An approach for measuring rdf data completeness

Fayçal Hamdi, Samira Si-Said Cherfi

▶ **To cite this version:**

Fayçal Hamdi, Samira Si-Said Cherfi. An approach for measuring rdf data completeness. BDA - Gestion de Données – Principes, Technologies et Applications, Sep 2015, Ile de Porquerolles, France. pp.32-41. hal-02476016

HAL Id: hal-02476016
https://hal.science/hal-02476016

Submitted on 10 Apr 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

ISSN 2429-4586

An approach for measuring RDF data completeness

Fayçal Hamdi
Cedric Lab, Conservatoire National des Arts et Métiers (CNAM)
Paris, France
faycal.hamdi@cnam.fr

Samira Si-Said Cherfi
Cedric Lab, Conservatoire National des Arts et Métiers (CNAM)
Paris, France
samira.cherfi@cnam.fr

ABSTRACT

With the increasing use of the Web of Data and the development of data based analysis applications, there is a real need for developing suitable methods and techniques for evaluating and help ensuring web data quality. Among Quality dimensions, completeness is recognized as difficult to evaluate, as it often relies on gold standards and/or a reference data schema that are neither always available nor realistic from a practical point of view. In this paper, we propose an approach for Linked Data completeness assessment. The approach is a two-step process: first, mining from a data source, a schema that reflects the actual representation of data, which, in the second step, is used for completeness evaluation. The paper presents both theoretical background and experimental results performed on real-world RDF linked datasets.

Keywords

Linked Data, RDF Data Quality, Completeness, Quality Evaluation

1. INTRODUCTION

Nowadays, with the development of web technologies, data is considered as a strategic asset for increasing companies' revenue. This phenomenon is enhanced thanks to the crowd sourced community efforts that led to the availability of huge structured data sets covering a wide variety from several sources such as civil society, local communities and domain experts. This data is being released continuously and, thanks to its digital format, it is mainly suitable for computerized usage leading to the development of new techniques, technologies, practices and methodologies aiming to better understand business and market needs. As a consequence, this data, accessible from a wide range of users, empowers and even influences decision making processes. According to a report by the McKinsey Global Institute [19], by using social technologies, companies could increase margins by as much as 60 percent. In research area, the exploration of semantic links between datasets will enable novel analyzes to be performed. For example, linked open drug data aims to connect previously unlinked results from clinical trials, gene expression assays, and chemical testing [27].

However, as data publishing does not require special expert knowledge or skills, availability of data does not always guarantees its usability. This means that the quality of published data is not as good as we could expect leading to a low added value and low reliability of the derived conclusions. Information quality has been extensively studied in the context of relational databases [29, 21]. However, in the context of Web of Data, this is an emerging issue and a real challenge for future research. Indeed, Linked Open Data[1] (LOD) through HTTP URI's offer an infrastructure for publishing and navigating through data without ensuring the quality of its usage [3]. As a consequence, there is a real need for developing suitable methods and techniques for evaluating and help ensuring web data quality.

Quality of the Web of Data attracted recent interest and main contributions go into two directions. The first one relies on user contribution to qualify the quality of data and is thus subjective. The second one, more objective, mainly concentrates on data provenance. In this paper we are more precisely concerned with data completeness as a dimension of data quality and the way to assess it. Indeed, data completeness is recognized as an important quality dimension and providing completeness information about a data source helps increasing the confidence of such a source.

Data completeness has two facets. The first one analyzes whether all data is available. Such completeness is known as structural completeness [1] and requires a reference benchmark or gold data set as completeness reference. This vision is more related to the amount of data provided and its relevance to answer users requests. The second facet entails having a value for each property of the data. This latter is then said complete if all necessary values are recorded. A traditional way to measure completeness in this case is the rate of missing values. This subsumes that data rely on an agreed and well-designed schema in which properties are equally relevant. Indeed, assessing completeness as the rate of missing values does not take into account the fact that missing a marginal property should have a less importance than missing an essential value. In the context of web data

[1] http://lod-cloud.net

(c) 2015, Copyright is with the authors. Published in the Proceedings of the BDA 2015 Conference (September 29-October 2, 2015, Ile de Porquerolles, France). Distribution of this paper is permitted under the terms of the Creative Commons license CC-by-nc-nd 4.0.
(c) 2015, Droits restant aux auteurs. Publié dans les actes de la conférence BDA 2015 (29 Septembre-02 Octobre 2015, Ile de Porquerolles, France). Redistribution de cet article autorisée selon les termes de la licence Creative Commons CC-by-nc-nd 4.0.
BDA 2015, 29 septembre au 02 octobre 2015, Ile de Porquerolles, France.

these two assumptions are not satisfied.

