RapidMiner
Linked Open Data Extension

Manual

Version 1.0, 09/13/13

Heiko Paulheim

Evgeny Mitichkin

Petar Ristoski

Christian Bizer

Contact: heiko@informatik.uni-mannheim.de

University of Mannheim

Data and Web Science Group

Table of Contents

If you use the RapidMiner Linked Open Data Extension for scientific purposes, please kindly include a reference to the following paper, which describes the original algorithms:

H. Paulheim, J. Fürnkranz (2012): *Unsupervised Generation of Data Mining Features from Linked Open Data.* In: International Conference on Web Intelligence, Mining, and Semantics (WIMS'12).

1 Introduction

Linked Open Data[1] is a collection of freely accessible, machine interpretable data on the web. It uses semantic web standards like RDF for data representation. As of today, there are more than 200 datasets from various domains, encompassing general knowledge as well as specific domains such as government, geographic, and life science data, scientific publications, and media. A detailed list of datasets that are available as Linked Open Data can be obtained at http://lod-cloud.net/.

Essentially, Linked Open Data contains data as graphs with labeled edges, where each entity is represented by a dereferencable URI. Each statement can be understood as a triple with subject, predicate, and object, such as

<http://dbpedia.org/resource/Germany> <http://dbpedia.org/ontology/capital> <http://dbpedia.org/resource/Berlin> .

The RapidMiner Linked Open Data extension leverages those data sources both to create useful datasets for Data Mining, as well as for adding background knowledge to existing datasets. For example, on a dataset for countries, data like the population, area, and GDP can be added from DBpedia[2] in order to improve the results of the data mining process.

In particular, the extension works with datasets that provide a SPARQL endpoint, i.e., a web service which delivers data using the query language SPARQL. A list of such datasets is available at datahub.io.

Furthermore, data from Linked Open Data sources can be used as input to the data mining process, e.g., reading data from the Eurostat Linked Open Data endpoint and running analysis on the data with RapidMiner.

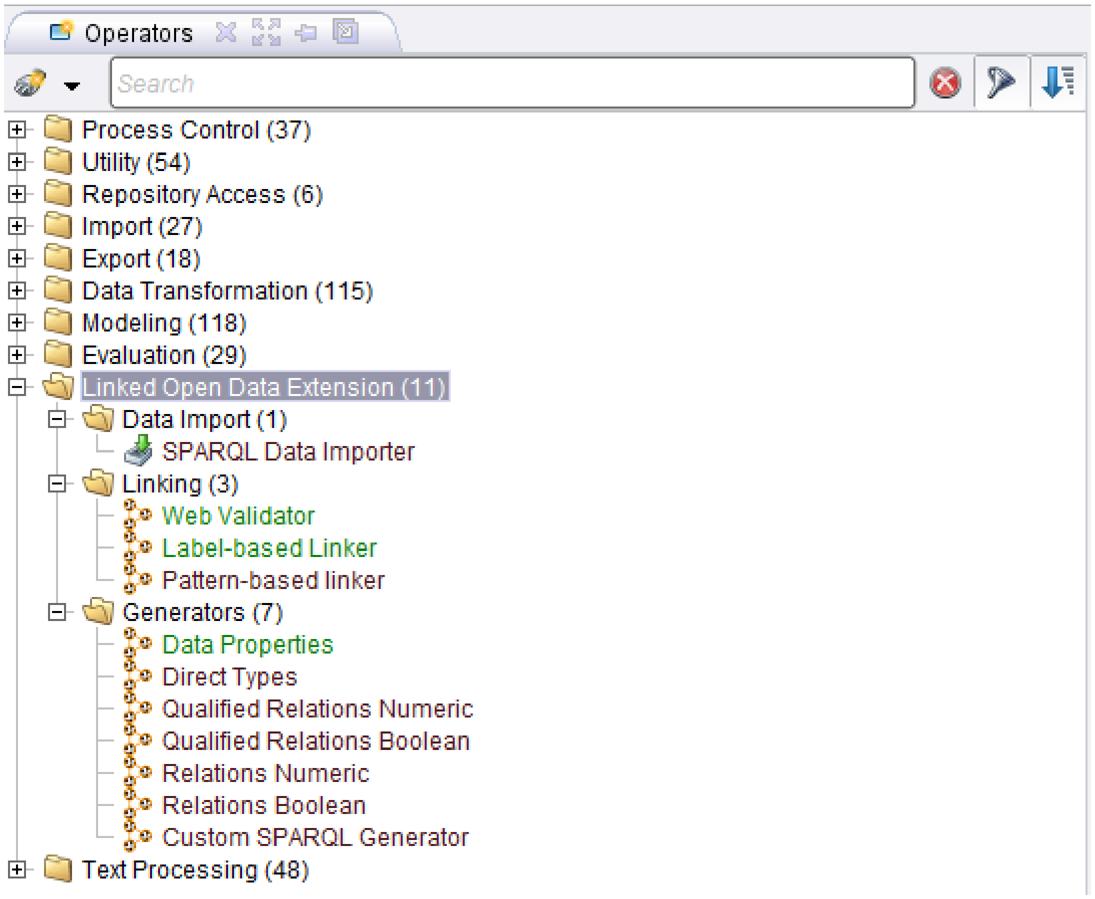
Important notice: The RapidMiner LOD Extension usually accesses live data from the web. This can result in longer operator runtimes.

