Can You Find All the Data You Expect in a Linked Dataset?

Walter Travassos Sarinho[1], Bernadette Farias Lóscio[1] and Damires Souza[2]
[1]Center for Informatics, Federal University of Pernambuco, Recife, Brazil
[2]Academic Unit of Informatics, Federal Institute of Education, Science and Technology of Paraiba, João Pessoa, Brazil

Keywords: Linked Dataset, Quality Information, Completeness Assessment.

Abstract: The huge volume of datasets available on the Web has motivated the development of a new class of Web applications, which allow users to perform complex queries on top of a set of predefined linked datasets. However, given the large number of available datasets and the lack of information about their quality, the selection of datasets for a particular application may become a very complex and time consuming task. In this work, we argue that one possible way of helping the selection of datasets for a given application consists of evaluating the completeness of the dataset with respect to the data considered as important by the application users. With this in mind, we propose an approach to assess the completeness of a linked dataset, which considers a set of specific data requirements and allows saving large amounts of query processing. To provide a more detailed evaluation, we propose three distinct types of completeness: schema, literal and instance completeness. We present the definitions underlying our approach and some results obtained with the accomplished evaluation.

1 INTRODUCTION

The Web has evolved into an interactive information network, allowing users and applications to share data on a massive scale. To help matters, the Linked Data principles define a set of practices for publishing structured data on the Web aiming to provide an interoperable Web of Data (Heath and Bizer, 2011). These principles are based on technologies such as HTTP, URI and RDF (Heath and Bizer, 2011). By using the RDF model, it is possible to publish data or resources on the Web in the form of triples (composed by a subject, a predicate and an object), where each resource is individually identified by means of URIs. In addition, ontology languages such as RDFS and OWL provide means for the creation of vocabularies to structure data domains.

The huge volume of datasets available on the Web has motivated the development of a new class of Web applications, which allow users to perform complex queries on top of a set of predefined datasets instead of performing simple queries through search engines. However, given the large number of available datasets and the lack of data quality and provenance information about them, the selection of datasets for a particular application may become a very complex task.

In this sense, in this paper, we are interested on how to evaluate the completeness of linked datasets to help the selection of datasets for a particular application. Completeness is a known Information Quality (IQ) criterion, usually defined on datasets in terms of the degree to which information is not missing (Pipino et al., 2002). Completeness is also considered as a contextual IQ criterion since it is usually measured within the context of a task at hand (Wang and Strong, 1996). This criterion is further classified into the following categories (Mendes et al., 2012): schema completeness and data completeness. On the former, a dataset is complete if it contains all of the attributes needed for a given task. On the latter, a dataset is complete if it contains all of the necessary objects for a given task.

Considering a given application A, the task at hand consists of providing the data that users expect to obtain when using the application A. In other words, the task at hand consists of attending to user requirements in terms of data, called as data requirements. Given that data requirements are represented through a set of queries Q, completeness of a given dataset D could be evaluated through the execution of queries from Q on D and then analyzing the corresponding results in order to verify

Travassos Sarinho W., Farias Lóscio B. and Souza D..
Can You Find All the Data You Expect in a Linked Dataset?.
DOI: 10.5220/0005381806480655
In *Proceedings of the 17th International Conference on Enterprise Information Systems* (ICEIS-2015), pages 648-655
ISBN: 978-989-758-097-0

if data provided by *D* meets the data requirements. However, this approach is not suitable when dealing with large amounts of data, i.e., large number of datasets and datasets with large amounts of data.

In this paper, we propose an approach to completeness assessment of linked datasets, which is suitable for large amounts of data. The proposed approach uses information extracted from *Q* to evaluate both schema and data completeness, and it doesn't require the execution of time consuming queries. To provide a more detailed evaluation, we propose two distinct types of data completeness: literal and instance completeness.

In order to evaluate our approach, a tool was implemented and some experiments were performed. The accomplished evaluation shows that our approach is able to produce similar results to the ones produced when considering a conventional approach. As a conventional approach we mean when the user has to submit queries over each dataset (by means of its endpoint) individually and analyze if that dataset meets his/her data requirements. This is usually a hard and time-consuming task.

