Sustainable Linked Data Generation
The case of DBpedia

Wouter Maroy, Anastasia Dimou, Dimitris Kontokostas, Ben De Meester, Ruben Verborgh, Jens Lehmann, Erik Mannens, Sebastian Hellman

Email: wouter.maroy@ugent.be
DBpedia describes 38.8 million entities.
DBpedia contains more than 3 billion triples.
DBpedia keeps growing, but there are quality issues
There are 2 types of quality issues

- Schema-level
- Data-level
What are the causes?
DBpedia’s Extraction Framework extracts Wikipedia’s infobases
The community created **mapping rules** from infobox properties to a schema.
Causes for the issues can be found in the Extraction Framework (EF)
Causes for the issues can be found in the mapping rules (MR)
Causes for the issues can be found in Wikipedia itself.
Our goal is to adjust the EF & the MR to provide a more sustainable framework.
We integrated a generic, modular, and sustainable mapping language.
The result is a framework that enables sustainable Linked Data generation
DBpedia is making the switch!
Sustainable Linked Data Generation
The case of DBpedia

Before

After

Progress
Limitations of the EF

- Hard-coded mapping rules
- No machine-interpretable mapping rules
- No other ontology
- No schema-validation on mapping rules
Hard-coded mapping rules

Mapping rules define triple generation
Hard-coded mapping rules

Mapping rules only influence the predicate and object due to implementation coupling
Hard-coded mapping rules

This also limits defining the predicate and object
Limitations of the EF

- Hard-coded mapping rules
- No machine-interpretable mapping rules
- No other ontology
- No schema-validation on mapping rules
No machine-interpretable MR

Mapping rules are in **Wikitext format**

The same format that is used for defining Wikipedia articles
No machine-interpretable MR

Mapping rules are in **Wikitext format**

This format cannot be interpreted automatically
No machine-interpretable MR

Mapping rules are in **Wikitext format**

- No querying
- No schema-validation
- No generation
Limitations of the EF

- Hard-coded mapping rules
- No machine-interpretable mapping rules
- Restricted to the DBpedia ontology
- No schema-validation on mapping rules
Restricted to the DBpedia ontology

One ontology is used
Restricted to the DBpedia ontology

That is also coupled with its implementation
Restricted to the DBpedia ontology

What if we want to change the ontology?
Restricted to the DBpedia ontology

All mappings need to be changed manually
Limitations of the EF

- Hard-coded mapping rules
- No machine-interpretable mapping rules
- Restricted to the DBpedia ontology
- No schema validation on mapping rules
No schema validation

Validating the dataset requires many resources

There are other options...

> 3 billion triples
An infobox template for defining persons on Wikipedia

```markdown
<table>
<thead>
<tr>
<th>name</th>
</tr>
</thead>
<tbody>
<tr>
<td>image</td>
</tr>
<tr>
<td>alt</td>
</tr>
<tr>
<td>caption</td>
</tr>
<tr>
<td>birth_name</td>
</tr>
<tr>
<td>birth_date</td>
</tr>
<tr>
<td>death_date</td>
</tr>
<tr>
<td>death_place</td>
</tr>
<tr>
<td>nationality</td>
</tr>
<tr>
<td>other_names</td>
</tr>
<tr>
<td>occupation</td>
</tr>
<tr>
<td>years_active</td>
</tr>
<tr>
<td>known_for</td>
</tr>
<tr>
<td>notable_works</td>
</tr>
</tbody>
</table>
...
```
{{Infobox person
| name          = Bill Gates
| alt           = William H. Gates III in June 2015
| caption       = 
| birth_name    = 
| birth_date   = October 28, 1955
| death_date    = 
| death_place   = 
| nationality   = 
| other_names   = 
| occupation    = 
| years_active  = 1975–present
| known_for     = 
| notable_works = 
}}

Born: William Henry Gates III
Seattle, Washington, US

Residence: Medina, Washington, US

Alma mater: Harvard University

Years active: 1975–present

Net worth: US$85.1 billion
September 2017

Title: Technology Advisor of Microsoft
Co-Chairman of the Bill & Melinda Gates Foundation
CEO of Cascade Investment
Chairman of Branded Entertainment Network
Chairman of TerraPower
Aristotle

Roman copy in marble of a Greek bronze bust of Aristotle by Lysippus, c. 330 BC. The alabaster mantle is modern.