1http://lod-cloud.net/

2http://dbpedia.org/

2 Enabling the RapidMiner LOD Extension

1. Add plugin (.jar file) to RapidMiner plugin directory (C:\Program Files\Rapid-I\RapidMiner5\lib\plugins by default), or add the extension via RapidMiner marketplace.
2. Create a new process. You will see a new entry (Linked Open Data Extension) in the catalog available:



Operators

Search

- Process Control (37)
- Utility (54)
- Repository Access (6)
- Import (27)
- Export (18)
- Data Transformation (115)
- Modeling (118)
- Evaluation (29)
- Linked Open Data Extension (11)
  - Data Import (1)
    - SPARQL Data Importer
  - Linking (3)
    - Web Validator
    - Label-based Linker
    - Pattern-based linker
  - Generators (7)
    - Data Properties
    - Direct Types
    - Qualified Relations Numeric
    - Qualified Relations Boolean
    - Relations Numeric
    - Relations Boolean
    - Custom SPARQL Generator
- Text Processing (48)

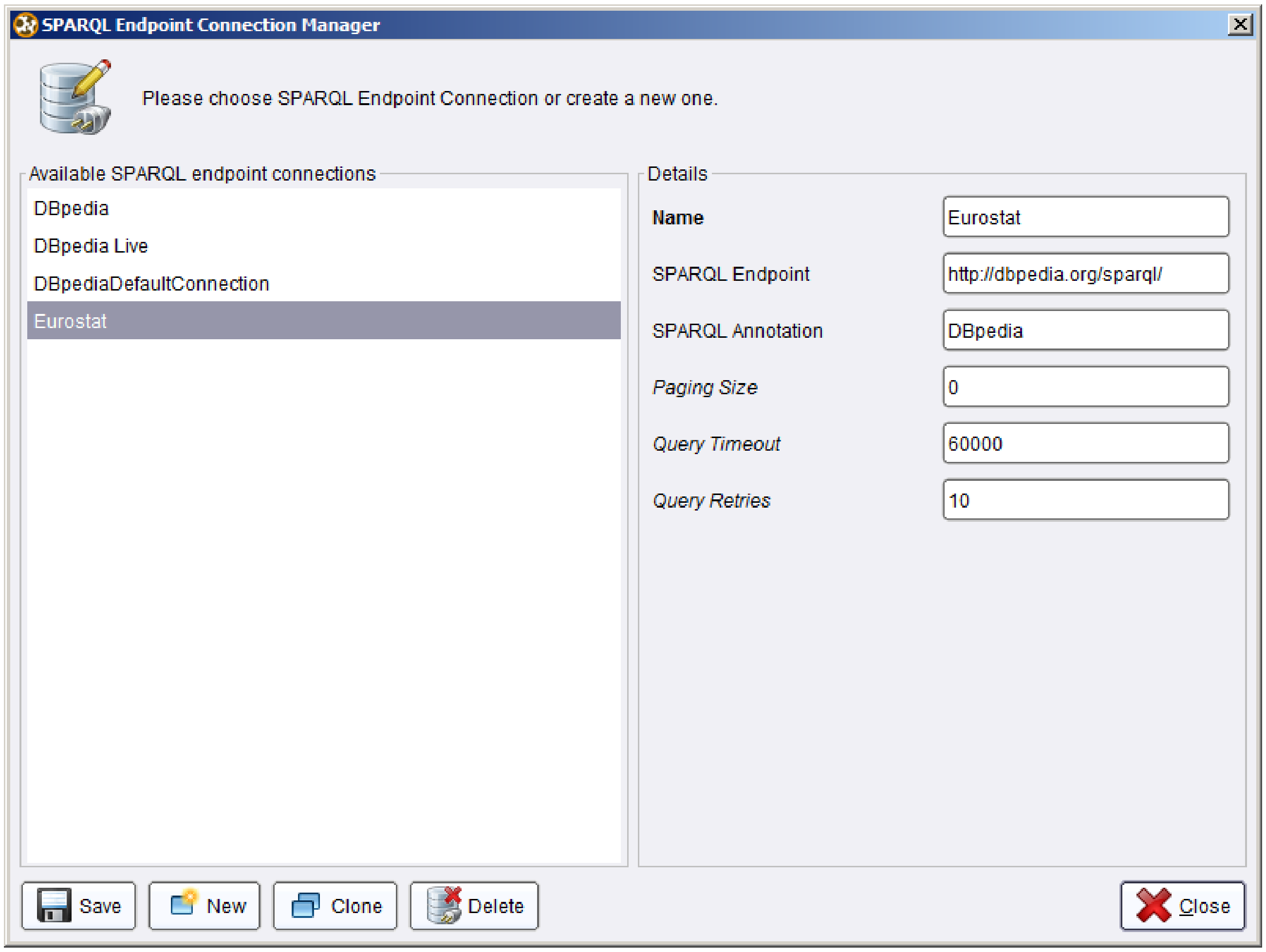
3 Configuring SPARQL Endpoints

Most Linked Open Data sources provide a SPARQL endpoint. These endpoints allow for querying the underlying dataset using a standardized query language, i.e., SPARQL[3].

Since most of the operators of the Linked Open Data extension use SPARQL, you will first have to define SPARQL Endpoints for the datasets you are planning to use.

To manage SPARQL Endpoints, go to the menu "Tools"->"Manage SPARQL Endpoint Connections…". The SPARQL Endpoint Configuration Dialog Opens:

3http://www.w3.org/TR/sparql11-overview/



Please choose SPARQL Endpoint Connection or create a new one.

Available SPARQL endpoint connections

- DBpedia
- DBpedia Live
- DBpediaDefaultConnection
- Eurostat

Details

Name	Eurostat
SPARQL Endpoint	http://dbpedia.org/sparql/
SPARQL Annotation	DBpedia
Paging Size	0
Query Timeout	60000
Query Retries	10

You can create new SPARQL Endpoints, as well as change the settings for existing ones.

