POLITECNICO DI TORINO

Dipartimento di Automatica e Informatica

PhD in Computer and Control Engineering

XXX cycle

Supervisor

*Prof. Marco Torchiano*

# Automated Quality Assessment and Validation for Evolving Knowledge Bases

PhD Candidate: *Mohammad Rifat Ahmmad Rashid*

## 1. Introduction

**Data quality** is one of the main issues for curating data on the Web. For example, the most common data quality issues are:

- **Persistency:** unexpected removal of information.
- **Completeness:** incomplete information.
- **Consistency:** facts being conflicting or contradictory**.**

**Linked Open Data** approach consists in exposing and connecting data from different sources on the Web.

**Knowledge Base** (**KB**) is a technology used to store both complex structured and unstructured, information representing domain knowledges.

**Resource Description Framework (RDF)** is a graph-based data model which is widely used in Linked Data applications.

**RDF shape** is a formal syntax for describing **how data is** and **how data should be.**

## 2. Goal

The goal of this Ph.D. is to **analyze and benchmark** the data quality issues in any knowledge base. In particular, we aim at answering the following research questions:

- *How to automatically detect and measure data quality issues using changes observed across different KB releases?*
- *How to automatically generate integrity constraints for RDF shape validation?*

## 3. Approach

The overall approach is showed in Figure 1. It consists of three main contributions:

- **KBQ:** A novel **K**nowledge **B**ase **Q**uality assessment approach which compares consecutive releases to compute quality measures that allow detecting quality issues [1].
- **RDF Shape Induction:** An automatic approach for **RDF** Shapes generation for a given dataset by applying machine learning techniques [2].
- **KBQ-Tool:** A tool that automates the detection and generates reports of quality issues for any knowledge base [2].

## 4. Results

We motivated this work by exploring the concept of Linked Data dynamics in the aspects of Knowledge Base quality analysis. We evaluated the KBQ approach both quantitatively and qualitatively on a series of releases from three knowledge bases, namely: DBpedia EN, DBpedia ES, and 3cixty. The study showed that the KBQ approach is extremely effective and it is able to achieve 95% precision in error detection. The results of the experiment are shown in Table 1.

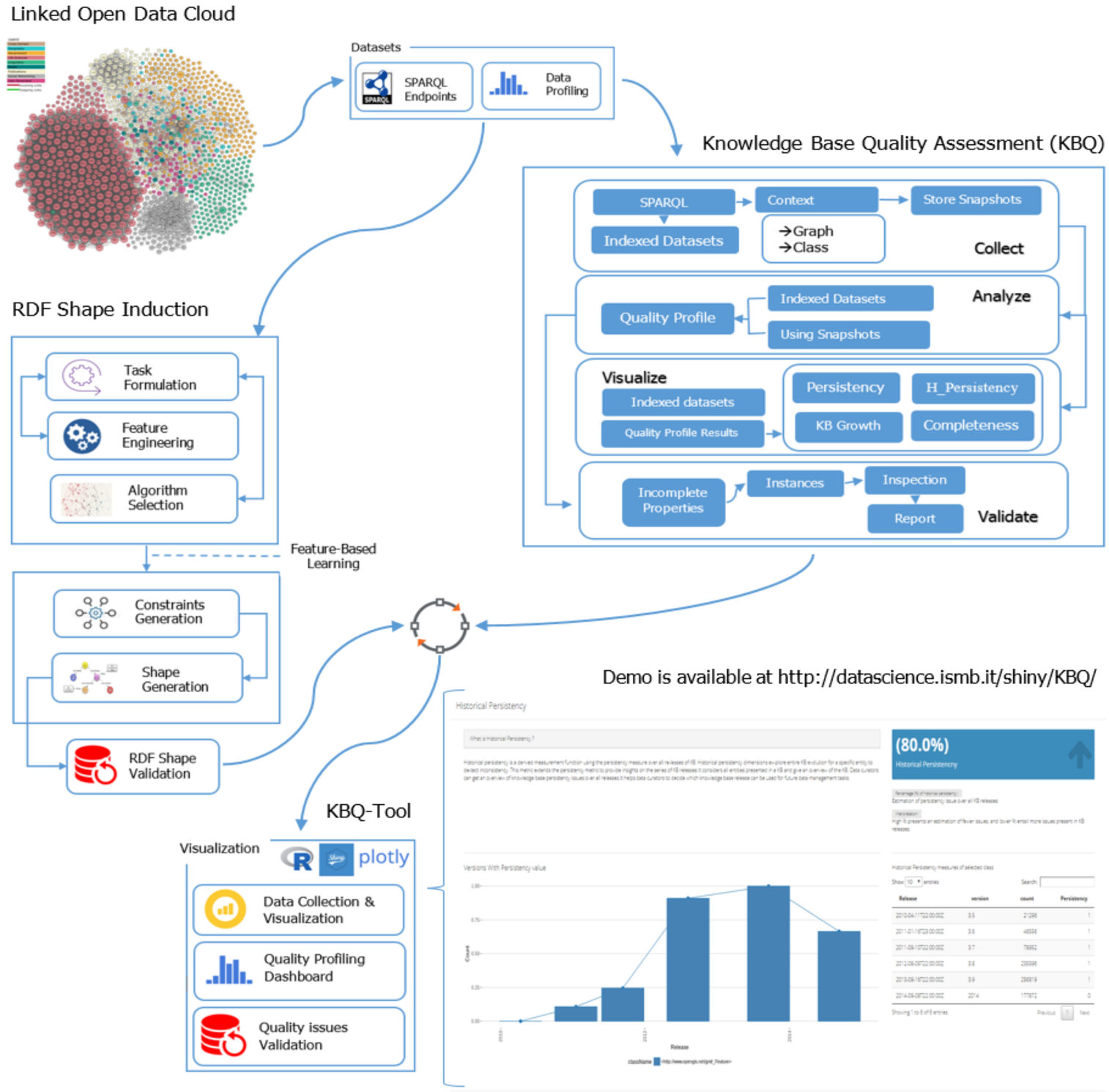


Figure 1: Overall Approach

| Knowledge Bases | Performance |
|---|---|
| | Precision |
| DBpedia EN | 95% |
| 3cixty | 94% |
| DBpedia ES | 89% |

Table 1: Evaluation result of KBQ approach

We evaluated our RDF shape induction approach based on cardinality constraints and range constraints from DBpedia KB.

| Machine Learning Models | DBpedia KB | |
|---|---|---|
| | Minimum Cardinality | Maximum Cardinality |
| Baseline | 76.4% | 52.9% |
| Random Forest | 97.61% | 87.29% |
| Logistic Regression | 92.21% | 84.74% |
| K-NN | 92.64% | 85.5% |
| SVM | 94.84% | 88.67% |

Table 2: Evaluation result of our RDF Shape Induction approach

The study showed that, when compared with the baseline model, our approach can significantly improve the prediction accuracy. The results based on cardinality constraints are shown in Table 2.

## 5. Conclusions

- The KBQ approach is Knowledge Base agnostic and delivered good performance for three different KBs.
- Our RDF Shape Induction approach is based on machine learning techniques that make the solution both flexible and scalable.

## 6. References

1. Rashid, M.; Torchiano, M.; Rizzo, G.; Nandana M.; Corcho, O.: Knowledge Base Quality Assessment Using Temporal Analysis. In: Semantic Web Journal, 2017. [Under Review]
2. Rashid, M.; Torchiano, M.; Rizzo, G.; Nandana M.; Corcho, O.: Automated Quality Assessment and Validation for Evolving Knowledge Bases. In: Journal of Web Semantics, 2017. [Under Review]