The remainder of this paper is organized as follows: Section 2 introduces information quality concepts; Section 3 presents our approach for linked datasets completeness assessment; Section 4 describes some experiments performed to evaluate our proposal; Section 5 discusses related work, and Section 6 points out some conclusions and indicates future work.

2 IQ AND THE WEB OF DATA

There has been an exponential growth in the availability of linked datasets on the Web and in the development of applications for querying and consuming these data. Due to that, the concept of Information Quality (IQ) is becoming more and more a necessity, instead of an optional requirement.

The notion of IQ has emerged during the past years and shows a steadily increasing interest. IQ is based on a set of dimensions or criteria. The role of each one is to assess and measure a specific quality aspect (Wang and Strong, 1996). In general, IQ researchers assume that there are some shared norms of quality, or quality expectations, and ways of measuring the extent of meeting those norms and expectations. For our purposes, we use the general definition of IQ – 'fitness for use' – that encompasses different aspects of quality (Wang and Strong, 1996; Zaveri et al., 2012).

It is important to distinguish the two concepts of Data Quality and Information Quality. IQ is a term to describe the quality of any element or content of information systems (Wang and Strong, 1996), not only the data. IQ assurance is the certainty that particular information meets some quality requirements. This leads us to think in a service-based perspective of quality, which focuses on the information consumer's response to his/her task-based interactions with the information system. The use of the term information rather than data implies that the use and delivery of the data must be considered in any quality judgments, i.e., the quality of delivered data represents its value to information consumers (Price and Shanksa, 2005). Thus, we use the definition of Information Quality as a set of criteria to indicate the overall quality degree associated with the information in the system (Pipino et al., 2002).

One of the most known quality dimensions classification is presented by Wang and Strong (1996). They empirically identified fifteen IQ criteria under the perspective of a set of users. An empirical approach analyzed the information collected from the users and determined the characteristics of useful data for their tasks. The aspects were grouped into four broad information quality classes: intrinsic, contextual, representational, and accessibility. Intrinsic data quality denotes the quality of data itself. Contextual data quality enforces that data quality must be considered within the context of a task at hand, i.e., data must be relevant, timely, complete and appropriate in terms of amount. The Representational data quality category is related to the format and the meaning of data. Accessibility defines if data are available or obtainable for the user.

Regarding linked datasets, Zaveri et al. (2012) compiled a list of data quality criteria applicable to Linked Data quality assessment. To this end, they gathered and compared some existing approaches and grouped them under a common classification scheme. As a result, they identified a core set of 26 different data quality, which have been grouped into six dimension classes: contextual, intrinsic, accessibility, representation, trust and dataset dynamicity. They argue that these groups are not strictly disjoint but can partially overlap. Also, the dimensions are not independent from each other but correlations exist among dimensions in the same group or between groups. The contextual, intrinsic, representational and accessibility dimensions are similarly defined as in the work of Wang and Strong

(1996). Trust dimensions are those that focus on the trustworthiness of the dataset, and the dataset dynamicity ones are related to data update over time. Particularly, this work introduced new dimensions such as interlinking, volatility and currency in order to cover some linked data problems.

While several dimensions for IQ assessment have been proposed in the linked data setting, the notion of dataset completeness is still open. The semantics of completeness is crucial for linked datasets, considering that, in general, these datasets are not capable of providing complete answers to a given set of queries.

In general, completeness is a context-dependent quality dimension that refers to "the extent to which data are of sufficient breadth, depth and scope for the task at hand" (Wang and Strong, 1996). It is usually defined in terms of two basic dimensions (Mendes et al., 2012): schema and data completeness. Regarding the former, a dataset is complete if it contains all of the attributes needed for a given task (e.g., a query). On the data (instance) level, a dataset is complete if it contains all of the necessary objects (individuals) for a given task (e.g., a query). Naturally, the completeness of a dataset can only be judged in the presence of a task such as a given query or a set of queries where the ideal set of attributes and objects are known (Mendes et al., 2012).

3 ASSESSING THE COMPLETENESS OF A LINKED DATASET

The completeness of a linked dataset can only be assessed in the presence of a given task. In this work, the task we are dealing with is concerned with what the user needs in terms of data with respect to a given linked dataset. Then, it becomes necessary to specify the user data requirements to guide the evaluation of the completeness of a linked dataset.