The configuration parameters for SPARQL Endpoints are:

- Name: the SPARQL Endpoint will be selectable later by that name
- SPARQL Endpoint: the URL of the service
- SPARQL Annotation: a string that will be used for naming attributes generated from this SPARQL endpoint

Three more expert parameters are available (i.e., they appear only if RapidMiner is executed in expert mode):

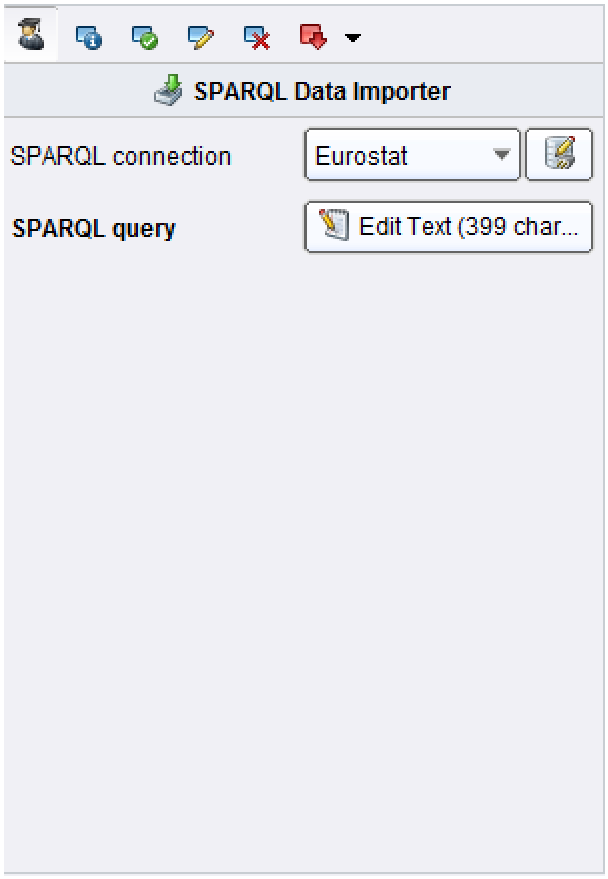
- Paging Size: many SPARQL endpoints have limitations w.r.t. to the amount of data they deliver. Paging ensures that all data is retrieved. For example, with a paging size value of 1000, a SELECT query is broken down into several queries like
  ```
  SELECT … LIMIT 1000 OFFSET 0
  SELECT … LIMIT 1000 OFFSET 1000
  SELECT … LIMIT 1000 OFFSET 2000
  ```
  etc.
  A Paging Size value of 0 disables paging.
- Query timeout defines the time which a SPARQL endpoint is given to respond (in milliseconds)

- Query retries defines the number of times a query is retried. If there are temporary errors, e.g., network problems, simply re-running the same query often resolves the issue.

Once the SPARQL Endpoint is defined, it may be used in subsequent operators. SPARQL Endpoint definitions for DBpedia and Eurostat are included as preconfigured endpoints.

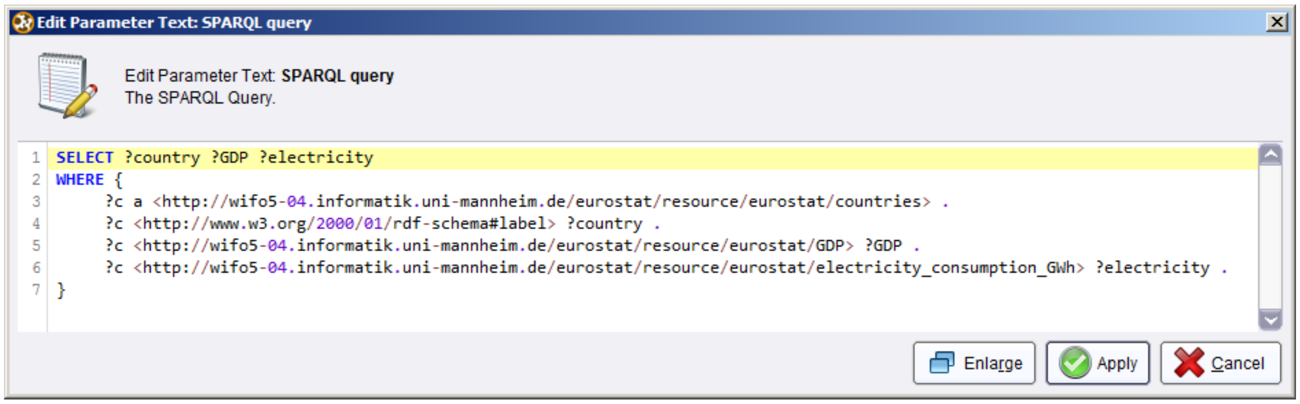
4 Using Linked Open Data as Input for Data Mining

The RapidMiner Linked Open Data Extension provides an import operator to read data from a Linked Open Data source into a RapidMiner table. The SPARQL Data Importer uses a defined SPARQL Endpoint and a custom query to generate a RapidMiner table:



