Synonym Predicate Discovery for Linked Data Quality Assessment Without Requiring the Ontology Semantic Relations

Samah Salem
Lire Laboratory, dept. of TLSI
Abdelhamid Mehri- Constantine 2 University
Constantine, Algeria
E-mail: samah.salem@univ-constantine2.dz

Fouzia Benchikha
Lire Laboratory, dept. of TLSI
Abdelhamid Mehri- Constantine 2 University
Constantine, Algeria
E-mail: fouzia.benchikha@ univ-constantine2.dz

***Abstract*—Over the past years, an increasing number of datasets have been published as part of the Web of Data, reaching more than 1,200 datasets in 2019. However, many datasets, totaling a large quantity of RDF triples, are without ontology or with an incomplete one. As a result, they suffer more and more from quality problems. Assessing linked data quality for fitness for use is a current research problem that we are interested in. In this paper, we propose a novel approach for the assessment of quality between RDF triples without requiring schema information. It allows assessing the quality of datasets by detecting errors and eventually measuring the error rate using synonym predicates techniques, profiling statistics, and quality verification cases. Promising results are obtained on the DBpedia dataset where several data quality issues have been detected, such as inaccurate values, redundant predicates, and redundant triples.**

Keywords-linked data; quality assessment; semantic relations; synonym predicates; profiling statistics; DBpedia.

I. INTRODUCTION

In the last decade, the number of datasets published in the Linked Data (LD) format had increased from 12 datasets in May 2007 to 1,239 datasets in March 2019. This huge growth leads to the appearance of many structured datasets on the Web of Data [18], such as DBpedia [19] and Wikidata [20]. However, many of these datasets do not have a well-developed ontology or do not have an ontology at all, and their qualities are highly variable, as in the case of DBpedia that is considered as the most well organized and widely used LD resource [2].

In the literature, data quality is usually defined as "fitness for use". It depends on several dimensions, such as accuracy, completeness, relevance, credibility, comprehensibility, consistency, and conciseness [1]. Several authors have proposed interesting approaches for quality assessment requiring ontology for datasets, which is not always available or may be incomplete. New approaches are thus required to deal with LD quality assessment by finding features that best represent the semantics of Resource Description Framework (RDF) triples without requiring ontologies, when each triple represents two entities (Subject and Object) linked with (Predicate). To achieve this goal, we propose an approach for quality assessment between RDF triples, independently of the semantic relationships of the ontology, using both techniques of synonym predicates discovery, profiling statistics, and predefined quality verification cases.

The remainder of this paper is structured as follows: in Section II, we discuss the related work. Section III presents our proposed approach. An evaluation is given in Section IV. Finally, we conclude with ideas for future work in Section V.

II. RELATED WORK

In this section, we present related work on quality assessment in the web of data as well as existing approaches for synonym predicates discovery.

A. Linked Data Quality Assessment

Several works on the quality assessment of linked data have been proposed. They focused on assessing the quality of different parts of datasets, namely literals, predicates, triples, and metadata. We present here the well-known methodologies and tools, which could be classified into two distinct categories: (1) those that use ontologies and (2) those that do not use ontologies.

In the first category, several approaches are proposed. Lei et al. [5] propose a framework that allows evaluating the accuracy, consistency, and conciseness of semantic metadata. SWIQA [4] allows automatically evaluating the quality of published data using a quality rule template. In addition, RDFUnit [3], a pattern-based approach for LD quality assessment, uses data schema and quality patterns are created from DBpedia user community feedback, Wikipedia maintenance system, and ontology analysis. Besides, another approach called ABSTAT [6], allows the use of data profiling and data mining techniques to explore LD and to detect quality issues at the schema level. Finally, a semi-automatic methodology for dataset quality assessment and improvement is proposed in [14]. Although, the previous works provide good support for LD quality assessment, none of them is focused on detecting errors by discovering semantic relations between properties in the dataset (that lacks a well-developed ontology or does not have ontology at all). Therefore, there is still a need for additional researches and efficient techniques to provide high quality for LD that do not require a lot of user expertise and ontology information.