Let D be a data domain (ex: "Biological", "Education"), and $DS=\{ds_1,...,ds_m\}$ a set of linked datasets, which belong to D. Each linked dataset $ds_j \in DS$ has a source description represented by a set of concepts and properties, which is formalized by means of an ontology or a vocabulary. Considering this, in our approach, user data requirements are defined as follows.

Definition 1 - User Data Requirements. We state that the user data requirements are represented by a set of queries $Q=\{q_1,...,q_n\}$,which provides what the user needs in terms of data belonging to D.

Since we are dealing with RDF linked datasets, we consider the SPARQL query language syntax and semantics to define queries. Thus, each query $q_i \in Q$ includes a number of statements that are defined as SPARQL Basic Graph Patterns (BGP). The BGP (?j rdf:type akt:Journal).(?j akt:published-by akt:ACM), for instance, looks for journals that have been published by ACM. Thus, the basic building blocks of a SPARQL query are triple patterns, which resemble RDF triples, except that in each position variables or resources are allowed. In the object position, literals can also be used. A resource may represent a class, a property or an instance of a class. We consider that classes are the resources that appear as objects in rdf:type patterns. In order to detect instances of classes, each resource present in each query $q_i \in Q$ will be checked using an ASK SPARQL as follows: ASK WHERE {**rcandidate** rdf:type rdfs:Class. optional {**rcandidate** rdf:type rdf:Property}}, where *rcandidate* is a candidate resource that will be evaluated by the ASK query. If the query result is false, the candidate resource is classified as an instance of a class.

In this sense, the assessment of the completeness criterion (and its subtypes) requires a detailed analysis of the triple patterns that compose the BGP (Basic Graph Pattern) of the SPARQL queries, which represent the user data requirements. To this end, for all $q_i \in Q$, we extract three specific sets, namely: (i) the set of triple patterns without literals, (ii) the set of triple patterns with literals, and (iii) the set of resources, which represent instances. These sets are defined as follows.

Definition 2 - Triple Patterns Set without literals (TP). Let $TP(Q) = \{tp_1,...,tp_n\}$ be a set of distinct triple patterns, which have been extracted from the BGP of each $q_i \in Q$. We state that each element tp_j represents a triple pattern that may be composed by SPARQL variables and resources.

Definition 3 - Triple Patterns Set with literals (TPL). Let $TPL(Q) = \{tpl_1,..., tpl_m\}$ be a set of distinct triple patterns, which have been extracted from the BGP of each $q_i \in Q$. We state that each element tpl_j represents a triple pattern where there is a literal in the object position.

Definition 4 – Set of Resources representing Instances (IR). Let $IR(Q) = \{ir_1,...,ir_k\}$ be a set of distinct resources, which have been extracted from the BGP of each $q_i \in Q$. We state that each element ir_j represents a class instance, and never a class or a property.

As an illustration, suppose a user is interested in data from the Bibliographic data domain (*D*), and that the AKT ontology[1] is used as the domain vocabulary to identify classes and properties. In order to represent the user requirements, the following SPARQL queries $Q = \{q_1, q_2\}$ were defined:

```
q1 = SELECT DISTINCT ?title
     WHERE {?article akt:has-title ?title .
            ?article akt:has-date akt-date:2000}
q2 = SELECT DISTINCT ?title
     WHERE {?article a akt:Article-Reference .
            ?article akt:has-title ?title .
            ?article akt:has-author ?author .
          ?author akt:full-name "Deborah Estrin"}
```

According to queries q_1 and q_2, the following three sets of triple patterns (triple patterns without literals, with literals and resources that represent instances) are identified: *TP = {[?article akt:has-title ?title], [?article a akt:Article-Reference], [?article akt:has-title ?title], [?article akt: has-author ?author], [?article akt:has-date ?var]}*; *TPL = {[?author akt:full-name 'Deborah Estrin']}* and *IR = {akt-date: 2000}*.

The triple pattern sets previously defined are fundamental for the completeness assessment. Based on these sets, a set of ASK SPARQL queries is defined, which is used to verify if a dataset meets the data requirements defined through *Q*. For each triple pattern $t \in \{TP \cup TPL\}$ is created an ASK query A_t, and for each $ir \in$ IR is created an ASK query A_{ir}.