Born: 384 BC
Stagira,
Chalcidice (Chalkidiki),
Chalcidian League,
Northern Greece

Res: 322 BC (aged 62)
Euboea, Greece,
Macedonian Empire

Net: Ancient philosophy
Region: Western philosophy
School: Peripatetic school
Aristotelianism

Main interests: Biology, Zoology
Physics, Metaphysics
Logic, Ethics, Rhetoric
Music, Poetry, Theatre
Politics, Government
<table>
<thead>
<tr>
<th>name</th>
<th>Albert Einstein</th>
</tr>
</thead>
<tbody>
<tr>
<td>image</td>
<td><a href="https://example.com/albert_einstein_image.jpg">https://example.com/albert_einstein_image.jpg</a></td>
</tr>
<tr>
<td>alt</td>
<td>Albert Einstein</td>
</tr>
<tr>
<td>caption</td>
<td>Albert Einstein in 1921</td>
</tr>
<tr>
<td>birth_name</td>
<td>Albert Einstein</td>
</tr>
<tr>
<td>birth_date</td>
<td>14 March 1879</td>
</tr>
<tr>
<td>death_date</td>
<td>18 April 1955 (aged 76)</td>
</tr>
<tr>
<td>death_place</td>
<td>Princeton, New Jersey, U.S.</td>
</tr>
<tr>
<td>nationality</td>
<td>German Empire, German Empire, German Empire, German Empire, German Empire, German Empire, German Empire, German Empire</td>
</tr>
<tr>
<td>other_names</td>
<td></td>
</tr>
<tr>
<td>occupation</td>
<td>Professor of Physics, Physicist, Theoretical Physicist</td>
</tr>
<tr>
<td>years_active</td>
<td>1879-1955</td>
</tr>
<tr>
<td>known_for</td>
<td>Physics, Theoretical Physics, Quantum Mechanics, Relativity Theory</td>
</tr>
<tr>
<td>notable_works</td>
<td></td>
</tr>
</tbody>
</table>

---

Albert Einstein in 1921

Pronunciation: /ˈalbərt ˈaɪntən/ (German: [ˈalbɐrt ˈaɪntɐn])

Born: 14 March 1879, Ulm, Kingdom of Württemberg, German Empire

Residence: Germany, Italy, Switzerland, Austria (present-day Czech Republic), Belgium, United States

Citizenship: Subject of the Kingdom of Württemberg during the German Empire (1879-1896), Stateless (1896-1901), Citizen of Switzerland (1901-1955), Austrian subject of the Austro-Hungarian Empire

School: University of Zürich, ETH Zürich, University of Gottingen, University of Berlin, University of Göttingen
{{Infobox person
| name          =
| image         =
| alt           =
| caption       =
| birth_name    =
| birth_date    =
| death_date    =
| death_place   =
| nationality   =
| other_names   =
| occupation    =
| years_active  =
| known_for     =
| notable_works =
}}

Marie Skłodowska Curie

Born Maria Salomea Skłodowska
7 November 1867
Warsaw, Kingdom of Poland, then part of Russian Empire[1]

Died 4 July 1934 (aged 66)
Passy, Haute-Savoie, France

Cause of death Aplastic anemia

Residence Poland, France
Citizenship Poland (by birth)
France (by marriage)
Alma mater University of Paris

...
| name          | = |
| image         | = |
| alt           | = |
| caption       | = |
| birth_name    | = |
| birth_date    | = |
| death_date    | = |
| death_place   | = |
| nationality   | = |
| other_names   | = |
| occupation    | = |
| years_active  | = |
| known_for     | = |
| notable_works | = |