Let's consider, for instance, a collaboratively built dataset. In this case, the traditional top down vision of a predefined schema is no true. Both data and underlying schema evolve continuously as data are described by several communities having different views and needs. The challenge is thus to provide a suitable completeness assessment method taking into account the absence of an agreed schema. Accordingly, we adopt in the remaining of the paper the definition from [28] for completeness as the ability to represent every meaningful state of the represented real world system. This definition highlights two interesting aspects. The first one is the notion of state considering the combination of properties rather than their individual values. The second is the notion of meaningful state' suggesting that completeness should take into account the relevance of properties when analyzing missing values.

We aim to provide an approach for RDF data completeness assessment. Our approach could be applied based on a reference schema that is inferred from data values. This schema will represent the meaningful state of the considered dataset. In summary, we make the following contributions:

1. We use a mining approach to infer a schema from data, as we consider that no predefined schema exists.

2. We introduce a novel approach for data completeness assessment based on both the inferred schema and the data.

3. We experimentally assess our approach by using real-world datasets as a case study. We analyzed the relationship between the completeness and the number of properties in the inferred schema, and the completeness and the robustness of the calculation when dataset size varies.

The remainder of this paper is organized as follows: section 2 presents a motivating example introducing the problem tackled in this article; section 3 gives a formal description of the problem; section 4 details the mining-based approach; section 5 presents and analyzes a set of experiments; section 6 summarizes a related literature on the subject while section 7 draws conclusions and future research directions.

2. MOTIVATING EXAMPLE

We illustrate in this section the main idea behind our approach through an example that shows the issues and the difficulties encountered in the calculation of a dataset completeness. Let consider the set of scientists described in the well-known open linked dataset, DBpedia. We would like to know, when querying this dataset about a particular scientist, if the information provided for this scientist are complete or not. We would like also to know if all the scientists in this dataset are well described.

To do so, the first intuition consists of comparing the properties used in the description of each scientist with a reference scientist schema (ontology). For example, in DBpedia, the class *Scientist*[2] has a list of 4 properties (e.g. *doctoralAdvisor*), but these properties are not the only ones used in the description of a scientist (e.g. the *birthdate* property is not present in this list). Indeed, the class *Scientist* has a super class called *Person*. So, the description of a scientist may also take into account the properties of this class. Therefore, to obtain an exhaustive list of the whole properties used in the description of a scientist, we have to calculate the union of the set of properties of the class *Scientist* and all its ancestors. For our example, the reference scientist schema that we called *Scientist_Schema* could be calculated as follows:

$$Scientist\_Schema = \{Properties\ on\ Scientist\} \cup \{Properties\ on\ Person\} \cup \{Properties\ on\ Agent\} \cup \{Properties\ on\ Thing\}$$

$$\text{such that: } Scientist \sqsubseteq Person \sqsubseteq Agent \sqsubseteq Thing$$

Thus, the completeness of a scientist description (e.g. *Albert_Einstein*) will be the proportion of properties used in the description of this scientist to the total number of properties in *Scientist_Schema*. In the case of DBpedia, with a simple SPARQL query[3], we can obtain the size of *Scientist_Schema*, which is equal to 664 (A-Box properties). So, the completeness of the description of *Albert_Einstein* could be calculated as follows:

$$Comp(Albert\_Einstein) = \frac{|Properties\ on\ Albert\_Einstein|}{|Scientist\_Schema|} = \frac{21}{664} = 4,21\%$$

However, on the one hand, we know that the description of a scientist does not possibly need all the properties of the schema. For example, the property *weapon* is far from being relevant in the description of *Albert_Einstein* (in DBpedia, this property is not used in any scientist instances, whether it is part of the Scientist_Schema) and cannot have the same importance as, for instance, the *name* or the *university* of a scientist. This is also the case of several properties in *Scientist_Schema*. In addition, a description, with 21 A-Box properties[4] (and 66 T-Box and other external properties), of a scientist seems to be actually a good representation and provides enough useful information about this scientist. Thus, considering that the description of *Albert_Einstein* has solely 4.21% as a value of completeness is quite reductive. On the other hand, when we calculate, for example, the completeness of the first 1000 scientists, we obtained a value of completeness equals to 1.37% (lower that the one obtained for the so "famous" scientist *Albert_Einstein*). This is due to the fact that several "not famous" scientist has a very few number of A-Box properties. Therefore, the completeness of the whole dataset gets lower regarding to the number of the "not famous" scientists (it certainly get even lower when it calculated for the 18,233 DBpedia scientists[5]).

[2] `http://mappings.dbpedia.org/server/ontology/classes/`
[3] Performed on: `http://dbpedia.org/sparql`
[4] `http://dbpedia.org/resource/Albert_Einstein`
[5] `http://wiki.dbpedia.org/Datasets/DatasetStatistics` (statistics of the DBpedia 2014 version for the english language)

We can finally conclude that, the completeness as calculated here, does not provide us with the relevant value regarding the real representation of scientists in the DBpedia dataset. Hence, to overcome this issue, there is a need to explore instances to get an idea about how they are actually describing and which properties, with the importance of each one, are used. Based on data mining, the approach that we propose in this paper, deals with this issue by extracting, from a set of instances (of the same class), the pattern of the most representative properties and calculates a completeness in respect to this pattern.