As mentioned earlier, the degree of completeness is commonly assessed by the schema completeness and the data completeness criteria. In our approach, we extend this idea and we propose to evaluate the completeness of a given dataset with respect to *Q* by considering three criteria - schema completeness, literal completeness and instance completeness, as described in the following.

Definition 5 - Schema Completeness (SC). We state that schema completeness regards the degree to which the classes and properties of a linked dataset ds_j (described by its ontology) are present with respect to *TP(Q)*. More precisely, the schema completeness is concerned with how much each linked dataset $ds_i \in DS$ can answer each $tp_n \in TP(Q)$.

Formula (1) measures the schema completeness of a dataset ds_i with respect to a set of queries *Q*, when ds_i and *Q* belong to the same domain.

$$SC_{ds_i} = \frac{TP_{dsi}}{TP(Q)} \quad (1)$$

Where TP_{dsi} is the number of ASK SPARQL queries that a linked dataset ds_i answered as true for triple patterns without literal; *TP(Q)* is the number of triple patterns without literal extracted from *Q*.

Definition 6 - Literal Completeness (LC). The literal completeness regards the degree to which literals are present in $ds_i \in DS$. More precisely, this is evaluated by considering the $tpl_k \in TPL(Q)$ with respect to $ds_i \in DS$.

In order to measure how much a linked dataset contributes to answer the set of triple patterns with literals, we provide formula (2).

$$LC_{ds_i} = \frac{TPL_{dsi}}{TPL(Q)} \quad (2)$$

Where TPL_{dsi} is the number of ASK SPARQL queries that a linked dataset ds_i answered as true for triple patterns with literal; *TPL(Q)* is the number of triple patterns with literal extracted from *Q*.

This type of completeness states that more triple patterns with literal (*TPL*) are present in a linked dataset ds_i, more complete is ds_i in terms of literals.

Definition 7 - Instance Completeness (IC). Instance completeness is stated as the degree of resources that represent class instances, which are present on a linked dataset ds_i. It is concerned with evaluating each $ir_k \in IR$ with respect to $ds_i \in DS$.

In order to measure the instance completeness, we use formula (3).

$$IC_{ds_i} = \frac{IR_{dsi}}{IR(Q)} \quad (3)$$

Where IR_{dsi} is the number of ASK SPARQL queries that a linked dataset ds_i answered as true for the resources that represent instances of classes; *IR* is the number of resources that represent instances of classes extracted from *Q*.

To clarify matters, considering queries q_1 and q_2 (from the example provided earlier), it is possible to identify triple patterns without literals, with literals and resources that represent class instances and generate sets *TP(Q)*, *TPL(Q)* and *IR(Q)*. Also, by considering these sets, we are able to obtain the ASK queries sets. Table 1 shows triple patterns that were extracted from q_1 and q_2. It also depicts the derived ASK queries sets and indicates to which kind of completeness assessment they can be used.

The presented ASK query sets are used to verify if a given linked dataset ds_i is able to answer *Q*. To this end, for each triple pattern with an instance, a

[1] http://lov.okfn.org/dataset/lov/details/vocabulary_akt.html

Table 1: ASK SPARQL queries corresponding to q_1 and q_2.

Criterion	Triple Pattern	ASK SPARQL Query Sets
Schema Completeness	*?article a akt:Article-Reference*	*ASK WHERE {?article a akt:Article-Reference}*
	?article akt:has-author ?author	*ASK WHERE {?article akt:has-author ?author}*
	?article akt:has-title ?title	*ASK WHERE {?article akt:has-title ?title}*
	*?article akt:has-date **?var***	*ASK WHERE {?article akt:has-date **?var**}*
Data Completeness	*?author akt:full-name "Deborah Estrin"*	*ASK WHERE {?author akt:full-name "Deborah Estrin"}*
Instance Completeness	*?article akt:has-date akt-date:2000*	*ASK WHERE { akt-date:2000 ?p ?o OPTIONAL {?s ?p akt-date:2000} }*