 ISBN: 978-1-61208-775-7

In contrast to the first category, in the second one, the most significant research work consists of the approach proposed by Jang et al. [2], which assesses the LD quality without using any data schema. It measures the quality of LD in terms of property's domain, range and data type through a semi-automatic generation of data quality patterns. The approach has been applied to Korean DBpedia, in which an error occurrence rate equal to 36.31% has been obtained. It seems to be an interesting approach, which will open new possibilities for researchers to develop efficient techniques for LD quality assessment without using data schema information. However, the quality assessment is done with only one triple and it does not give the exact domain/range (i.e., the generation of an upper-class type). Moreover, no quality improvement after detecting quality problems is incorporated.

In the context of our work, we consider datasets without ontologies. We propose an approach for LD quality assessment by understanding semantics between properties and considering assessing quality between triples. Table I gives a comparative study.

The proposed approach is based on synonym predicates discovery to efficiently assess data quality through detecting errors between triples. The next subsection will present some existing techniques of synonym predicates discovery that has been used for different purposes.

B. Synonym Predicates Discovery

In the literature, some work use synonym predicates discovery techniques in LD. For instance, Abedjan and Naumann [8] propose an approach that allows discovering synonymously used predicates. The main objective is to expand queries, by aggregating positive and negative association rules at the statement level based on the concept of mining configurations. However, it discovers only predicates that could substitute each other, such as *starring* and *artist*, which is usually not suitable since the predicate expansion operation is different from the predicate unification operation.

Another work for knowledge graph consolidation is proposed in [9]. It is a data-driven method to identify existed synonymous relationships in the knowledge graph using knowledge embedding methods, such as RESCAL [11], ComlEX [12], and ANALOGY [13], and without making any assumptions on the data.

In addition, Issa [10] proposed an approach to assess the completeness and the conciseness of LD. It is based on Abedjan et al. [8] approach, in which synonymous relationships are used to detect redundant predicates in datasets and so to ensure their conciseness.

Broadly, in the existing approaches, the synonym predicates are used for query expansion [8], graph consolidation [9], and redundancy detection [10], but in our approach, we discover the synonym predicates for a holistic detection of quality issues at subject-level, predicate-level, and object-level. Since in our opinion, the discovery of synonyms may reveal several problems in the data. As well, the methods used for the discovery of synonyms are different from our natural language processing method. Table II highlights their main limitations compared with our approach. The next section will give more details on the proposed approach.

TABLE I. COMPARISON BETWEEN LINKED DATA QUALITY ASSESSMENT APPROACHES.

Approaches	Goal	Quality of	Quality dimensions	With/ without ontology
Lei et al., 2007	Quality assessment of semantic metadata	Metadata	Accuracy, consistency, conciseness	With ontology
Fürber and Hepp, 2011	Quality assessment of published data	Literal	Accuracy, completeness, uniqueness, timeliness	With ontology
Kontokostas et al., 2014	DBpedia quality assessment	Triple	-	With ontology
Spahiu et al., 2016	Summarize the content of a dataset and reveal data quality problems	Predicate	Accuracy, completeness, timeliness	With ontology
Jang et al., 2015	Linked data quality assessment	Triple	Accuracy and consistency	Without ontology
Our approach	Assess the quality between RDF triples Understand the semantics between properties	Predicate, object, triple	Accuracy and conciseness	Without ontology

TABLE II. COMPARISON BETWEEN OUR APPROACH AND SYNONYM PREDICATE DISCOVERY APPROACHES.

Approaches	Goal	Based on	Techniques
Abedjan and Naumann, 2013	Query expansion	Synonymously used predicates	Association rules mining
Issa, 2018	Dataset conciseness	Synonymously used predicates	Abedjan and Naumann. [8] approach
Kalo et al., 2019	Graph consolidation	Synonym predicates	Knowledge embedding
Our approach	Measure the accuracy and the conciseness of the datasets that do not have an ontology	Synonym predicates	Natural language processing-based methods

III. THE PROPOSED APPROACH

We propose a novel approach for the assessment of quality between RDF triples without requiring schema information. The approach consists of three main steps (as shown in Figure 1): (1) synonym predicates discovery, (2) profiling statistics generation, and (3) quality assessment. It assesses the quality of datasets by detecting errors and eventually measuring the error rate using synonym predicates techniques, profiling statistics, and quality verification cases.