3. PROBLEM STATEMENT

The problem of completeness evaluation is not to find an absolute value but to compute the more suitable one for the context of use. In our situation, the context is a dataset representing a category or a set of categories such as *Actor* or *Organization*. Table 1 illustrates some instances of the *Actor* category in form of triples, taken from DBpedia.

Table 1: A sample of triples from DBpedia

Subject	Predicate	Object
Ben_Affleck	birthDate	1972-01-01
Ben_Affleck	residence	Los Angeles
Angelina_Jolie	birthDate	1975-06-04
Angelina_Jolie	citizenship	United_States
Adam_West	birthDate	1928-09-19
Adam_West	citizenship	American
Adam_West	residence	Ketchum,_Idaho

Each Category is described by a set of properties (predicates) and an instance of this category could have a value for all the properties or only for a subset of these properties. This subset is called transaction. Table 2 represents the set of transactions constructed from the triples of the table 1.

Table 2: Transactions created from triples

Instance	Transaction
Ben_Affleck	birthDate, residence
Angelina_Jolie	birthDate, citizenship
Adam_West	birthDate, citizenship, residence

More formally, lets define a dataset $\mathcal{D}$ to be triple (C, I_C, P), where C is the set of categories (e.g., *Actor*, *City*), I_C is the set of instances for categories in C (e.g., *Ben_Affleck* is an instance of the *Actor* category), and $P = \{p_1, p_2, ..., p_n\}$ is the set of properties (e.g $residence(Person, Place)$).

Let $\mathcal{T} = \{t_1, t_2, ..., t_m\}$ be a set of transactions with $\forall k, 1 \leq k \leq m : t_k \subseteq P$ be a vector of transactions over P, and $E(t_k)$ be the set of items in transaction t_k. Each transaction is a set of properties used in the description of the instances of the subset $\mathcal{I}' = \{i_1, i_2, ..., i_m\}$ with $\mathcal{I}' \subseteq I_C$ (e.g., properties used to describe the *Ben_Affleck* instance are: *birthDate* and *residence*). We consider $\mathcal{CP}$ the completeness of $\mathcal{I}'$ against properties used in the description of each of its instances.

Problem definition Given a dataset $\mathcal{D}$, a subset of instances $\mathcal{I}'$, and a set of transactions $\mathcal{T}$ constructed from $\mathcal{I}'$, we would like to calculate the completeness $\mathcal{CP}$ of $\mathcal{I}'$.

4. THE MINING-BASED APPROACH

Completeness at the data level assesses missing values [26]. This vision requires a schema that needs to be inferred from the data source. However, as we mentioned in the motivating example, considering a schema as the union of all properties used for a category description is not relevant. Indeed, this vision neglects the fact that missing values could express inapplicability [6]. A missing value is said inapplicable when it represents a property that is inapplicable to the particular object or instance. In this case, we could conclude that all properties are not equally important.

To take this aspect into consideration, we propose an approach that calculates the completeness of an input dataset by posing the problem as an itemset mining problem. Therefore, the process of inferring a reference schema will consider the real data, actually contained in the dataset. Our mining-based approach includes two steps:

1. **Properties mining**: Given a dataset $\mathcal{D}$, we first represent the properties, used for the description of the $\mathcal{D}$ instances, as a transactions vector. We then apply the well known FP-growth algorithm [13, 14] for mining frequent itemsets (we chose FP-growth for efficiency reasons. Any other itemset mining algorithm could, obviously, be used). Only a subset of these frequent itemsets, called "Maximal" [20, 11, 12], is captured. This choice is motivated by the fact that, on the one hand, we are interested in the *expression* of the frequent pattern and, on the other hand, the number of frequent patterns could be exponential when the transaction vector is very large (see Section 4.1 for details).
2. **Completeness calculation**: Once the set of maximal frequent itemsets $\mathcal{MFP}$ is generated, we use the apparition frequency of items (properties) in $\mathcal{MFP}$, to give each of them a weight that reflects how important the property is considered for the description of instances. Weights are then exploited to calculate the completeness of each transaction (regarding the presence or absence of properties) and, hence, the completeness of the whole dataset.

In the following we give a detailed description of each step.