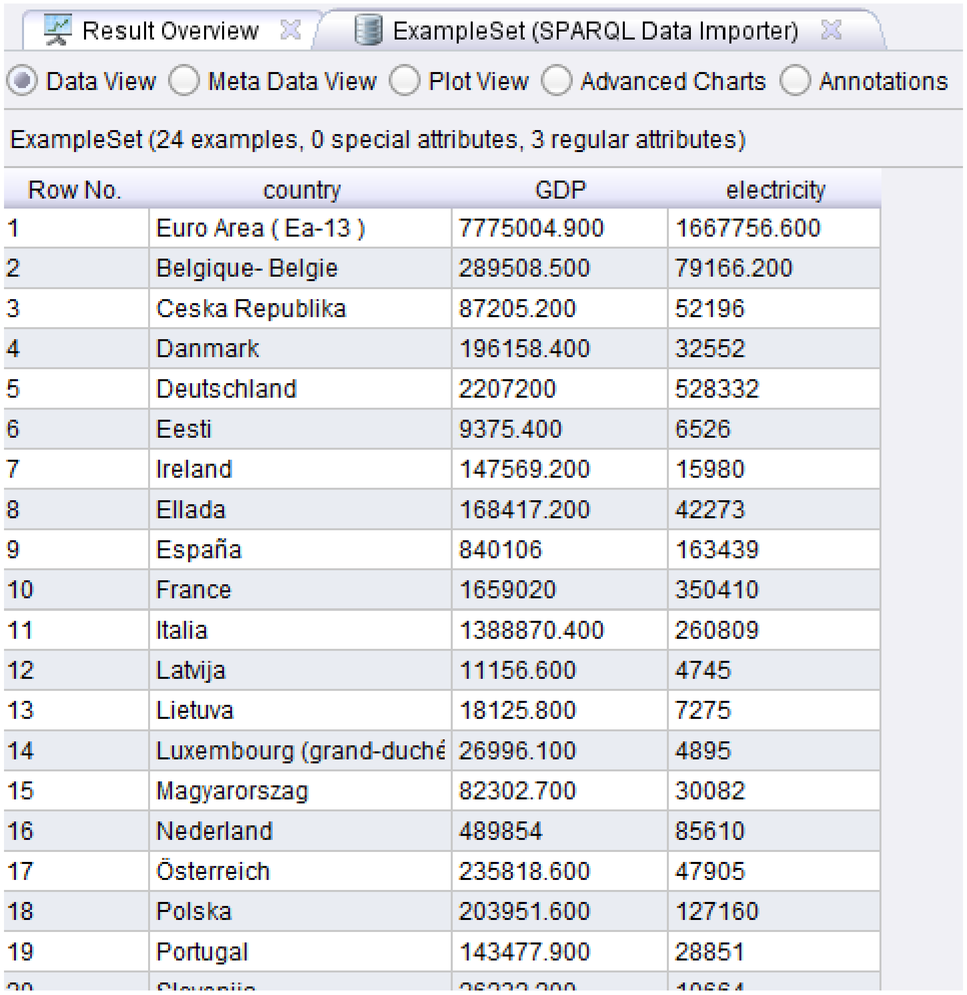
SPARQL connection: Eurostat
SPARQL query

The SPARQL query is defined as follows:


Edit Parameter Text: SPARQL query
Edit Parameter Text: SPARQL query
The SPARQL Query.
Enlarge
Apply
Cancel

```
SELECT ?country ?GDP ?electricity
WHERE {
     ?c a <http://wifo5-04.informatik.uni-mannheim.de/eurostat/resource/eurostat/countries> .
     ?c <http://www.w3.org/2000/01/rdf-schema#label> ?country .
     ?c <http://wifo5-04.informatik.uni-mannheim.de/eurostat/resource/eurostat/GDP> ?GDP .
     ?c <http://wifo5-04.informatik.uni-mannheim.de/eurostat/resource/eurostat/electricity_consumption_GWh> ?electricity .
}
```

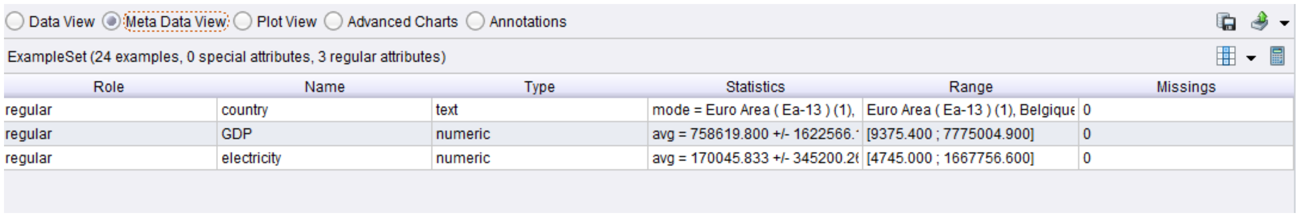
The table which is generated in the above example has three columns, named after the SPARQL query variables, i.e., country, GDP, and electricity. The output of the operator thus looks as follows:



ExampleSet (24 examples, 0 special attributes, 3 regular attributes)

Row No.	country	GDP	electricity
1	Euro Area (Ea-13)	7775004.900	1667756.600
2	Belgique- Belgie	289508.500	79166.200
3	Ceska Republika	87205.200	52196
4	Danmark	196158.400	32552
5	Deutschland	2207200	528332
6	Eesti	9375.400	6526
7	Ireland	147569.200	15980
8	Ellada	168417.200	42273
9	España	840106	163439
10	France	1659020	350410
11	Italia	1388870.400	260809
12	Latvija	11156.600	4745
13	Lietuva	18125.800	7275
14	Luxembourg (grand-duché	26996.100	4895
15	Magyarorszag	82302.700	30082
16	Nederland	489854	85610
17	Österreich	235818.600	47905
18	Polska	203951.600	127160
19	Portugal	143477.900	28851

In the meta data view, you will observe that the operator automatically assigns meaningful data types to the values read from the SPARQL Endpoint. In that case, the country name is a text attribute, while GDP and electricity are numeric:



ExampleSet (24 examples, 0 special attributes, 3 regular attributes)

Role	Name	Type	Statistics	Range	Missings
regular	country	text	mode = Euro Area (Ea-13) (1),	Euro Area (Ea-13) (1), Belgique	0
regular	GDP	numeric	avg = 758619.800 +/- 1622566.	[9375.400 ; 7775004.900]	0
regular	electricity	numeric	avg = 170045.833 +/- 345200.2	[4745.000 ; 1667756.600]	0

5 Linking your Data to Existing Datasets

For using Linked Open Data as background knowledge for existing datasets, the first step is to link your data to a dataset from which you want to use additional data. For example, your dataset may contain a column “Country” with values “Germany”, “France”, etc. Linking your dataset to Linked Open Data means that an additional column is added, which contains URIs identifying those entities in Linked Open Data, such as http://dbpedia.org/resource/Germany, http://dbpedia.org/resource/France, etc.