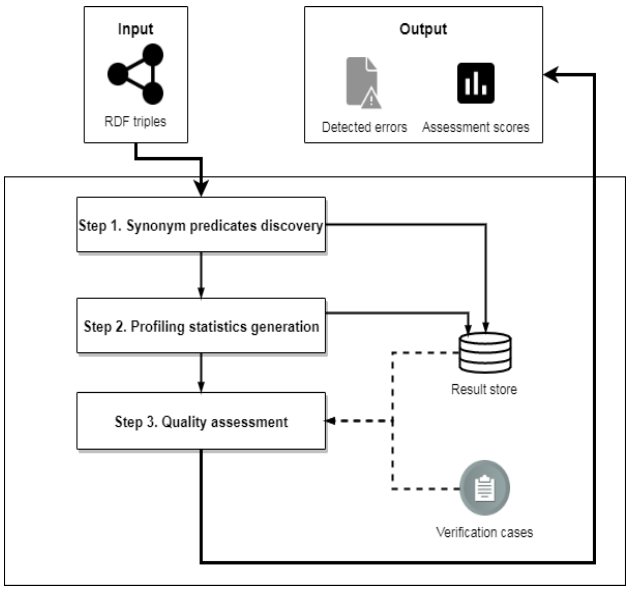

Figure 1. A three-step approach for quality assessment.

A. Step 1. Synonym Predicates Discovery

In a dataset without schema, there is no definition of entities, data types, and semantics of the properties. However, the possibility of finding two or more predicates, which have the same meaning is very high (as revealed after a study on DBpedia, for example *foaf:nick* and *dbp:nickname*). For this purpose, we are interested particularly in synonym predicates discovery for the creation of synonym-pattern (cf. III.B) and for the detection of quality problems (cf. III.C).

Our research on discovering the synonym predicates is based on the natural language processing methods. Indeed, as it is known that the web of data uses complex identifiers for naming predicates and not literals, then we adapt the natural language processing methods to our validation needs. An RDF graph G is a set of triples T s.t. G = {T}, where each triple T has the form of subject, predicate, and object s.t. T = (s,p,o).

$$\exists\, o_i, \exists\, s_i \mid p_i(o_i, s_i) \quad (1)$$

$$\exists\, o_j, \exists\, s_j \mid p_j(o_j, s_j) \quad (2)$$

As in our case, we are interested in the errors that occur between triples, therefore we will focus on the discovery of the synonym predicates. (3) Gives a predicate p_i of triple t_i and predicate p_j candidate synonym of triple t_j.

$$p_i \in t_i \wedge p_j \in t_j \mid p_i \equiv_{syn} p_j \quad (3)$$

We focus on the thesaurus-based methods, WordNet, due to their high precision in the synonym identification that is necessary in our case study. However, there are several synonyms that are not indexed by WordNet, and the problem of the predicates with spelling errors that are not detected by the WordNet, such as *dbp:birthPace*, *dbp:birthPaxes*, and *dbp:nbirthPlace*. For these reasons, we use a check spelling method [21] that suggests corrections for misspelt words based on many popular spells checking packages, such as Ispell [15], Aspell [16], and MySpell [17]. This is a semi-automatic step, and a user (not necessarily a domain expert) must confirm each detected synonym pair.

B. Step 2. Profiling Statistics Generation

Data profiling is about examining and collecting information from datasets. In our approach, it is very useful to generate some profiling tasks to prepare for quality score estimation. The principal goal of this step is to generate synonym-pattern based on the results of the first step, and to calculate simple profiling statistics, such as the total number of triples in a dataset, and the property occurrence (i.e., how many times the property defined as a synonym occurs in the dataset). The synonym-pattern is a summary that provides a global view of the synonym predicates existing in the dataset and the predicate frequency. A predicate-pattern has the following form:

$$<p_i\,(\Sigma p_i) \equiv_{syn} p_j(\Sigma p_j) \equiv_{syn} p_n(\Sigma p_n)>. \quad (4)$$

For example, we can have as result <dbo:birthplace (13), dbp:birthCity (2)>, where the pattern shows two predicates synonym (dbo:birthplace and dbp:birthCity) with the frequency of each predicate (13 and 2 respectively) in the dataset.