4.1 Properties mining

Let $\mathcal{D}(C, I_C, P)$ be a dataset and $\mathcal{I}'$ be a subset of instances with $\mathcal{I}' \subseteq I_C$. We first initialize $\mathcal{T} = \phi$, $\mathcal{MFP} = \phi$. For each $i \in \mathcal{I}'$ we generate a transaction t. Indeed, each instance i is related to values (either resources or literals) through a set of properties. Therefore, a transaction t_k of an instance i_k is a set of properties such that $t_k \subseteq P$. Transactions generated for all the instances of $\mathcal{I}'$ are then added to the $\mathcal{T}$ set.

Example 1. Taking table 1, let $\mathcal{I}'$ be a subset of instances such that: $\mathcal{I}' = \{$*Ben_Affleck, Angelina_Jolie, Adam_West*$\}$. The set of transaction $\mathcal{T}$ would be:

$$\mathcal{T} = \{\{birthDate, residence\}, \{birthDate, citizenship\}, \{birthDate, citizenship, residence\}\}$$

The objective is then to compute the set of frequent patterns $\mathcal{FP}$ from the transaction vector $\mathcal{T}$.

Definition 1. (*Pattern*) Let $\mathcal{T}$ be a set of transactions. A pattern $\hat{P}$ is a sequence of properties shared by one or several transactions t in $\mathcal{T}$.

For any pattern $\hat{P}$, let $E(\hat{P})$ be the corresponding set of items (constitutes, in our case, of properties), and $T(\hat{P}) = \{t \in \mathcal{T} \mid E(\hat{P}) \subseteq E(t)\}$ be the corresponding set of transactions. $E(\hat{P})$ designates the *expression* of $\hat{P}$, and $|T(\hat{P})|$ the *support* of $\hat{P}$. A pattern $\hat{P}$ is frequent if $\frac{1}{|\mathcal{T}|}|T(\hat{P})| \geq \xi$, where ξ is a user-specified threshold.

Example 2. Taking table 2, let $\hat{P} = \{birthDate, residence\}$ and $\xi = 60\%$. $\hat{P}$ is frequent as its relative support (66.7%) is greater than ξ.

To find all the frequent patterns $\mathcal{FP}$, we used, as we motivated above, the FP-growth itemsets mining algorithm. However, according to the size of the transactions vector, the FP-growth algorithm could generate a very large $\mathcal{FP}$ set. As our objective is to see how a transaction (a description of an instance) is *complete* against a set of properties, we focus on the pattern *expression* (in terms of items it contains) instead of its *support*.

For completeness calculation, we need to select a pattern to serve as reference schema. This pattern should present a right balance between frequency and expressiveness. In itemset mining, a concept, called "Maximal" frequent patterns, allow us finding this subset. Thus, to reduce $\mathcal{FP}$, we generate a subset containing only "Maximal" patterns.

Definition 2. ($\mathcal{MFP}$) Let $\hat{P}$ be a frequent pattern. $\hat{P}$ is maximal if none of its proper superset is frequent. We define the set of Maximal Frequent Patterns $\mathcal{MFP}$ as:

$$\mathcal{MFP} = \{\hat{P} \in \mathcal{FP} \mid \forall \hat{P}' \supsetneq \hat{P} : \frac{|T(\hat{P}')|}{|\mathcal{T}|} < \xi\}$$

Example 3. Taking table 2, let $\xi = 60\%$ and the set of frequent patterns $\mathcal{FP} = \{\{birthDate\},\{residence\},\{citizenship\}, \{birthDate,residence\}\{birthDate,citizenship\}\}$. The $\mathcal{MFP}$ set would be:

$$\mathcal{MFP} = \{\{birthDate,residence\}, \{birthDate,citizenship\}\}$$

4.2 Completeness calculation

In this step our goal is to find a unique pattern that allows us evaluating the completeness of a transaction regarding this pattern. However, the maximal frequent patterns $\mathcal{MFP}$, that we generated in the previous steps, contains often several candidate patterns. In this case, our strategy consists of reducing the $\mathcal{MFP}$ set to obtain only one pattern. However, all the properties from $\mathcal{MFP}$ are not equally important. We consider for that two aspects; the apparition frequency of an item (a property) in $\mathcal{MFP}$, and the support of each $\hat{P} \in \mathcal{FP}$ containing this item. The result will be a *weighted* frequent pattern that we denote $\hat{P}_w$. This pattern will be used then as a reference schema in the completeness calculation.

Definition 3. (*Weighted frequent pattern*) Let $\mathcal{MFP}$ be a set of maximal frequent pattern. A weighted frequent pattern $\hat{P}_w$ is a set of couples $< p, w(p) >$ composed of a property p and a weight w associated to this property. The weight w is calculated as follows:

$$\forall p \in \bigcup_{i=1}^{n} E(\hat{P}_i) : w(p) = \frac{1}{|\mathcal{MFP}|} \sum_{i=1}^{n} \delta(p,i).\frac{|T(\hat{P}_i)|}{|\mathcal{T}|} \quad (1)$$

such that: p a *singleton*, $\hat{P}_i \in \mathcal{MFP}$ and

$$\delta(p,i) = \begin{cases} 1 & \text{if } p \in E(\hat{P}_i) \\ 0 & \text{otherwise} \end{cases}$$

The equation 1 takes into account, as we mentioned above, the frequency of a property p by checking, via the δ parameter, its presence on the itemset of each maximum frequent pattern. As the weight is calculated relatively to the number of transactions in $\mathcal{T}$ (to get a relative *support*), its value will range between 0 and 1.

