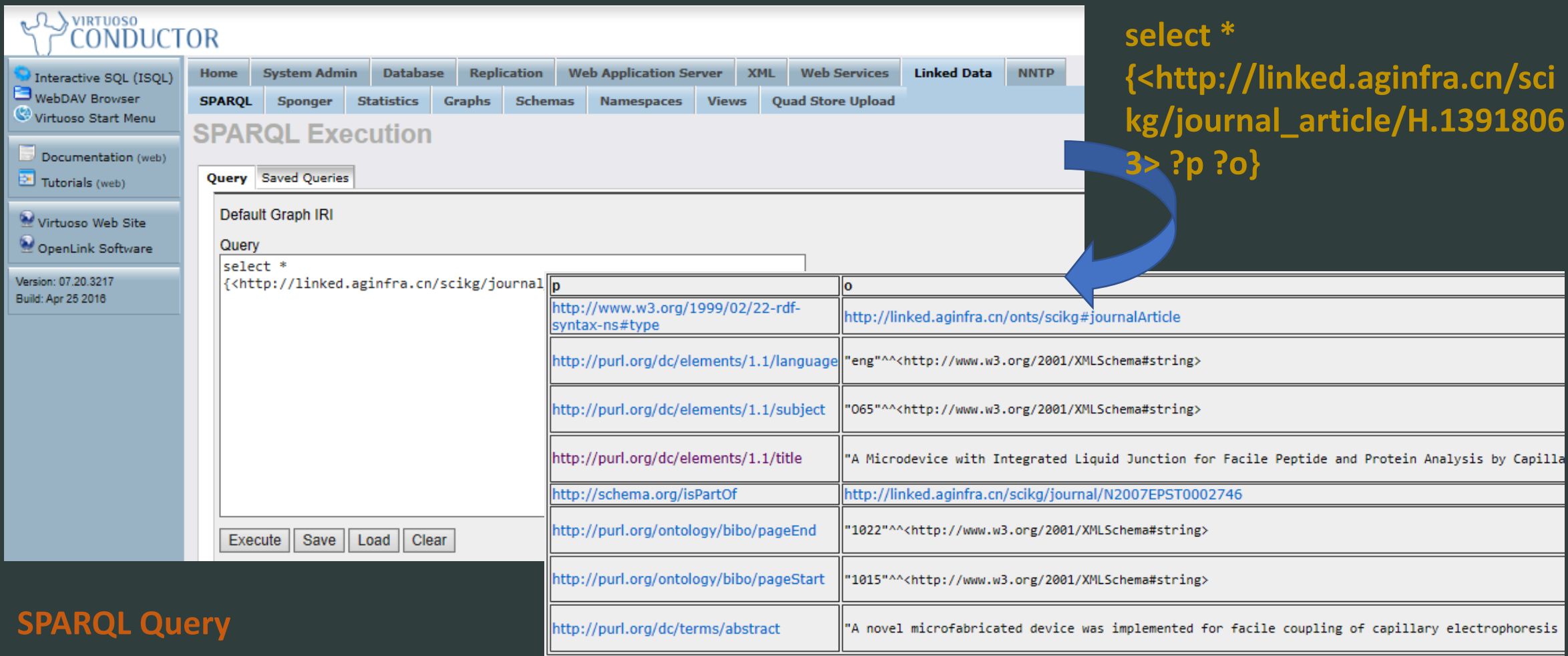
```

1 <?xml version="1.0" encoding="UTF-8"?>
2 <rdf:RDF
3   xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
4   xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema/"
5   xmlns:xsd="http://www.w3.org/2001/XMLSchema#"
6   xmlns:owl="http://www.w3.org/2002/07/owl#"
7   xmlns:dc="http://purl.org/dc/elements/1.1/"
8   xmlns:skos="http://www.w3.org/2004/02/skos/core#"
9   xmlns:dcterms="http://purl.org/dc/terms/"
10  xmlns:npg="http://ns.nature.com/terms/"
11  xmlns:bibo="http://purl.org/ontology/bibo/"
12  xmlns:scikg="http://linked.aginfra.cn/onts/scikg#"
13  xmlns:prism="http://prismstandard.org/namespaces/basic/3.0/"
14  xmlns:schema="http://schema.org/">
15  ...
16 <rdf:Description rdf:about="http://linked.aginfra.cn/scikg/geneExpression">
17   <rdf:type rdf:resource="http://ns.nature.com/terms/Journal"/>
18   <dc:title>NSTL英文期刊论文</dc:title>
19   <owl:sameAs rdf:resource="http://linked.aginfra.cn/scikg/journalArticles/001"/>
20   <rdf:seeAlso rdf:resource="http://linked.aginfra.cn/scikg/journalArticles/001"/>
21 </rdf:Description>
22  ...
23 <rdf:Description rdf:about="http://linked.aginfra.cn/scikg/journal_article/H.19406373">
24   <rdf:type rdf:resource="http://linked.aginfra.cn/onts/scikg#journalArticle"/>
25   <dc:title xml:lang="zh-CN">Age-Dating of Slope Failures on The Sigsbee Escarpment</dc:title>
26   <dc:language>eng</dc:language>
27   <bibo:pageStart>30,32-35</bibo:pageStart>
28   <bibo:pageEnd>0</bibo:pageEnd>
29   <dcterms:abstract>A large number of slope failures have occurred in the geologic past along the Sigsbee
30   <dc:subject>P7</dc:subject>
  
```

SqlServer

RDF--Local File System

Triple store--Virtuoso



The screenshot shows the Virtuoso Conductor web interface. The main area is titled "SPARQL Execution" and contains a query editor and a results table. A blue arrow points from the SPARQL query text to the results table.

SPARQL Query:

```
select *
{<http://linked.aginfra.cn/scikg/journal
```

Results Table:

p	o
http://www.w3.org/1999/02/22-rdf-syntax-ns#type	http://linked.aginfra.cn/onts/scikg#journalArticle
http://purl.org/dc/elements/1.1/language	"eng"^^<http://www.w3.org/2001/XMLSchema#string>
http://purl.org/dc/elements/1.1/subject	"065"^^<http://www.w3.org/2001/XMLSchema#string>
http://purl.org/dc/elements/1.1/title	"A Microdevice with Integrated Liquid Junction for Facile Peptide and Protein Analysis by Capilla
http://schema.org/isPartOf	http://linked.aginfra.cn/scikg/journal/N2007EPST0002746
http://purl.org/ontology/bibo/pageEnd	"1022"^^<http://www.w3.org/2001/XMLSchema#string>
http://purl.org/ontology/bibo/pageStart	"1015"^^<http://www.w3.org/2001/XMLSchema#string>
http://purl.org/dc/terms/abstract	"A novel microfabricated device was implemented for facile coupling of capillary electrophoresis

SPARQL Query

Future View

- Multi-format Data Conversion and Loading (between different serialization formats or Endpoints)
- Remote RDF Data Migration
- RDF Graph Update (by using SPARQL 1.1 update)

Thank you!

Questions/Comments?

lijiao@caas.cn

xianguojian@caas.cn