Assessing the Quality of RDF Mappings with EvaMap

Benjamin Moreau, Patricia Serrano-Alvarado

Assessing the Quality of RDF Mappings with EvaMap

Benjamin Moreau[1,2] and Patricia Serrano-Alvarado[2]

[1] OpenDataSoft {Name.Lastname}@opendatasoft.com
[2] Nantes University, LS2N, CNRS, UMR6004, 44000 Nantes, France
{Name.LastName@}univ-nantes.fr *

Abstract. Linked Data (LD) is a set of best practices to publish reusable data on the web in RDF format. Despite the benefits of LD, many datasets are not published as RDF. Transforming structured datasets into RDF datasets is possible thanks to RDF Mappings. But, for the same dataset, different mappings can be proposed. We believe that a tool capable of evaluating the quality of an RDF mapping would make the creation of mappings easier. In this paper, we present EvaMap, a framework to assess the quality of RDF mappings. The demonstration shows how EvaMap can be used to evaluate and improve RDF mappings.

1 Introduction and Motivation

Linked Data (LD) is a set of best practices to publish reusable data on the web in RDF format. Despite the benefits of LD, many datasets are not published as RDF. Transforming structured datasets into RDF datasets is possible thanks to *RDF Mappings*.

An RDF Mapping is an RDF graph where triples are transformation rules. The set of transformation rules defines the transformation of the initial dataset into an RDF dataset.

Making a relevant RDF mapping for a dataset is a challenging task because it requires to answer several questions:

1. What are the different resources described in the dataset (e.g., cars, persons, cities, places, etc.)?
2. What are the attributes of these resources (e.g., price, age, etc.)?
3. How should the URI of resources be defined?
4. What are the possible relations between the different resources (e.g., the city is the birthplace of the person)?
5. Which ontology, classes, and properties should be used?

In addition to possible errors by the user, different answers are possible for some of these questions and, thus, different RDF mappings are possible for the same dataset.

For example, Figure 1 represents two possible mappings for the dataset in Table 1. Unlike mapping 1(a), mapping 1(b) does not include a class description in resource URIs and does not reference the *Birth Province* column.

Name	Birth	Birth City	Birth Province
Augustus	0062-09-23	Rome	
Caligula	0012-08-31	Antitum	
Claudius	0009-08-01	Lugdunum	Gallia Lugdunensis
...	...	...	...

Table 1. Excerpt from a structured dataset describing roman emperors.

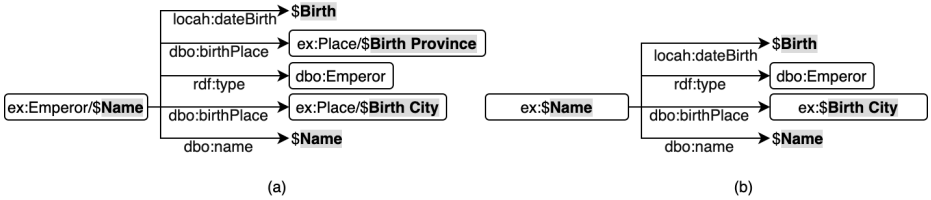

Fig. 1. Two RDF mapping for Roman emperors dataset. Bold text starting with $ are references to a column in the dataset.

How to automatically help users to create RDF mappings without errors and how to choose the best mapping from a set of RDF mappings?

We believe that a tool capable of evaluating the quality of an RDF mapping would make the creation and the choice of RDF mappings easier. [1] proposes a framework that assesses and refines RML mappings. However, this work focuses on logical errors due to incorrect usage of ontologies (e.g., violation of domain, range, disjoin classes, etc.). [3] proposes a framework to assess the quality of RDF datasets through metrics. Metrics are organized in dimensions evaluating different aspects of a dataset (e.g., availability, interlinking, etc.). However, this work does not propose to assess the quality of an RDF mapping. In our work, like in [1], we evaluate metrics on the RDF mapping instead of the resulting RDF dataset. This identifies errors at the beginning of the publishing process and saves time.