Example 4. Let $\mathcal{MFP}$ = {{*birthDate,residence*}, {*birthDate,citizenship*}} where both itemsets have a *support* of 67%. The set of the weighted frequent patterns $\hat{P}_w$ would be:

$$\hat{P}_w = \{\{birthDate, 0.67\}, \{residence, 0.33\}, \{citizenship, 0.33\}\}$$

Once we obtained the weighted pattern $\hat{P}_w$, we carry out for each transaction, a comparison between its corresponding properties and the weighted properties of $\hat{P}_w$. We get, therefore, an indication about the completeness of each transaction $t \in \mathcal{T}$.

Definition 4. (*Completeness* $\mathcal{C}$) Let $\mathcal{I}'$ a subset of instances, $\mathcal{T}$ the set of transactions constructed from $\mathcal{I}'$, and $\hat{P}_w$ the set of weighted frequent properties. The completeness of $\mathcal{I}'$ corresponds to the completeness of its transaction vector $\mathcal{T}$ obtained by calculating the average of the completenesses of each transaction regarding $\hat{P}_w$. Therefore, we define the completeness $\mathcal{CP}$ of a subset of instance $\mathcal{I}'$ as follows:

$$\mathcal{CP}(\mathcal{I}') = \frac{1}{|\mathcal{T}|} \sum_{k=1}^{|\mathcal{T}|} \sum_{j=1}^{n} \frac{w(p_j \cap E(t_k))}{\mathcal{W}} \quad (2)$$

such that:

$$w(\emptyset) = 0,\ \mathcal{W} = \sum_{j=1}^{n} w(p_j),\ p_j \in \bigcup_{j=1}^{n} E(\hat{P}_j), \text{ and } \hat{P}_j \in \mathcal{MFP}$$

It is worth noting that, the weights of the properties allow us to evaluate the completeness of a given transaction with respect to a context, which is in our case, the properties used in the all transactions. The Algorithm 1 shows the pseudo-codes for calculating $\mathcal{CP}(\mathcal{I}')$.

Example 5. Let $\xi = 60\%$. The completeness of the subset of instances in table 1 regarding $\hat{P}_w$ = {{birthDate,0.67}, {residence,0.33},{citizenship,0.33}}, would be:

$\mathcal{CP}(\mathcal{I}')$ = (2∗(0.33+0.67)+(0.33+0.33+0.67))/1.33∗3 = 0.83

This value corresponds to the completeness average value for the whole dataset regarding the inferred schema $\hat{P}_w$.

Algorithm 1 Completeness calculation

Input: $\mathcal{D}, \mathcal{I}', \xi$
Output: $\mathcal{CP}(\mathcal{I}')$
  for each $i \in \mathcal{I}'$ **do**
    $t_i = \lfloor p_1 \quad p_2 \quad \dots \quad p_n \rfloor$
    $\mathcal{T} = \mathcal{T} + t_i$
  ▷ *Properties mining*
  $\mathcal{MFP} = Maximal(FP\text{-}growth(\mathcal{T}, \xi))$
  $\hat{P}_w =< nil, nil >$
  for each $\hat{P} \in \mathcal{MFP}$ **do**
    for each $p \in \hat{P}$ **do**
      ▷ *Using equation 1*
      $w(p_i) = CalculateWeight(p, \mathcal{MFP}, \mathcal{T})$
      $\hat{P}_w.put(p_i, w(p_i))$
  ▷ *Using equation 2*
  return $CP(\mathcal{I}') = CalculateCompleteness(\mathcal{I}', \mathcal{T}, \hat{P}_w)$

5. EMPIRICAL EVALUATION

This section is devoted to a set of experiments aiming to evaluate our approach against a variation of its underlying parameters. The evaluation methodology explores the behavior of completeness metric regarding two sides. The first concerns the impact of the number of instances whereas the second is related to the user-specified threshold ξ. The experiments were performed on two well-known real-world datasets, publicly available on the Linked Open Data (LOD). The first one, DBpedia, is a large knowledge base composed of structured information extracted collaboratively from Wikipedia. It describes currently 4.58 million things. The second dataset, Freebase, is a large collaborative knowledge base of the world's information, built mainly from data provided by its community members. The last version of Freebase (before its migration to Wikidata) includes approximately 47.3 million topics and 2.9 billion facts.

