Incorporating Functions in Mappings to Facilitate the Uplift of CSV Files into RDF

Ademar Crotti Junior
crottija@scss.tcd.ie

Christophe Debruyne
debruync@scss.tcd.ie

Declan O'Sullivan
declan.osullivan@scss.tcd.ie

Trinity College Dublin

Introduction

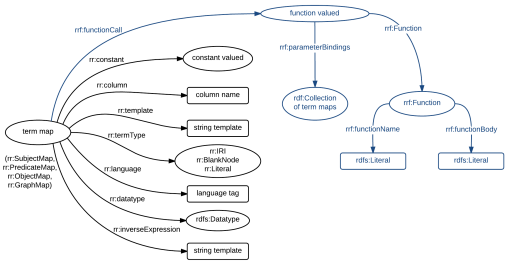
Many solutions have been developed to convert non-RDF data to RDF. A common task during this conversion is applying data manipulation functions to obtain the desired output. Depending on the data format of the source to be transformed, one can rely on the underlying technology, such as RDBMS for relational databases or XQuery for XML, to manipulate data - to a certain extent - while generating RDF. For CSV files, however, there is no such underlying technology. Instead, one has to resort to more elaborate Extract, Transform and Load (ETL) processes, which can render the generation of RDF more complex, and therefore less traceable and transparent. One solution to this problem is the declaration and inclusion of functions in mappings of non-RDF data to RDF.

Incorporating Functions in Mapping Languages

Starting from work presented in [1], we define functions as resources with function names and function bodies. Function names are unique and each function must have one function name and one function body. A function body defines a function with a return statement; parameters are optional.

Functions can be used to capture both domain knowledge (e.g., transforming units) and other – more syntactic – data manipulation tasks (e.g., transforming values to create valid URIs).

Our implementation extends R2RML's vocabulary and RML's [2] engine.



Demonstration

CSV Input

```
NGA,Polity,Section,Variable,Value From
Latium,ItRomPr,General variables,Capital,Rome
Latium,ItRomPr,General variables,Language,Latin
Latium,ItRomPr,General variables,Supracultural entity,Greco-Roman
...
```

Mapping

```
<#Variable>
 rml:logicalSource[rml:source "data.csv";rml:referenceFormulation ql:CSV];
 rr:subjectMap [ rr:termType rr:BlankNode; ];
 rr:predicateObjectMap [
   rr:predicateMap [ rrf:functionCall [ rrf:function <#Camelize> ; rrf:parameterBindings ( [ rml:reference "Variable" ] );]; ];
  rr:objectMap [ rr:parentTriplesMap <#Value> ] ].

<#Camelize>
 rrf:functionName "camelize" ;
 rrf:functionBody """ function camelize (str) { var camelCaseString = str.toLowerCase().replace( /[-_]+/g, ' ').replace(
/[^\\w\\s]/g, ' ').replace( / (.)/g, function($1) { return $1.toUpperCase(); }).replace( / /g, '' );
return "http://dacura.cs.tcd.ie/data/seshat/" + camelCaseString; } """ ; .
```

Uplift Engine

Output

```
<http://dacura.cs.tcd.ie/data/seshat/ItRomPr> <http://dacura.cs.tcd.ie/data/seshat#hasVariable> _:d3SRTs6uGh .
_:d3SRTs6uGh <http://dacura.cs.tcd.ie/data/seshat/capital> _:d8fF2wuiam .
_:d8fF2wuiam <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <http://dacura.cs.tcd.ie/data/seshat#NameVariable> .
_:d8fF2wuiam <http://dacura.cs.tcd.ie/data/seshat#definiteDataValue> "Rome" .
<http://dacura.cs.tcd.ie/data/seshat/ItRomPr> <http://dacura.cs.tcd.ie/data/seshat#hasVariable> _:genid344T94t1y4CS .
_:genid344T94t1y4CS <http://dacura.cs.tcd.ie/data/seshat/language> _:qjeWSDRDcd .
_:qjeWSDRDcd <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <http://dacura.cs.tcd.ie/data/seshat#NameVariable> .
_:qjeWSDRDcd <http://dacura.cs.tcd.ie/data/seshat#definiteDataValue> "Latin" .
<http://dacura.cs.tcd.ie/data/seshat/ItRomPr> <http://dacura.cs.tcd.ie/data/seshat#hasVariable> _:genid388sbqxWXT4T .
_:genid388sbqxWXT4T <http://dacura.cs.tcd.ie/data/seshat/supraculturalEntity> _:yeEQUyN1yW .
_:yeEQUyN1yW <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <http://dacura.cs.tcd.ie/data/seshat#NameVariable> .
_:yeEQUyN1yW <http://dacura.cs.tcd.ie/data/seshat#definiteDataValue> "Greco-Roman" .
...
```

Code and examples available at https://www.scss.tcd.ie/~crottija/funul/

References

[1] Debruyne, C., O'Sullivan, D.: R2RML-F: Towards Sharing and Executing Domain Logic in R2RML Mappings. In: Workshop on Linked Data on the Web (2016).

[2] Dimou, A., Vander Sande, M., Colpaert, P., Verborgh, R., Mannens, E., Van de Walle, R.: RML: A Generic Language for Integrated RDF Mappings of Heterogeneous Data. In: Workshop on Linked Data on the Web. (2014).

Ireland's European Structural and Investment Funds Programmes 2014-2020 Co-funded by the Irish Government and the European Union
European Union European Regional Development Fund
SFI Science Foundation Ireland

The ADAPT Centre for Digital Content Technology is funded under the SFI Research Centres Programme (Grant 13/RC/2106) and is co-funded under the European Regional Development Fund.