new triple pattern is created with a SPARQL variable *(?var)* in the position of the instance. This new triple pattern is used to query ds_i, with respect to the assessment of the instance Completeness. Regarding schema completeness, as shown in Table 1, four triple patterns are used to query ds_i. Regarding literal completeness, there is only one triple pattern with literal, and, thereby only one ask query to be checked in ds_i. Also, we have found only one resource that represents an instance of a class on queries q_1 and q_2, i.e., the akt-date:2000. To assess the instance completeness we use only the resources that represent instances of classes rather than the complete triple pattern that contains it. Thus, the ASK query that evaluates the instance completeness checks if the instance exists in the linked dataset both in the position of the subject and of the object. Thereby, the modifier OPTIONAL is added in that ASK query.

4 EXPERIMENTS AND RESULTS

Some experiments were conducted to verify the effectiveness of our approach not only in terms of the proposed completeness metrics but also with respect to its cost-efficiency. The goal was twofold: (i) to check results obtained with each kind of completeness criterion and also with their combination into a single measure, and (ii) to verify if the approach, implemented in a semi automatic way, is cost-efficient. Our baseline regarded the situation where the user submits queries (SELECT ones) directly over each dataset endpoint. In this particular evaluation, the *Bibliographic data domain* (*D*) was considered. Data requirements, in terms of a set of queries *Q*, were identified and are presented in Table 2.

At first, a set of 32 public endpoints (*DS*) from Bibliographic data domain was selected for our experiments. In order to produce a ranking with these datasets, the set of queries *Q* was evaluated in each one of them. Then, based on the corresponding results, a ranking, called R_Q, was created to show datasets that are likely to be suitable to meet the user requirements, i.e., capable of answering queries from *Q*. For the sake of space, we present the first ten ranked datasets in Table 3[2]. This ranking, produced based on the SPARQL SELECT queries (Q) over the datasets, will be used as a baseline for the evaluation of our approach results.

Table 2: Data Requirements (Q).

Q	SPARQL Query	Query Description
q_1	SELECT DISTINCT ?title WHERE { ?journal a akt:Journal . ?journal akt:has-publication-reference ?publication . ?publication akt:has-title ?title }	Retrieves all names of existing journals.
q_2	SELECT DISTINCT ?name WHERE { ?article akt:has-author ?author . ?author akt:full-name ?name}	Selects all article author names.
q_3	SELECT DISTINCT ?title WHERE { ?article a akt:Article-Reference. ?article akt:has-title ?title . ?article akt:has-author ?author . ?author akt:full-name "Takeo Kanade" }	Selects all titles of articles, which belong to the author *Takeo Kanade*.
q_4	SELECT DISTINCT ?title ?web WHERE { ?article a akt:Article-Reference . ?article akt:has-title ?title . ?article akt:has-web-address ?web }	Retrieves all article titles and its corresponding web addresses.
q_5	SELECT DISTINCT ?title WHERE { ?article akt:has-title ?title . ?article akt:has-date akt-date:2000 }	Selects article titles which have been published in 2000.

Table 3: Baseline Ranking R_Q.

Number	Public Endpoints	q_1	q_2	q_3	q_4	q_5
1	http://lod.openlinksw.com/sparql	■	■	■	■	■
2	http://dblp.rkbexplorer.com/sparql/		■	■	■	■
	http://newcastle.rkbexplorer.com/sparql/	■	■		■	■
3	http://budapest.rkbexplorer.com/sparql/	■	■			■
	http://deploy.rkbexplorer.com/sparql/	■	■		■	
	http://ibm.rkbexplorer.com/sparql/	■	■			■
	http://irit.rkbexplorer.com/sparql/	■	■			■
	http://kaunas.rkbexplorer.com/sparql/	■	■			■
	http://laas.rkbexplorer.com/sparql/	■	■			■
	http://pisa.rkbexplorer.com/sparql/	■	■			■

In order to evaluate schema completeness, literal completeness and instance completeness, the sets *TP(Q)*, *TPL(Q)* and *IR(Q)*, presented in Table 4, were produced from the extraction of triple patterns from *Q*. For each triple pattern, a corresponding

[2] The complete evaluation is available at http://www.cin.ufpe.br/~dayse/qualitystamp.html

ASK SPARQL query was generated. Then, completeness assessment was performed based on formulas (1), (2) and (3). For each criterion, a new dataset ranking was created according to the completeness assessment results. The ranking results regarding schema, literal and instance completeness are presented in Tables 5, 6, and 7 respectively. We show the first ten ranked datasets for each measurement.