5.1 Dataset description

For the evaluation of the robustness of our approach we are looking to investigate its behavior regardless the nature of data. To do so, for each dataset we have chosen a couple of categories from different natures. For DBpedia, we studied the completeness of instances that have as types the following categories: $C = \{$*Populated Place*, *Organisation*, *Actor*, *Athlete*$\}$, and for Freebase the instances that have as types the following ones (which are equivalent or close to those of DBpedia): $C = \{$*Citytown*, *Organization*, *Football Player*, *Museum*$\}$. Indeed, our objective is not to compute an absolute completeness of a source. In practice, the completeness of the whole source is of marginal interest since queries often concern a subset of data. We thus analyzed the behavior and robustness of our completeness measure for various categories with different sizes. This is motivated by the fact that good completeness values, for a desired category, remain attractive even if the overall completeness (average or other aggregate measure) or the completeness of categories that do not interest the user are not satisfactory. Besides, a completeness of 99% may be uninteresting if the schema is ridiculously small.

In the first step of our experiments, we performed queries (SPARQL for DBpedia and MQL for Freebase) on datasets endpoints to extract data concerning each category. We then constructed the set of corresponding transactions $\mathcal{T}$. A transaction vector is constituted of sequences of properties deduced from instances belonging to a single category (e.g. the set of *Actors* in DBpedia). The set of transactions[6] is then used as an input to generate the frequent patterns and to compute the completeness. Experiments were run on a Dell XPS 27 with an Intel Core i7-4770S processor and 16GB of DDR3 RAM. The execution time of each experiment is insignificant (less than 5 seconds).

5.2 Impact of the number of instances

In this experiment we compare the completeness values, obtained for each category when varying the number of transactions (values range between 100 and 10000). Besides, we explain the completeness values that we obtained using a brute force approach (BFA), which measure the completeness according to the whole properties used in the descriptions of the instances. The results of DBpedia and Freebase categories are given in Figures 1 and 3.

We observe in these two figures that, concerning our metric, the obtained completeness values are relatively stable regarding the variation of the number of transactions (e.g. they range between 0.92 and 0.95 for the DBpedia category *Populated Place* with $\xi = 60\%$). For Freebase categories, the values are less regular than those of DBpedia. This means that the distribution of missing values in DBpedia is more regular than the one of Freebase. This observation is interesting as it expresses the fact that, for the used categories at least(e.g. *organization*), the instances have a widely divergent descriptions (despite the fact that they have the same type).

For the number of properties, represented in Figure 2 and 4, the conclusions are the same. The number of properties in $\hat{P}_w$ remains stable for the different itemsets sizes. This expresses a relative robustness of the metrics regarding the number of instances.

Concerning the completeness computed using the brute force approach, the values that we obtained are very low (less than 0.26 for both DBpedia and Freebase categories) and get lower (tend to zero) when the number of transactions grows (hence, we do not represent those values in the two figures). This is due to the fact that the instances from DBpedia or Freebase categories do not use all possible properties present in the selected set of instances. As we explained above in the motivating example, this approach could not provide us with a relevant value of completeness, as properties have not all the same importance.

5.3 Impact of the user-specified threshold ξ

To measure the impact of the minimum support threshold we rank correlation between ξ and the completeness, and between ξ and the number of properties in $\hat{P}_w$. We used for that the Spearman's rank correlation coefficient ρ.

[6]itemsets used in these experiments are available at: `http://cedric.cnam.fr/~hamdif/upload/cpmining2/`

