High Quality Linked Data Generation

Dr Anastasia Dimou
Post-Doc researcher

🌐 imec.be - IDLab.technology
✉️ Anastasia.Dimou@imec.be
🐦 @natadimou
Linked Data

best practices for semantically annotating & connecting structured data on the (Semantic) Web
Linked Data Generation

Linked Data is (still) not intuitively available on the Web
Linked Data Generation from Heterogeneous Data

Linked Data is derived from heterogeneous data sources, structures & formats
High Quality Linked Data Generation from Heterogeneous Data

Linked Data needs to be consistent
Human agents do not want to put in effort to provide Linked Data until there are software agents that use it.
Why is Linked Data Generation Such a difficult task?
63% of all SWIB participants have created tools to convert MARC to RDF #swib17

@FakeLibStats

Linked Data

custom implementation
data owner

- DB
  - Linked Data
- CSV
  - Linked Data
- XML
  - Linked Data
- JSON
  - Linked Data

format-specific implementation
facilitate Linked Data generation by reducing data & semantic heterogeneity & increasing Linked Data quality
data owner

- DB
- CSV
- XML
- JSON

Linked Data

multiple data sources
data owner

DB, CSV, XML, JSON

Linked Data

uniform execution
data owner

- DB
- CSV
- XML
- JSON

uniform declaration
data owner

DB
CSV
XML
JSON

Linked Data

uniform declaration

Osma Suominen @OsmaSuominen 18h
Replying to @richtich @FakeLibStats
And 89% of them created their own ontology to represent the result
declaration
execution
assessment
Extending R2RML to a source-independent mapping language for RDF
A. Dimou et al.
http://RML.io

<table>
<thead>
<tr>
<th>id</th>
<th>title</th>
<th>venue</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Assessing &amp; Refining Mappings to RDF to improve Dataset Quality</td>
<td>ISWC 2015</td>
</tr>
<tr>
<td>2</td>
<td>RMLEditor : a graph-based editor for Linked Data mappings</td>
<td>ESWC 2016</td>
</tr>
<tr>
<td>3</td>
<td>An ontology to semantically declare and describe functions</td>
<td>ESWC 2016</td>
</tr>
<tr>
<td>4</td>
<td>Modeling, Generating, Publishing Knowledge as Linked Data</td>
<td>EKAW 2017</td>
</tr>
<tr>
<td>5</td>
<td>Semi-automatic example-driven linked data mapping creation</td>
<td>ISWC 2017</td>
</tr>
<tr>
<td>id</td>
<td>title</td>
<td>venue</td>
</tr>
<tr>
<td>----</td>
<td>-----------------------------------------------------------------------</td>
<td>----------------</td>
</tr>
<tr>
<td>1</td>
<td>Assessing &amp; Refining Mappings to RDF to improve Dataset Quality</td>
<td>ISWC 2015</td>
</tr>
<tr>
<td>2</td>
<td>RMLEditor: a graph-based editor for Linked Data mappings</td>
<td>ESWC 2016</td>
</tr>
<tr>
<td>3</td>
<td>An ontology to semantically declare and describe functions</td>
<td>ESWC 2016</td>
</tr>
<tr>
<td>4</td>
<td>Modeling, Generating, Publishing Knowledge as Linked Data</td>
<td>EKAW 2017</td>
</tr>
<tr>
<td>5</td>
<td>Semi-automatic example-driven linked data mapping creation</td>
<td>ISWC 2017</td>
</tr>
</tbody>
</table>
Who defines those rules?
rr:subjectMap [  
  rr:template "http://data.kb.nl/ppn/{repox:metadata/record/datafield[@tag="003@"]/subfield[@code="0"]}";
  rr:class ex:Record ];

rr:predicateObjectMap [  
  rr:predicate rdfs:label;
  rr:objectMap [  
    rml:reference "repox:metadata/record/datafield[@tag="021A"]/subfield[@code="a"]";
    rr:language "nl" ] ];

rr:predicateObjectMap [  
  rr:predicate dcterms:type;  
  rr:objectMap [ rml:reference "repox:metadata/record/datafield[@tag="002@"]/subfield[@code="0"]" ] ];

rr:predicateObjectMap [  
  rr:predicate dcterms:publisher;  
  rr:objectMap [ rml:reference "repox:metadata/record/datafield[@tag="033A"]/subfield[@code="n"]" ] ];

rr:predicateObjectMap [  
  rr:predicate rdfs:seeAlso;  
  rr:objectMap [ rr:template "http://www.worldcat.org/oclc/{metadata/record/datafield[@tag="003O"]/subfield[@code="0"]}" ] ];
RMLEditor: A Graph-Based Mapping Editor for Linked Data Mappings
P. Heyvaert, et al.
declaration
execution
assessment

declare
mapping rules
RML Language
A. Dimou et al.