Based on the framework proposed in [3], we propose **EvaMap**. EvaMap is a framework to **Eva**luate RDF **Map**pings. The goal is to control the quality of the resulting dataset through its mapping without having to generate the RDF dataset.

2 EvaMap: A Framework to Evaluate RDF Mappings

EvaMap uses a set of metrics organized in 7 dimensions. Each metric is evaluated on the RDF mapping or on the resulting RDF dataset when instances are needed. For example, the *available resource URIs* metric needs RDF dataset to check if generated URIs are dereferenceable. In this case, EvaMap generates a sample such that applying each mapping rule to the entire input dataset is not necessary. Table 2 describes each dimension of EvaMap. These dimensions are based on [3]. We propose only the *Coverability* dimension that detects the lose of data between the input dataset and the resulting RDF dataset. Table 3 describes the new metrics we introduce in EvaMap.

Dimension	Description
Availability	Checks if URIs are dereferenceable
Clarity	Checks human-readability of the mapping and the resulting dataset
Conciseness	Checks if the mapping and the resulting dataset is minimal while being complete
Consistency	Checks if the mapping is free from logical errors
Metadata	Checks metadata quality (license, date, creator, etc.)
Connectability	Checks if links exist between local and external resources
Coverability	Checks if the RDF mapping is exhaustive compared to the initial dataset

Table 2. Dimensions used by EvaMap.

Dimension	Metric	Description
Conciseness	Redundant rule	Checks if multiple rules generate the same triples
Connectability	Local links	Checks if resources described in the mapping are connected
Consistency	Datatype inheritance	Checks if datatype in the mapping correspond to datatype in the initial dataset
Metadata	License compatibility	Checks if the license of the initial dataset is compatible with the license of the resulting dataset

Table 3. New metrics proposed in EvaMap.

In order to compute the quality of a mapping, M_i applied on a raw dataset D, we propose a function $q(M_i, D) \in [0, 1]$ that is the weighted mean of the quality of each metric $m_j(M_i, D)$:

$$q(M_i, D) = \frac{\sum_{j=1}^{n} w_j . m_j(M_i, D)}{\sum_{j=1}^{n} w_j}$$

Weights w_j associated with metrics can be used to give more or less importance to each metric.

We use the same function to compute the score for a specific dimension. To do that, we only consider the subset of metrics for the corresponding dimension.

3 Demonstration

We implemented EvaMap to evaluate YARRRML [2] mappings for datasets of the OpenDataSoft's data network[3]. Our tool is available as a web service at `https://evamap.herokuapp.com/`. The source code of our tool[4] and web service[5] are available on GitHub under the MIT license.

During the demonstration, attendees will be able to select different mappings and use EvaMap to compare them. For each mapping, the global quality score will be computed as well as the quality score for each dimension. Our tool will also give feedback to improve RDF mapping.

In our tool, users can assess two mappings for the dataset *football-ligue*. Users can see that the mapping *football-ligue* obtains a worse global score than the mapping *football-ligue-fixed*. In the detailed report, users can analyze by dimension why these scores are different.

Acknowledgments Authors thank Chanez Amri, Alan Baron, and Marion Hunault (Master students of the University of Nantes) for their participation in this work.

References

1. Dimou, A., Kontokostas, D., Freudenberg, M., Verborgh, R., Lehmann, J., Mannens, E., Hellmann, S., Van de Walle, R.: Assessing and Refining Mappings to RDF to Improve Dataset Quality. In: International Semantic Web Conference (ISWC) (2015)
2. Heyvaert, P., De Meester, B., Dimou, A., Verborgh, R.: Declarative Rules for Linked Data Generation at Your Fingertips! In: Extended Semantic Web Conference (ESWC), Poster&Demo (2018)
3. Zaveri, A., Rula, A., Maurino, A., Pietrobon, R., Lehmann, J., Auer, S.: Quality Assessment for Linked Data: a Survey. Journal of Semantic Web (2016)

[3] `https://data.opendatasoft.com`
[4] `https://github.com/benjimor/EvaMap`
[5] `https://github.com/benjimor/EvaMap-Web`