Table 4: Triple patterns and ASK queries regarding Q.

Criteria	Triple Pattern	ASK SPARQL Query
Schema Completeness	?journal a akt:Journal	ASK WHERE { ?journal a akt:Journal }
	?journal akt:has-publication-reference ?publication	ASK WHERE { ?journal akt:has-publication-reference ?publication }
	?publication akt:has-title ?title	ASK WHERE { ?publicacao akt:has-title ?title }
	?article akt:has-author ?author	ASK WHERE { ?article akt:has-author ?author }
	?author akt:full-name ?name	ASK WHERE { ?author akt:full-name ?name }
	?artigo a akt:Article-Reference	ASK WHERE { ?artigo a akt:Article-Reference }
	?article akt:has-title ?title	ASK WHERE { ?article akt:has-title ?title }
	?article akt:has-web-address ?web	ASK WHERE { ?article akt:has-web-address ?web }
	?article akt:has-date ?var	ASK WHERE { ?article akt:has-date ?var }
Literal Completeness	?author akt:full-name "Takeo Kanade"	ASK WHERE { ?author akt:full-name "Takeo Kanade" }
Instance Completeness	?article akt:has-date akt-date:2000	ASK WHERE {akt-date:2000 ?p ?o . OPTIONAL {?s ?p akt-date:2000} }

Table 5: Endpoints' ranking according to Schema Completeness.

Public Endpoints	SC	Ranking
http://lod.openlinksw.com/sparql/	1.0	1
http://newcastle.rkbexplorer.com/sparql/	1.0	
http://roma.rkbexplorer.com/sparql/	1.0	
http://budapest.rkbexplorer.com/sparql/	1.0	
http://irit.rkbexplorer.com/sparql/	1.0	
http://laas.rkbexplorer.com/sparql/	1.0	
http://deploy.rkbexplorer.com/sparql/	1.0	
http://ulm.rkbexplorer.com/sparql/	1.0	
http://dblp.rkbexplorer.com/sparql/	0.85714	2
http://rae2001.rkbexplorer.com/sparql/	0.85714	

Table 6: Endpoints' ranking according to Literal Completeness.

Endpoints	LC	Ranking
http://lod.openlinksw.com/sparql/	1.0	1
http://dblp.rkbexplorer.com/sparql/	1.0	
http://acm.rkbexplorer.com/sparql/	1.0	
http://citeseer.rkbexplorer.com/sparql/	1.0	
http://nsf.rkbexplorer.com/sparql/	1.0	

Table 7: Endpoints' ranking according to Instance Completeness.

Endpoints	IC	Ranking
http://lod.openlinksw.com/sparql/	1.0	1
http://dblp.rkbexplorer.com/sparql/	1.0	
http://rae2001.rkbexplorer.com/sparql/	1.0	
http://newcastle.rkbexplorer.com/sparql/	1.0	
http://roma.rkbexplorer.com/sparql/	1.0	
http://budapest.rkbexplorer.com/sparql/	1.0	
http://irit.rkbexplorer.com/sparql/	1.0	
http://laas.rkbexplorer.com/sparql/	1.0	
http://kaunas.rkbexplorer.com/sparql/	1.0	
http://ibm.rkbexplorer.com/sparql/	1.0	

Comparing results of R_Q with results from Table 5, Table 6 and Table 7, we can observe that the ranking obtained with our approach is very similar to the one obtained in the baseline (Rq). This means that our approach is able to point out, by considering any of the completeness measures, the datasets that meet the user data requirements. However, in our approach there is no need to perform complex and time consuming queries.

Then, in order to obtain a single completeness measure (SCM), individual completeness measures were combined. To this end, a weighted sum of scores was considered (Naumann, 1998). These results are depicted in Table 8. Comparing results of R_Q (Table 3) with results from Table 8, we observe that they are very similar. More precisely, these results show that the three types of completeness assessment may be accomplished individually, although we perceive that they are indeed complementary and, when combined, produce similar results. Nevertheless, in our approach, the user may verify each one individually depending on their application needs.