C. Step 3. Quality Assessment

In the previous steps, we generated the synonym predicates and the profiling statistics. This step involves the actual quality assessment including: (1) the detection of quality problems that may occur between RDF triples, and (2) the estimation of quality scores. For the first task, we will use the synonym predicates defined in the first step, and predefined quality verification cases (more details are provided below). For the second task, we will use the profiling statistics generated in the second step for the estimation of quality scores. Note that, in this first version of the proposed approach, we allow just to reveal the errors existing between RDF triples, in the future, we will incorporate the treatment of errors once identified.

 ISBN: 978-1-61208-775-7

1) Quality Problems Detection: In order to detect quality issues, we will verify the similarity or the difference between the subject and the object of each predicate synonyms pair to detect the errors between RDF triples. Note that there are only four possible cases that could occur between two triples.

a) Case 01:

$$\text{If } s_i = s_j \wedge o_i = o_j \Rightarrow \{p_i(o_i,s_i) \Leftrightarrow p_j(o_j,s_j)\}. \quad (5)$$

If the synonym predicates p_i and p_j have the same subject and the same object, then the triple t_i is equivalent to the triple t_j, which mean that one of these triple t_i or t_j is a redundant one (see TABLE IV).

b) Case 02:

$$\text{If } s_i = s_j \wedge o_i \neq o_j \Rightarrow \{p_i \Leftrightarrow p_j\}. \quad (6)$$

If the synonym predicates p_i and p_j have the same subject and totally different object, then (see TABLE IV), there are two types of errors:

- The predicate p_i is equivalent to the predicate p_j, which means that two predicates having the same meaning are defined differently in graph G, thus duplicating the information (i.e., redundant terms to represent the same predicate).
- We can ensure that the object value o_i and/ or o_j is an inaccurate value.

c) Case 03:

$$\text{If } s_i \neq s_j \wedge o_i = o_j \Rightarrow \{p_i \Leftrightarrow p_j\}. \quad (7)$$

If the synonym predicates p_i and p_j have different subject and the same object, then, it is possible to find two types of errors (see TABLE IV):

- The predicates p_i and p_j are defined differently, despite that they had the same meaning, since their equivalence.
- We can *assume* that the object value o_i and/ or o_j is an inaccurate value. If the predicate must contain a unique object value, then, we can *ensure* that the object value o_i and/ or o_j is an inaccurate value.

d) Case 04:

$$\text{If } s_i \neq s_j \wedge o_i \neq o_j \Rightarrow \{p_i \Leftrightarrow p_j\}. \quad (8)$$

If the synonym predicates p_i and p_j have different subjects and different objects, then, we can say that in this case there is duplicate information in order to define the same predicate in the dataset (see TABLE IV).

2) Quality Scores Estimation: After detecting the abnormal triples, it is suitable to measure the quality in terms of numbers. Based on the data quality score metrics [4][14] and the generated profiling statistics, the quality scores according to our needs are calculated, in particularly quality score (QScore), accuracy (Acc-QS), and conciseness (Co-QS). For instance, QScore is the ratio between the number of abnormal triples A_t and the total number of triples T_t, as the following formula shows:

$$\text{QScore} = A_t / T_t. \quad (9)$$

In addition, in order to differentiate between the detected errors, we calculate Acc-QS to measure the percentage of inaccurate values, and Co-QS for duplicate predicates and triples.

$$\text{Acc-QS} = PA_t / A_t. \quad (10)$$

$$\text{Co-QS} = PC_t / A_t. \quad (11)$$

Where PA_t is the number of inaccurate values, and PC_t represents the number of redundant predicates plus the number of redundant triples. The obtained results present the accuracy/ conciseness error occurrence rate compared to the total number of errors in the dataset.