Machine-interpretable dataset & service descriptions for heterogeneous data access and retrieval.
Machine-interpretable dataset & service descriptions for heterogeneous data access and retrieval. A. Dimou et al.
An Ontology to Semantically Declare & Describe Functions

B. De Meester, A. Dimou, R. Verborgh, E. Mannens & R. Van De Walle
RML Mapper: a tool for uniform Linked Data generation from heterogeneous data. A. Dimou et al.
declaration
execution
assessment
<table>
<thead>
<tr>
<th>id</th>
<th>title</th>
<th>venue</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Assessing &amp; Refining Mappings to RDF to improve Dataset Quality</td>
<td>ISWC 2015</td>
</tr>
<tr>
<td>2</td>
<td>RMLEditor: a graph-based editor for Linked Data mappings</td>
<td>ESWC 2016</td>
</tr>
<tr>
<td>3</td>
<td>An ontology to semantically declare and describe functions</td>
<td>ESWC 2016</td>
</tr>
<tr>
<td>4</td>
<td>Modeling, Generating, Publishing Knowledge as Linked Data</td>
<td>EKAW 2017</td>
</tr>
<tr>
<td>5</td>
<td>Semi-automatic example-driven linked data mapping creation</td>
<td>ISWC 2017</td>
</tr>
<tr>
<td>id</td>
<td>title</td>
<td>venue</td>
</tr>
<tr>
<td>----</td>
<td>----------------------------------------------------------------------</td>
<td>-------------------</td>
</tr>
<tr>
<td>1</td>
<td>Assessing &amp; Refining Mappings to RDF to improve Dataset Quality</td>
<td>ISWC 2015</td>
</tr>
<tr>
<td>2</td>
<td>RMLEditor : a graph-based editor for Linked Data mappings</td>
<td>ESWC 2016</td>
</tr>
<tr>
<td>3</td>
<td>An ontology to semantically declare and describe functions</td>
<td>ESWC 2016</td>
</tr>
<tr>
<td>4</td>
<td>Modeling, Generating, Publishing Knowledge as Linked Data</td>
<td>EKAW 2017</td>
</tr>
<tr>
<td>5</td>
<td>Semi-automatic example-driven linked data mapping creation</td>
<td>ISWC 2017</td>
</tr>
<tr>
<td>id</td>
<td>title</td>
<td>venue</td>
</tr>
<tr>
<td>----</td>
<td>----------------------------------------------------------------------</td>
<td>---------------</td>
</tr>
<tr>
<td>1</td>
<td>Assessing &amp; Refining Mappings to RDF to improve Dataset Quality</td>
<td>ISWC 2015</td>
</tr>
<tr>
<td>2</td>
<td>RMLEditor: a graph-based editor for Linked Data mappings</td>
<td>ESWC 2016</td>
</tr>
<tr>
<td>3</td>
<td>An ontology to semantically declare and describe functions</td>
<td>ESWC 2016</td>
</tr>
<tr>
<td>4</td>
<td>Modeling, Generating, Publishing Knowledge as Linked Data</td>
<td>EKAW 2017</td>
</tr>
<tr>
<td>5</td>
<td>Semi-automatic example-driven linked data mapping creation</td>
<td>ISWC 2017</td>
</tr>
</tbody>
</table>

![Diagram](image-url)
What happens then?
<table>
<thead>
<tr>
<th>id</th>
<th>title</th>
<th>venue</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Assessing &amp; Refining Mappings to RDF to improve Dataset Quality</td>
<td>ISWC 2015</td>
</tr>
<tr>
<td>2</td>
<td>RMLEditor : a graph-based editor for Linked Data mappings</td>
<td>ESWC 2016</td>
</tr>
<tr>
<td>3</td>
<td>An ontology to semantically declare and describe functions</td>
<td>ESWC 2016</td>
</tr>
<tr>
<td>4</td>
<td>Modeling, Generating, Publishing Knowledge as Linked Data</td>
<td>EKAW 2017</td>
</tr>
<tr>
<td>5</td>
<td>Semi-automatic example-driven linked data mapping creation</td>
<td>ISWC 2017</td>
</tr>
</tbody>
</table>

100 triples
2 violations/triple

200 violations!
<table>
<thead>
<tr>
<th>id</th>
<th>title</th>
<th>venue</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Assessing &amp; Refining Mappings to RDF to improve Dataset Quality</td>
<td>ISWC 2015</td>
</tr>
<tr>
<td>2</td>
<td>RMLEditor : a graph-based editor for Linked Data mappings</td>
<td>ESWC 2016</td>
</tr>
<tr>
<td>3</td>
<td>An ontology to semantically declare and describe functions</td>
<td>ESWC 2016</td>
</tr>
<tr>
<td>4</td>
<td>Modeling, Generating, Publishing Knowledge as Linked Data</td>
<td>EKAW 2017</td>
</tr>
<tr>
<td>5</td>
<td>Semi-automatic example-driven linked data mapping creation</td>
<td>ISWC 2017</td>
</tr>
</tbody>
</table>