Regarding the second goal – if our approach is cost-efficient, Figure 1 illustrates the overall time for executing each one of the queries from Q over the 32 select datasets. As showed in Figure 1, our approach requires less time (290247ms) to evaluate the datasets completeness when compared to the manual approach (798866ms).

Finally, we verify that our approach for linked dataset completeness assessment is a promising and affordable way to help the selection of datasets for particular applications. This is due to the fact that the user may verify if a given dataset can really meet his data requirements and in which degree the completeness can be achieved. Moreover, a more automatic approach avoids the user to accomplish the hard and time consuming task of verifying each dataset individually in a manual way and producing

measures by himself.

Table 8: Endpoints' ranking according to SCM.

Endpoint	SCM	Ranking
http://lod.openlinksw.com/sparql/	1.0	1
http://dblp.rkbexplorer.com/sparql/	0.95913	2
http://newcastle.rkbexplorer.com/sparql/	0.64303	3
http://roma.rkbexplorer.com/sparql/	0.64303	
http://budapest.rkbexplorer.com/sparql/	0.64303	
http://irit.rkbexplorer.com/sparql/	0.64303	
http://laas.rkbexplorer.com/sparql/	0.64303	
http://ulm.rkbexplorer.com/sparql/	0.64303	
http://rae2001.rkbexplorer.com/sparql/	0.60216	4
http://kaunas.rkbexplorer.com/sparql/	0.60216	

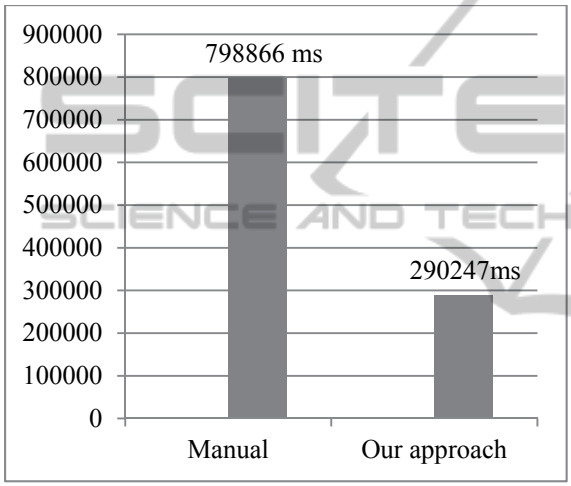

Figure 1: Manual and Approach Assessment Time.

5 RELATED WORK

Although there is some research concerning completeness assessment in distributed environments (e.g., the work of Roth and Naumann (2007)), few works discuss the issue of completeness of RDF datasets (Mendes et al., 2012; Harth and Speiser, 2012; Dadari et al., 2013).

Dadari *et al.* (2013) introduce a theoretical framework for describing data sources in terms of their completeness. They focus on the problem of the completeness of query answering over plain and RDFS data sources augmented with completeness statements expressed in RDF. In (Mendes et al., 2012) a solution for flexibly expressing quality assessment methods as well as fusion methods is presented. The quality assessment module allows users to define relevant indicators and respective scoring functions for their specific quality assessment task. The data fusion module uses quality scores to decide on which values to keep, discard or transform when a data integration task is performed. In the work of Harth and Speiser (2012), several notions of completeness for queries over Linked Data are presented. They also introduce the notion of authoritativeness to specify which information resources are necessary to have complete information about an identifier. Then, different types of completeness are defined based on that idea of authoritativeness of sources.

Differently from these related works, our work applies completeness assessment to help the identification of linked dataset capable to meet specific data requirements of linked data applications. To this end, it uses information extracted from a set of queries (which defines the user requirements) to evaluate both schema and data completeness. Particularly, it doesn't require execution of time consuming queries in a conventional (or manual) way. In addition, it measures two distinct types of data completeness (literal and instance ones) to help achieving the overall completeness of a linked dataset.