IV. VALIDATION

In order to evaluate our proposed approach, which is available on GitHub repository [22], several studies are carried out on the latest version of DBpedia released in 2019. The experiment revealed several cases of unknown synonymous relationships. Table III illustrates some synonym pairs discovered by applying our approach to entities of type *Person*. Quality problems between triples are detected as shown in Table IV. We used properties of 449 triples, and we found 50 abnormal triples that present an error rate equal to 11 %. In order to better evaluate the performance of the proposed approach, it will be applied to even larger and more complex datasets (which is left for future work).

TABLE III. TOP 5 OF SYNONYM PAIRS.

DBpedia Person	
foaf:name	dbp:name
dbo:birthplace	dbp:birthCity
dbo:birthDate	dbp:birthdate
foaf:gender	dbo:gender
dbo:occupation	dbp:occupation

The abnormal triples may contain several errors, such as redundant predicates, redundant triples, and inaccurate values. Through the detection of these errors, we could measure two quality dimensions, namely accuracy, and conciseness of the dataset. Note that we omit the blank node from our approach and leave it for future work.

 ISBN: 978-1-61208-775-7

TABLE I. QUALITY ISSUES DETECTED BETWEEN TRIPLES ON DBPEDIA.

Triples pairs with synonym predicates	Error type	Quality dimension
dbr:Duduka_da_Fonseca, **dbo:birthplace**, dbr:Rio_de_Janeiro dbr:Duduka_da_Fonseca, **dbp:birthCity**, dbr:Rio_de_Janeiro	***Case 01:*** The results show that the two triples are equivalent, which means that one of these two triples is redundant.	*Conciseness*
dbr:Paulie_Pennino, **foaf:gender**, "female"@en dbr:Paulie_Pennino, **dbo:gender**, dbr:Male	***Case 02:*** The sex of the entity **dbr:Paulie_Pennino** is inaccurate in one of these two triples since once is defined as "female", and once is defined as dbr:Male	*Accuracy/ Conciseness*
dbr:Cornelia_(wife_of_Caesar), **dbp:diedPlace**, dbr:Rome dbr:Aloysius_Lilius, **dbo:deathPlace**, dbr:Rome	***Case 03:*** The predicates **dbp:diedPlace** and **dbo:deathPlace** are defined differently despite that they have the same meaning	*Conciseness*
dbr:Alice_Walker, **foaf:gender**, "female"@en dbr:Zack_Addy, **dbo:gender**, dbr:Male	***Case 04:*** In this case, there is duplicate information in order to define the same predicate in the dataset	*Conciseness*

V. CONCLUSION

The Web of Data allows publishing data that includes its semantics using shared vocabularies and data annotations described in ontologies [4]. Unfortunately, there are a large number of datasets without ontology or with an incomplete one. Therefore, it is necessary to generate ontologies from the target LD. However, constructing an ontology for a large amount of data that may contain quality problems is a difficult and time-wasting task. For these reasons, we propose an approach based on the discovery of the semantic links between properties to assess the quality of RDF triples without requiring the existence of the ontology information. This work guides the users to evaluate the quality between RDF triples through the discovery of synonym predicates and the generation of profiling statistics, and predefined quality verification cases.

Similar to [9][10] approaches, we are interested in the discovery of the synonym predicates, but in our approach we work with RDF triples without using the ontology information. We present the discovered synonym predicates as a synonym-pattern in order to (i) understand the semantics between properties, (ii) detect quality problems and (iii) estimate the quality scores. Our approach allows to efficiently detecting the errors between RDF triples without using the ontology information at all. The obtained results show that there is an important number of inaccurate values in the DBpedia dataset, as well as, duplicate predicates due to the usage of synonym predicates discovery. Despite the fact that the proposed approach shows interesting results in the field of quality problem detection, some exceptions will be handled in the future. For example, when the predicate values are represented with different patterns, such as (dbr:Julius_Caesar, dbo:birthdate, '−100 - 07 - 13') and (dbr:Julius_Caesar, dbo:birthdate, '− 100 - 7 - 13') these triples are identified in Case 02, however, they should be identified in Case 01.