> 253 000 pages
use the “person” infobox template
Only **one** mapping is responsible for extraction
```json
{{Infobox person
  | name          =
  | image         =
  | alt           =
  | caption       =
  | birth_name    =
  | birth_date    =
  | death_date    =
  | death_place   =
  | nationality   =
  | other_names   =
  | occupation    =
  | years_active  =
  | known_for     =
  | notable_works =
  | Changes at least 250 000 times
  | dbo:name =
  | dbo:givenName =
}}
```
Wrong at least 250,000 times
Validating mappings is more feasible, more efficient & sustainable
No schema validation

But there is no validation for the current mapping rules!
Sustainable Linked Data Generation
The case of DBpedia
A sustainable framework is needed that provides

- ✓ Declarative mapping rules
- ✓ Machine-interpretable format
- ✓ Schema validation
- ✓ Usage of other ontologies
Our solution

The **RDF Mapping language (RML)**

A generic scalable mapping language defined to express rules that map data in heterogeneous structures and serializations to the **RDF** data model
Our solution

The **RDF Mapping language (RML)**

Mapping rules in **RML** are **RDF**
A sustainable framework is needed that provides:

- Declarative mapping rules
- Machine-interpretable format
- Schema validation
- Allows alternative ontology
Declarative mapping rules

Mapping rules are decoupled from their implementation
Declarative mapping rules

- Subject can be mapped
- Subject
- Predicate
- Object

Subject can be defined
Declarative mapping rules

- Subject can be mapped
- Multiple ontologies

Other ontologies can be defined
Declarative mapping rules

Data transformations can be defined

Declarative data transformations for linked data generation: the case of DBpedia
Ben De Meester, Wouter Maroy, Anastasia Dimou, Ruben Verborgh and Erik Mannens
A sustainable framework is needed that provides:

- Declarative mapping rules
- Machine-interpretable format
- Schema validation
- Usage of other ontologies
Machine-interpretable format
Machine-interpretable format

Automated processing with RDF tools

- Querying
- Schema validation
- Generating
A sustainable framework is needed that provides:

- Declarative mapping rules
- Machine-interpretable format
- Schema validation
- Usage of other ontologies
Test-driven Evaluation of Linked Data Quality
Dimitris Kontokostas, Patrick Westphal, Sören Auer, Sebastian Hellmann, Jens Lehmann, Roland Cornelissen

http://rdfunit.aksw.org/
A sustainable framework is needed that provides

- Declarative mapping rules
- Machine-interpretable format
- Schema validation
- Usage of other ontologies
Usage of other ontologies

Mappings don’t work for other ontologies
Usage of other ontologies

Mappings can be changed automatically!
Sustainable Linked Data Generation
The case of DBpedia
Switching to RML

Translating all DBpedia mappings to RML documents
Switching to RML

The Extraction Framework needed to process RML documents
Evaluation: coverage

Complete extraction was done on the English Wikipedia with 98% coverage in comparison with the original dataset.
Evaluation: performance

The framework offers more sustainable mapping rules at a performance cost of 35%.

Next step is optimizing!
Evaluation: flexibility

Mapping rules in RDF can be (automatically) updated

Other datasets can be generated from Wikipedia because of the ontology independency
Evaluation: flexibility

We extracted a dataset of all persons on Wikipedia with the schema.org vocabulary by only changing mapping rules.
A sustainable framework that has

- Declarative mapping rules
- Machine-interpretable format
- Schema validation
- Usage of other ontologies
The future is bright!
Sustainable
Linked Data Generation
The case of DBpedia

Wouter Maroy, Anastasia Dimou, Dimitris Kontokostas,
Ben De Meester, Ruben Verborgh, Jens Lehmann,
Erik Mannens, Sebastian Hellman

wouter.maroy@ugent.be
Made possible by