It is also important to note that our proposal for ranking linked data sets is also similar to other approaches proposed to select data sources in the context of federated queries (Schwarte et al. 2011). (Schwarte et al. 2011) uses ASK queries to select data sources that are able to answer specific triple patterns. In our approach, we also use ASK queries to select data sources that are able to answer specific triple patterns. However, ASK queries are also used to select data sources that have a specific resource (i.e. an instance of a given class). It is important to note that the goal of our approach is to generate a ranking of sources (linked datasets). It is not concerned with neither query optimization nor query decomposition. Our approach evaluates completeness by using IQ metrics, thus providing a ranking which indicates the best sources.

6 CONCLUSIONS AND FUTURE WORK

With the ever growing availability of linked datasets on the Web and the development of applications for consuming these data, the selection of which datasets better meet application requirements has become a key issue. Usually, this is done in a manual way by searching each dataset individually, what represents a hard and time consuming task for the users. In this scenario, the approach for assessing

the completeness of linked datasets, presented in this paper, is an option to find suitable linked datasets to specific applications with respect to their completeness. This is accomplished by taking into account the application data requirements in terms of a set of queries.

Experiments have shown that the approach is able to produce particular measures for each one of the defined completeness metrics, namely: schema, literal and instance. In addition, it is able to combine the three metrics into one single measure. In both situations, the approach produces similar results as the ones obtained in a manual search. The main difference is that in our approach, this task is done in a more automatic way, thus enabling the user to select the datasets in less time and considering IQ measurements.

Future work includes considering other application and domain scenarios and accomplishing performance and scalability experiments.

REFERENCES

Darari, F., Nutt, W., Pirró, G. and Razniewski, S. 2013. *Completeness statements about RDF data sources and their use for query answering*. In Proceedings of the 12th International Semantic Web Conference, ISWC 2013, Sydney, NSW, Australia, October 21-25.

Fürber, C., and Hepp, M. 2011. *Swiqa - a semantic web information quality assessment framework*. In Proceedings of the 19th European Conference on Information Systems, ECIS 2011, Helsinki, Finland, June 9-11, 2011.

Harth, A. and Speiser, S. 2012. *On Completeness Classes for Query Evaluation on Linked Data*. In Proceedings of the 26th AAAI Conference, AAAI 2012, Toronto, Canada, July 22–26.

Heath, T. and Bizer, C. 2011. *Linked Data*. (1st ed.). Morgan & Claypool Publishers.

Mendes, P., Mühleisen, H., and Bizer, C. 2012. *Sieve: linked data quality assessment and fusion*. In Proceedings of the 2012 Joint EDBT/ICDT Workshops (EDBT-ICDT '12), Divesh Srivastava and Ismail Ari (Eds.). ACM, New York, NY, USA, 116-123. DOI=http://doi.acm.org/10.1145/2320765.2320803.

Naumann F. 1998. *Data Fusion and Data Quality*, In: New Techniques and Technologies for Statistics Seminar (NTTS'98). Sorrent, Italy, 1998.

Pipino, L. L., Lee, Y. and Wang, R. 2002. *Data quality assessment*. Commun. ACM 45, 4 (April 2002), 211-218. DOI=http://doi.acm.org/10.1145/505248.506010.

Price, R. and Shanksa, G. 2005. *A semiotic information quality framework: development and comparative analysis*. JIT 20, 2 (June 2005), 88-102. DOI=10.1057/palgrave.jit.2000038.

Roth, A. and Naumann, F. 2007. *System P: Completeness-driven Query Answering in Peer Data Management Systems*. In Proceedings of the Datenbanksysteme in Business, Technologie und Web, BTW 2007, Aachen, Germany, March 7-9.

Schwarte, A., Haase, P., Hose, K, Schenkel, R. and Schmidt, M. 2011. *FedX: Optimization Techniques for Federated Query Processing on Linked Data*. In Proceeding of the 10th international conference on The semantic web (ISWC'11), Bonn (Germany).

Wang, R. Y. and Strong, D. M. 1996. *Beyond accuracy: what data quality means to data consumers*. J. Manage. Inf. Syst. 12, 4 (March 1996), 5-33.

Zaveri, A., Rula, A., Maurino, A., Pietrobon, R., Lehmann, J., Auer, S. 2012. Quality assessment methodologies for linked open data. Available at: http://www.semantic-web-journal.net/content/ quality-assessment-methodologies-linked-open-data.