The current version of the RapidMiner Linked Open Data extension comprises two linking methods:

1. The Pattern-based Linker builds links using a defined URI pattern. For example, the contents of the column "Country" are appended to a fixed string, e.g., "http://dbpedia.org/resource/".

2. The Label-based linker searches for the contents of a column (and optionally word n-grams) in a data source, and finds the most suitable resource based on their labels. For example, for a string "United States of America", the linker would search for word n-grams such as "United States" as well, and from all the results, use the best suited one.

Each linker has two output ports:

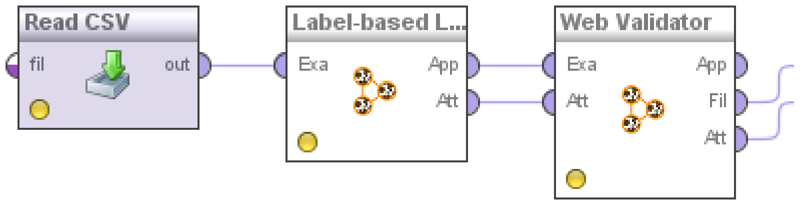
1. App contains the data with the new column depicting the URI

2. Att contains the attribute name which holds the URIs

The column is named ORIGINAL_COLUMN_link_to_SPARQL_ENDPOINT_ANNOTATION, where ORIGINAL_COLUMN is the name of the column where the values come from (e.g., City), and SPARQL_ENDPOINT_ANNOTATION is the value set when configuring the SPARQL endpoint. For example, linking a column City to DBpedia will result in a new column City_link_to_DBpedia.

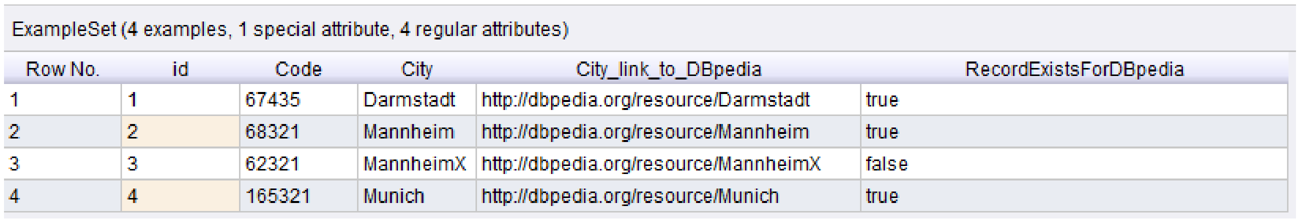
The Att port has to be connected to subsequent steps, e.g., generators, to inform them on which attribute to perform data generation.

Since some naïve linkers (such as the Label-based Linker) may generate links that actually do not exist, the Web Validator checks the existence of each link and removes instances for which the link does not exist. Furthermore, it removes all instances for which no link could be found. The full pipeline of linking looks as follows:



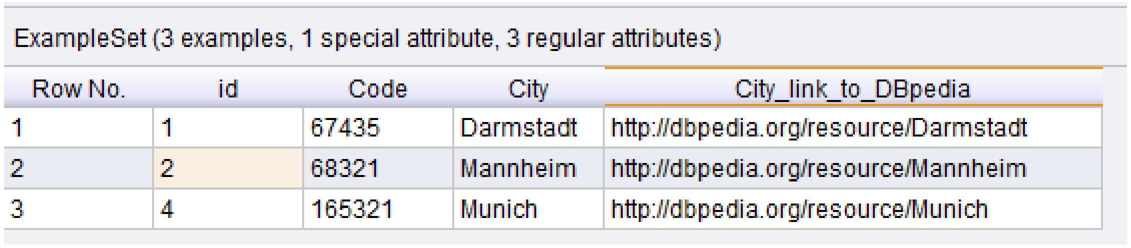
The Web Validator has three outputs: App is the original data with an additional attribute boolean flag indicating whether the link is valid or not, Fil is the filtered data with instances removed for which no link is found, and Att bypasses the original Att input port.

Assume that the original set contains an entity which cannot be resolved in DBpedia, i.e., "MannheimX". Linking that dataset to DBpedia with the simple linker results in the following table:


ExampleSet (4 examples, 1 special attribute, 4 regular attributes)

Row No.	id	Code	City	City_link_to_DBpedia	RecordExistsForDBpedia
1	1	67435	Darmstadt	http://dbpedia.org/resource/Darmstadt	true
2	2	68321	Mannheim	http://dbpedia.org/resource/Mannheim	true
3	3	62321	MannheimX	http://dbpedia.org/resource/MannheimX	false
4	4	165321	Munich	http://dbpedia.org/resource/Munich	true

The output produced by the Web Validator operator then filters the table as follows:



ExampleSet (3 examples, 1 special attribute, 3 regular attributes)

Row No.	id	Code	City	City_link_to_DBpedia
1	1	67435	Darmstadt	http://dbpedia.org/resource/Darmstadt
2	2	68321	Mannheim	http://dbpedia.org/resource/Mannheim
3	4	165321	Munich	http://dbpedia.org/resource/Munich