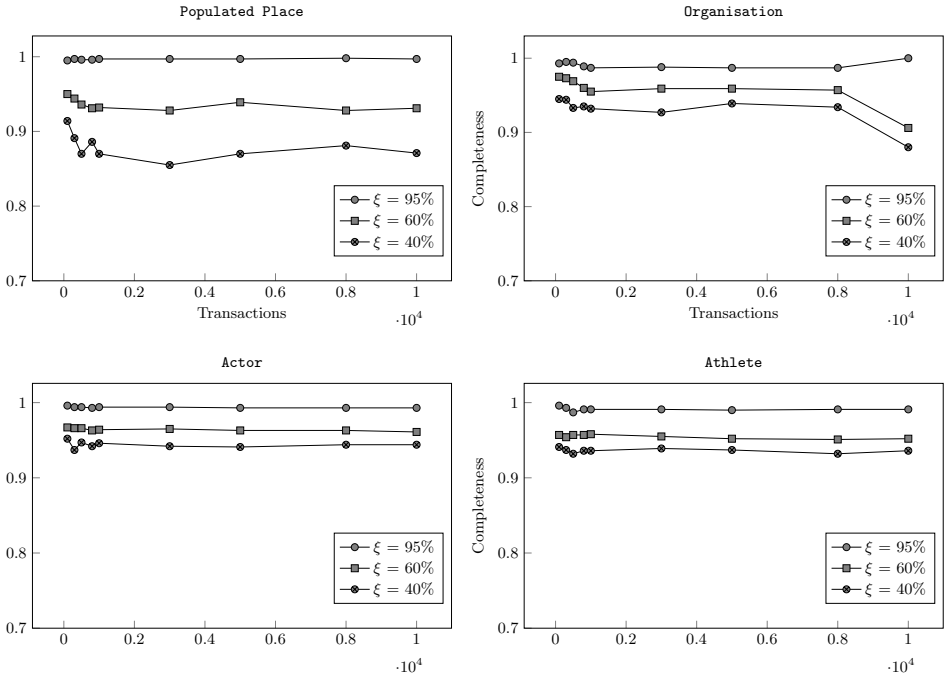

Figure 1: Completeness of DBpedia categories when varying the number of transactions and the minimum support ξ

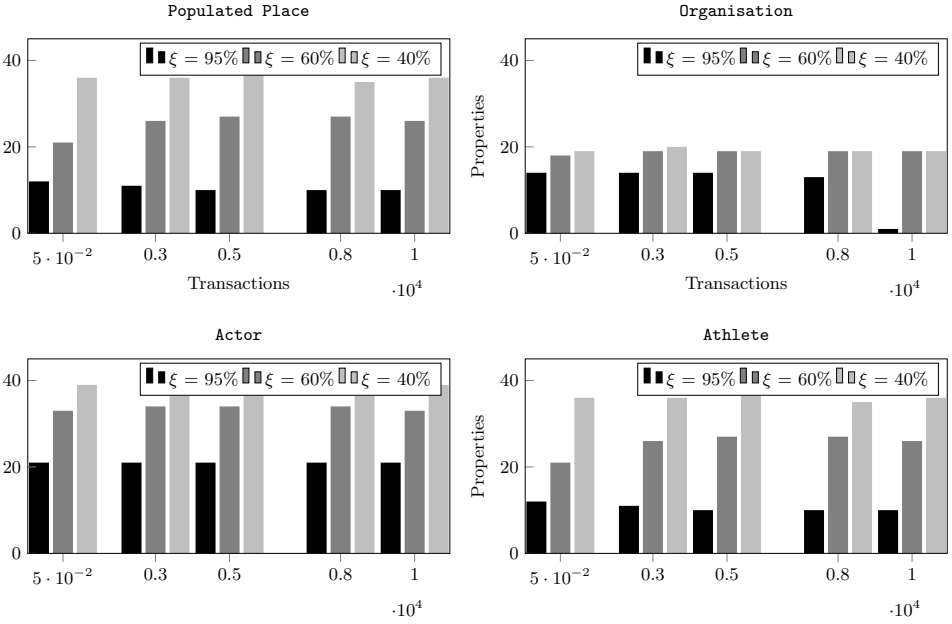

Figure 2: The number of properties in $\hat{P}_w$, for each DBpedia category, when varying the number of transactions and the minimum support ξ

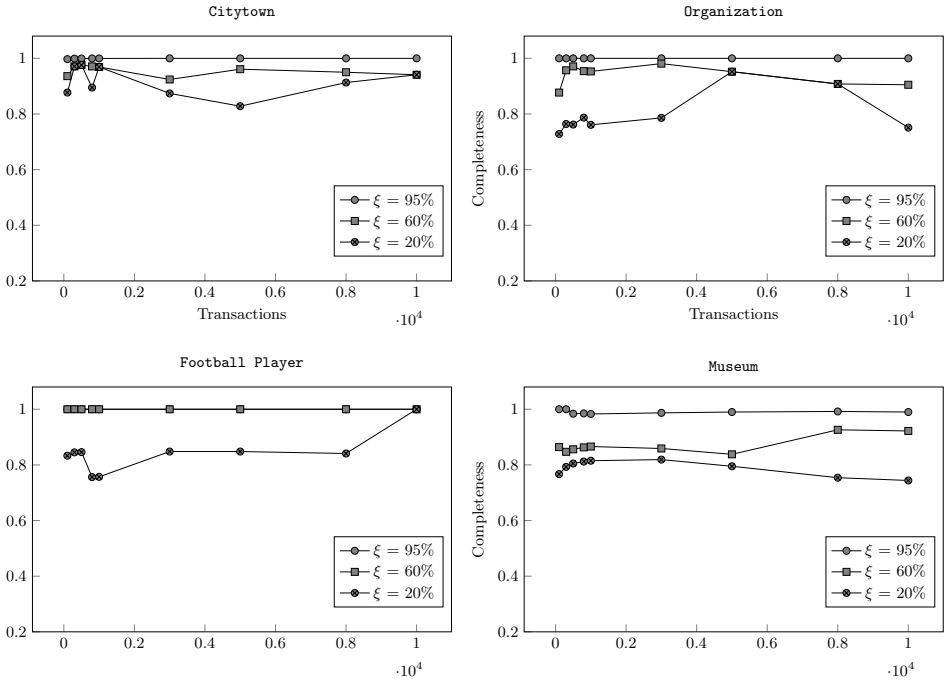

Figure 3: Completeness of Freebase categories when varying the number of transactions and the minimum support ξ