For further work, we intend to define more varied metrics for linked data quality assessment mainly for dataset without ontology. We plan to improve the quality of data and to improve the performance of our approach through the treatment of the blank node identifiers.

REFERENCES

[1] A. Zaverii et al., "Quality assessment for linked data: A survey," Semantic Web, 7(1), pp. 63-93, january 2016.

[2] S. Jang, M. Megawati, J. Choi, and M. Y. Yi, "Semi-automatic quality assessment of linked data without requiring ontology," In NLP-DBPEDIA@ ISWC, pp. 45-55, October 2015.

[3] D. Kontokostas et al., "Test-driven evaluation of linked data quality," In Proceedings of the 23rd international conference on World Wide Web, pp. 747-758, ACM, April 2014.

[4] C. Fürber and M. Hepp, "Swiqa-a semantic web information quality assessment framework," In ECIS, Vol. 15, pp. 19-31, 2011.

[5] Y. Lei, V. Uren, and E. Motta, "A framework for evaluating semantic metadata," Proceedings of the 4th international conference on Knowledge capture, ACM, pp. 135-142, October 2007.

[6] B. Spahiu, "Profiling the Linked (Open) Data," International Semantic Web Conference, Vol. 1491, October 2015.

[7] B. Spahiu, R. Porrini, M. Palmonari, A. Rula, and A. Maurino, "ABSTAT: ontology-driven linked data summaries with pattern minimalization," In European Semantic Web Conference, pp. 381-395. Springer, Cham, 2016.

[8] Z. Abedjan and F. Naumann, "Synonym analysis for predicate expansion." In Extended semantic web conference, pp. 140-154. Springer, Berlin, Heidelberg, May 2013.

[9] C. Kalo, P. Ehler, and W. T. Balke, "Knowledge Graph Consolidation by Unifying Synonymous Relationships," International Semantic Web Conference. Springer, pp. 276-292, Cham, October 2019.

[10] S. Issa, "Linked Data Quality," In DC@ ISWC, pp. 37-45, 2018.

[11] M. Nickel, V. Tresp, and H. P. Kriegel, "A Three-Way Model for Collective Learning on Multi-Relational Data," In ICML, vol. 11, pp. 809-816, June 2011.

[12] T. Trouillon, J. Welbl, S. Riedel, E. Gaussier, and G. Bouchard, "Complex embeddings for simple link prediction," In International Conference on Machine Learning, pp. 2071-2080, June 2016.

[13] H. Liu, Y. Wu, and Y. Yang, " Analogical inference for multi-relational embeddings." In Proceedings of the 34th International Conference on Machine Learning-Volume 70, pp. 2168-2178, JMLR. org, August 2017.

[14] A. Rula and A. Zaveri, "Methodology for Assessment of Linked Data Quality," In Proceedings of the 1st Workshop on Linked Data Quality co-located with 10th International

Conference on Semantic Systems, LDQ@SEMANTiCS 2014, Leipzig, Germany, September 2nd, 2014.

[15] R. E. Gorin, P. Willisson, W. Buehring, and G. Kuenning, "Ispell, a free software package for spell checking files," The UNIX community, 1971.

[16] K. Atkinson, "GNU Aspell," 2003, URL http://aspell. Net, 2011.

[17] C. Andrea, "My spell-checker's «weigh» with words," The Christian Science Monitor, August 2002.

[18] The Linked Open Data Cloud. [Online]. Available from: https://lod-cloud.net/, last accessed: December 19,2019.

[19] DBpedia. [Online]. Available from: https://wiki.dbpedia.org/, last accessed: December 19,2019.

[20] Wikidata. [Online]. Available from: https://www.wikidata.org/wiki/Wikidata:Main_Page, last accessed: December 19,2019.

[21] Spellchecking library for Python. [Online]. Available from: https://github.com/pyenchant/pyenchant, last accessed: December 19,2019.

[22] https://github.com/SalemSamah/SPDiscovery.

 ISBN: 978-1-61208-775-7