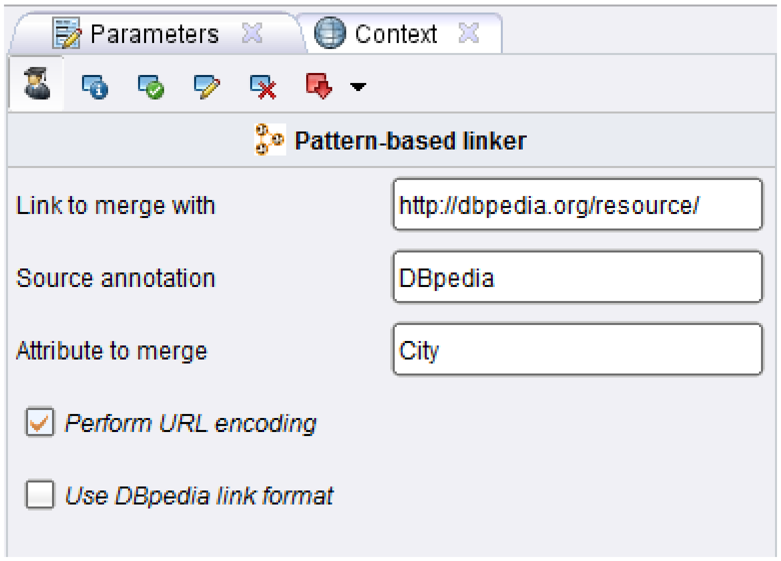
Observation: the entity MannheimX, for which the DBpedia link is wrong, is removed from the dataset.

5.1 The Pattern-based Linker

The Pattern-based linker provides a simple method for creating links, which simply guesses links by forming simple URI patterns. The linker takes three arguments:

1. The link prefix
2. The source annotation
3. The attribute to link

Links are created by concatenating the link prefix with the attribute to link. The source annotation field is used for creating a name for the attribute containing the link.



Link to merge with	http://dbpedia.org/resource/
Source annotation	DBpedia
Attribute to merge	City

[✓] Perform URL encoding

[ ] Use DBpedia link format

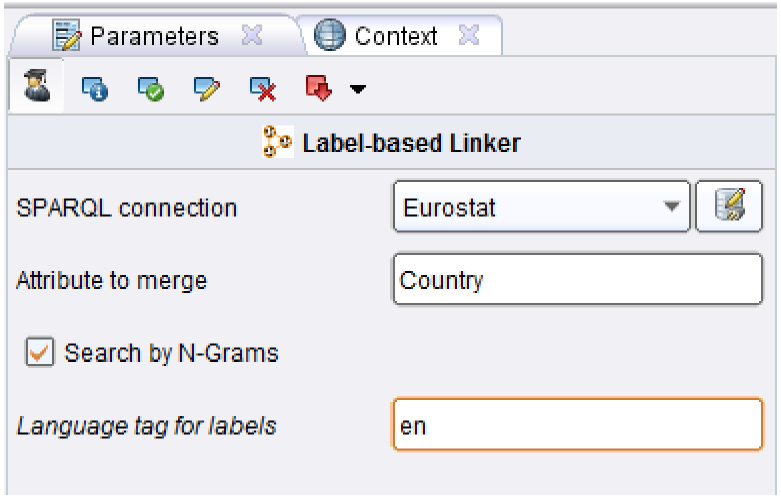
Further attributes that can be set are whether to perform URL Encoding (i.e., replacing special characters and creating proper UTF-8 links), and whether to use a specific link format for DBpedia (e.g., representing blanks as underscores).

In the example shown above, a column "City" containing values such as "Mannheim", "Darmstadt" etc. would result in the corresponding links "http://dbpedia.org/resource/Mannheim", "http://dbpedia.org/resource/Darmstadt", etc.

5.2 The Label-based Linker

The Label-based Linker identifies entities by issuing SPARQL queries against an endpoint, trying to find entities whose labels contain a given string. It takes four parameters:

1. A SPARQL endpoint
2. The attribute to merge (see Pattern-based linker)
3. A flag indicating whether the search should include n-grams.



SPARQL connection: Eurostat
Attribute to merge: Country
[✓] Search by N-Grams
Language tag for labels: en

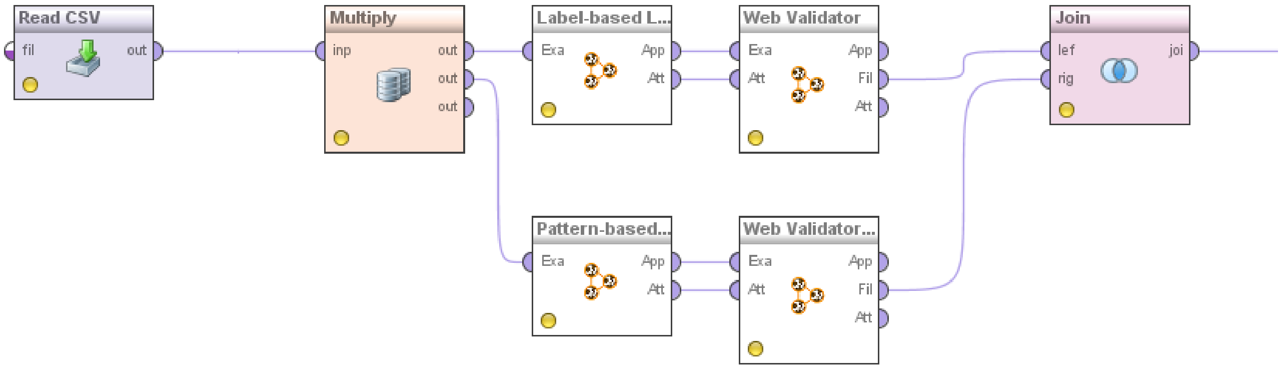
When including n-grams in the search, the linker will first try to identify entities whose labels contain the string in its entirety, e.g., “United States of America”. If this does not succeed, it will continue with sub tokens, eventually also finding entities whose label is only “United States”.