- 1,000,000 triples
- 2 violations/triple
- 2,000,000 violations!
<table>
<thead>
<tr>
<th>id</th>
<th>title</th>
<th>venue</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Assessing &amp; Refining Mappings to RDF to improve Dataset Quality</td>
<td>ISWC 2015</td>
</tr>
<tr>
<td>2</td>
<td>RMLEditor: a graph-based editor for Linked Data mappings</td>
<td>ESWC 2016</td>
</tr>
<tr>
<td>3</td>
<td>An ontology to semantically declare and describe functions</td>
<td>ESWC 2016</td>
</tr>
<tr>
<td>4</td>
<td>Modeling, Generating, Publishing Knowledge as Linked Data</td>
<td>EKAW 2017</td>
</tr>
<tr>
<td>5</td>
<td>Semi-automatic example-driven linked data mapping creation</td>
<td>ISWC 2017</td>
</tr>
</tbody>
</table>

1,000,000 triples
2 violations/triple

2,000,000 violations!
You think this doesn’t happen?!

<table>
<thead>
<tr>
<th>Year</th>
<th>2014</th>
<th>2015</th>
</tr>
</thead>
<tbody>
<tr>
<td>Solution 1.1</td>
<td>swc:OrganizedEvent</td>
<td>swc:OrganizedEvent</td>
</tr>
<tr>
<td>Solution 1.2</td>
<td>swc:Event</td>
<td>bibo:Conference</td>
</tr>
<tr>
<td>Solution 1.3</td>
<td>swrc:Event</td>
<td>swrc:Event</td>
</tr>
<tr>
<td>Solution 1.4</td>
<td></td>
<td>swrc:Event</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Year</th>
<th>Solution 1.1</th>
<th>Solution 1.2</th>
<th>Solution 1.3</th>
<th>Solution 1.4</th>
</tr>
</thead>
<tbody>
<tr>
<td>2014</td>
<td>bibo:Workshop</td>
<td>swc:Event</td>
<td>swrc:Event</td>
<td></td>
</tr>
<tr>
<td>2015</td>
<td>bibo:Workshop</td>
<td>bibo:Workshop</td>
<td>swrc:Workshop</td>
<td>swrc:Section</td>
</tr>
<tr>
<td>Year</td>
<td>2014</td>
<td>2015</td>
<td></td>
<td></td>
</tr>
<tr>
<td>------</td>
<td>---------------</td>
<td>---------------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Solution 1.1</td>
<td>swrc:InProceedings</td>
<td>foaf:Document</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Solution 1.2</td>
<td>bibo:Article</td>
<td>swrc:InProceedings</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Solution 1.3</td>
<td>swrc:Publication</td>
<td>swrc:Publication</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Solution 1.4</td>
<td></td>
<td>swc:Paper</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Can we prevent violations?
Assessing and Refining Mappings to RDF to Improve Dataset Quality. A. Dimou et al.
new RML

RML Mapper

linked Data

RML

MDQA violations

Mapping Refinements

(new RML)
DBpedia
Use Case
Sustainable Linked Data Generation: The case of DBpedia: W. Maroy et al.
DBpedia Quality Assessment
Linked Data: 16h
RML rules: 32s

Certain test cases require a complete Linked Data set
new RML

Linked Data

RML Mapper

RML

MDQA

MDQA violations

Mapping Refinements

(optional)

new RML
declaration
execution
assessment

declare RML rules
RML Language
execute RML rules
Linked Data

validate RML rules
RML Validator
RML rules
validated RML rules

validate RML rules
RML rules
execute RML rules
Linked Data
data
declaration
execution
assessment

administrate Linked Data generation workflow

declare RML rules
RML Language

validate RML rules
RML Validator

execute RML rules
RML Mapper

Linked Data
data

RML Mapper
execute RML rules

validated RML rules

RML Language
declare RML rules

RML Workbench
### Overview schedules

<table>
<thead>
<tr>
<th>Date</th>
<th>Mapping file</th>
<th>Output file</th>
<th>#Triples</th>
<th>Publish</th>
<th>Description</th>
<th>Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>July 7th 2016, 11:30:00 am</td>
<td>mapping.rml.ttl</td>
<td>testMpping</td>
<td>1</td>
<td>Yes</td>
<td></td>
<td>Planned</td>
</tr>
</tbody>
</table>
administrate Linked Data generation workflow

RML Workbench

declare RML rules

RML Language

validate RML rules

RML Validator

execute RML rules

RML Mapper

Linked Data

data

RML rules

validated RML rules
Human agents do not need to put in much effort to provide Linked Data
Human agents do not need to put in much effort to provide Linked Data

Intelligent software agents which function with Semantic Web technologies will have enough Linked Data to work with
High Quality Linked Data Generation

Dr Anastasia Dimou
Post-Doc researcher

🌐 imec.be - IDLab.technology
✉️ Anastasia.Dimou@imec.be
🐦 @natadimou