Note that the SPARQL Based linker, in particular when using the “Search by N-Grams” option, can take some time to link a dataset, since it performs full text search via SPARQL, which, depending on the endpoint used, may not be performed very fast.

In expert mode, you can also restrict the search to labels with a certain language tag, e.g., “en”. If the attribute is left empty (which is the default), labels in all languages, as well as without any language tag, are included in the search.

5.3 Combining Linkers

In some cases, it might be useful to combine linkers – e.g., to add information from multiple sources, or to add information about different entities in a dataset, e.g., both the film as well as its director. In that case, the data can be multiplied with the Multiply operator (in “Process control”), and the output of both linkers is joined with the Join operator (in “Data Transformation”). Note that for the latter step, the original input data needs to define at least one ID column.



6 Using generators

The RapidMiner Linked Open Data extension supports different generators, which perform different knowledge extraction and aggregation steps.

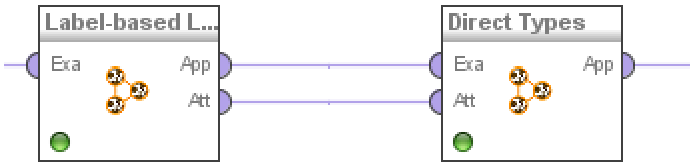
6.1 Direct Types Generator

The Direct Types Generator extracts all statements with predicate `rdf:type` and creates a boolean attribute for each type. For example, from a statement like

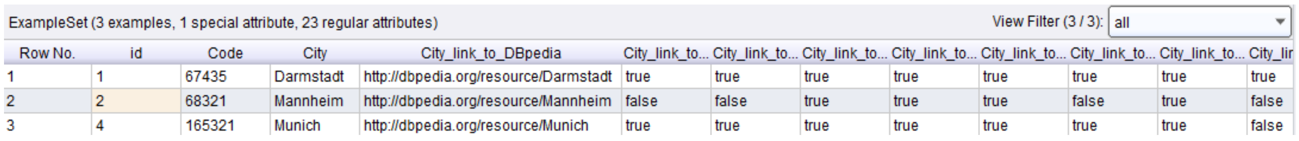
```
<http://dbpedia.org/resource/Mannheim> rdf:type
<http://dbpedia.org/ontology/City> .
```

an attribute with name City_link_to_DBpedia would be created, and the value for the instance linked to <http://dbpedia.org/resource/Mannheim> would be set to true.

The process using the Simple Types Generator looks as follows:



Note that both outputs of the linker – App and Att – need to wired to the respective input ports of the generator. The output looks as follows:



ExampleSet (3 examples, 1 special attribute, 23 regular attributes)

View Filter (3 / 3): all

Row No.	id	Code	City	City_link_to_DBpedia	City_link_to...	City_link_to...	City_link_to...	City_link_to...	City_link_to...	City_link_to...	City_link_to...	City_lir
1	1	67435	Darmstadt	http://dbpedia.org/resource/Darmstadt	true	true	true	true	true	true	true	true
2	2	68321	Mannheim	http://dbpedia.org/resource/Mannheim	false	false	true	true	true	false	true	false
3	4	165321	Munich	http://dbpedia.org/resource/Munich	true	true	true	true	true	true	true	false

6.2 Data Properties Generator

The Data Properties Generator creates an attribute for each literal value that the linked instances has. For example, from the property

```
<http://dbpedia.org/resource/Mannheim>
<http://dbpedia.org/ontology/population> "123873"^^xsd:integer .
```

the generator would create an attribute City_link_to_DBpedia_data_http://dbpedia.org/ontology/population and set the value for the instance linked to Mannheim to 123873. Some basic guessing of data types is performed, e.g., numeric values are marked as such, and numeric literals without an explicit type are parsed into numbers, if possible.

6.3 Unqualified and Qualified Relation Generators

The RapidMiner Linked Open Data extension supports generators exploiting relations to other resources. The *unqualified* relations generators create attributes from the existence of relations, while the *qualified* relations generators also take the types of the related resources into account. Both generators exist in two variants, one generating boolean, one generating numeric features.

Consider the following triples:

```
<http://dbpedia.org/resource/Berlin>
<http://dbpedia.org/ontology/capitalOf>
<http://dbpedia.org/resource/Germany> .

<http://dbpedia.org/resource/Germany> rdf:type
<http://dbpedia.org/ontology/Country> .
```

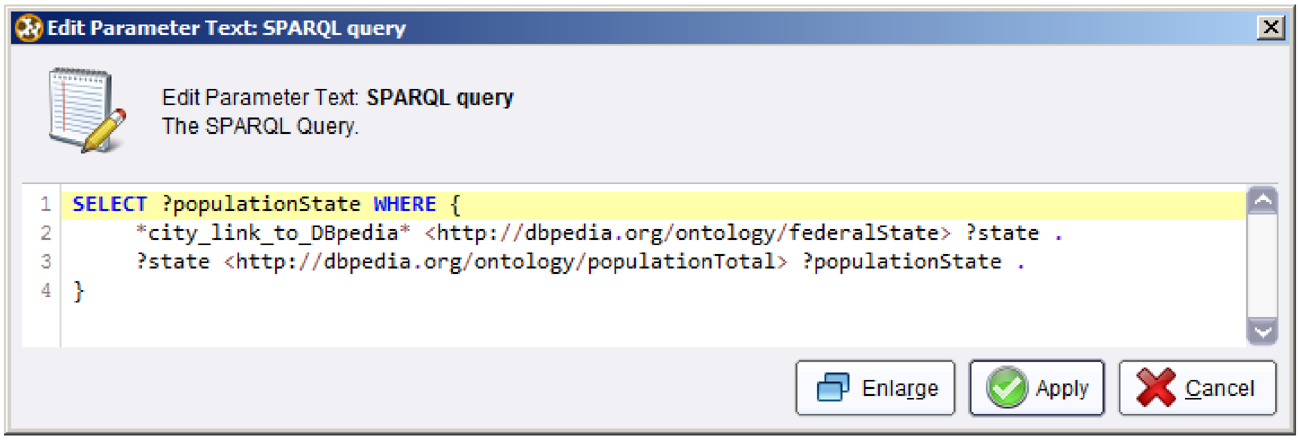
The Relation Boolean generator would create an attribute City_link_to_DBpedia_in_boolean_http://dbpedia.org/ontology/capitalOf and set its value to *true* for an instance linked to <http://dbpedia.org/resource/Berlin> . The Relation Numeric generator would create an attribute City_link_to_DBpedia_in_numeric_http://dbpedia.org/ontology/capitalOf and set its value to *1* for an instance linked to <http://dbpedia.org/resource/Berlin>.

The Qualified Relation Boolean generator would create an attribute City_link_to_DBpedia_in_boolean_http://dbpedia.org/ontology/capitalOf_type_http://dbpedia.org/ontology/Country and set its value to *true* for an instance linked to <http://dbpedia.org/resource/Berlin> . The Qualified Relation Numeric generator would create an attribute City_link_to_DBpedia_in_numeric_http://dbpedia.org/ontology/capitalOf_type_http://dbpedia.org/ontology/Country and set its value to *1* for an instance linked to <http://dbpedia.org/resource/Berlin>.

Note that in especially the qualified relation generators, while very powerful, may create a very large number of attributes, which can cause problems for subsequent processing steps.

6.4 The Custom SPARQL Generator

In cases where you know the dataset you are using and want to add specific data not covered by any of the default generators, you can also define your own SPARQL statements. For example, for a dataset of cities, you might want to add the population of the state that the city is in. This can be done with the custom SPARQL generator, which allows to create a custom statement:

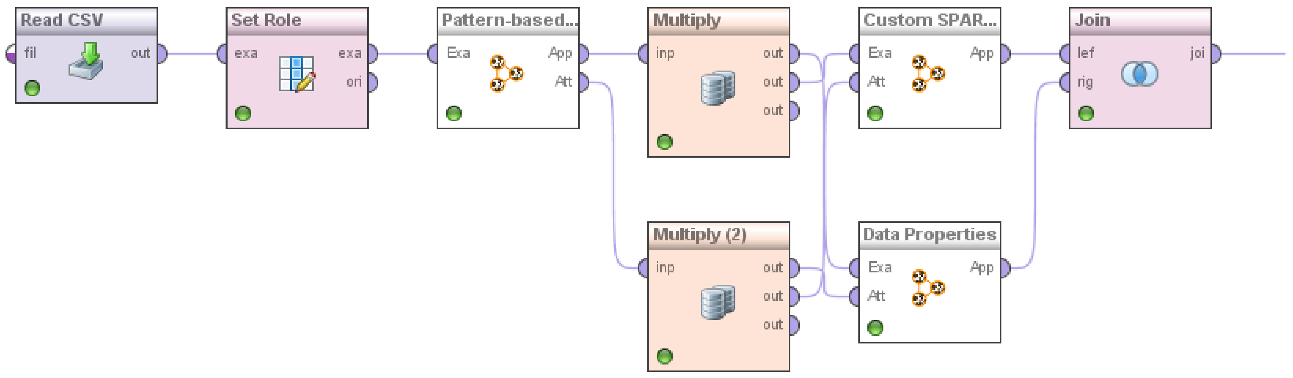

Edit Parameter Text: SPARQL query
Edit Parameter Text: SPARQL query
The SPARQL Query.
Enlarge
Apply
Cancel

```
SELECT ?populationState WHERE {
      *city_link_to_DBpedia* <http://dbpedia.org/ontology/federalState> ?state .
      ?state <http://dbpedia.org/ontology/populationTotal> ?populationState .
}
```

Within the statement, you can use links generated by linkers as placeholders, enclosed in asterisks. In the above example, a linker has generated the link attribute city_link_to_DBpedia, which is used in the query. The generator then generates a new attribute called “populationState” (using the variable name for naming the attribute) and adds it to the dataset.

6.5 Combining Generators

For combining generators on the same link, e.g., extracting both data properties and types, the following process setup has to be used:



Two Multiply operators are used, one for multiplying the data, one for multiplying the attribute information. Like for combining linkers, the input data needs to define at least one ID attribute to allow the final Join. The Set Role operator can be used for defining IDs.

Note that you can also combine generators with data from different data sources. To that end, you have to run a linker and one or more generators for each data source.

7 Storing Results

For larger datasets, the runtimes of linkers and generators can be quite high, depending on the dataset and specific operators used. Since in a typical workflow, you will first generate data and then experiment with a larger number of subsequent operators (e.g., classification or regression algorithms), we suggest that you store the output dataset of the LOD operator chain in an intermediate file (e.g., CSV) or database, and perform all subsequent analysis on that stored dataset. That way, you will have to run the time consuming access to Linked Open Data only once.

8 Release Notes

Version 1.0.071, released on September 13th, 2013.