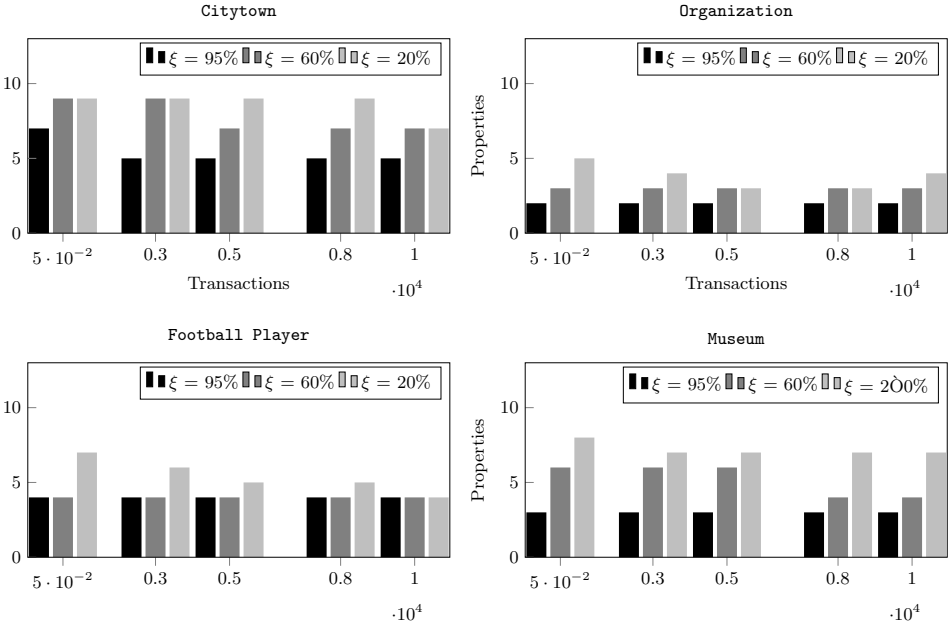

Figure 4: The number of properties in $\hat{P}_w$, for each Freebase category, when varying the number of transactions and the minimum support ξ

The values obtained for the sample of our experiments (by considering all DBpedia and Freebase categories, and all transactions sizes) are presented in the table 3.

To determine the significance of ρ, we used a t-distribution with $n-2$ degrees of freedom, and the standardized t statistic $t = \rho\sqrt{\frac{n-2}{1-\rho^2}}$ (where $n = 27$ is the size of our sample). For both DBpedia and Freebase categories , the positive correlation of ranks between ξ and the completeness is statistically significant at the 0.0005 level of significance (confidence level 99.9%), which is very well within the region of rejection of the null hypothesis. As a result, we have a large positive correlation between the user-specified threshold and the completeness. Concerning $|\hat{P}_w|$ a negative correlation with ξ seems to be obvious. Indeed, the negative correlation of the DBpedia categories is statically significant at the 0.005 level of significance. However, when we interpret ρ for the Freebase categories, we found that it is only significant at 0.05 (which still a good correlation). This is due to the fact that, as we can see in figure 4, there is no significant variation in the number of properties when ξ varies (especially for the *Football Player* category).

5.4 Discussion

Experiments that we carried out, show that the completeness metric provides meaningful values according to a pattern (or reference schema) mined from a dataset.

The first observation regarding the experiments concerns the behavior of the completeness metrics. The results obtained demonstrated the robustness of the proposed measure regarding the size of the dataset and even the nature of data. The higher stability in the values obtained for DBpedia compared to those of Freebase could be related to the fact that DBpedia has an underlying ontology that has been created based on the most commonly used infoboxes within Wikipedia. Freebase however, imports data from a wide variety of sources leading to less homogeneity in data descriptions. Moreover, we notice that values are better for categories such as *Actors* that are popular and probably attracts more publication effort than for *organization*.

The second observation is related to the user-defined threshold. This parameter refers to the desired expressiveness of the mined schema. The experiments showed a high correlation with the completeness of the datasets regarding this schema. This means that our approach is able to provide the user with the set of properties that ensures a degree of completeness meeting his/her requirements. Such a characteristics is very interesting and has several practical impacts. First, it provides a way to help expressing meaningful queries over RDF datasets when a schema describing the data source structure is missing. This schema, in addition to providing a query structure, ensures a degree of completeness that the user could control through the specification of the threshold. Note that the inferred pattern/schema, besides assuring good completeness values, has also a good expressiveness as it contains a significant number of properties (around 40 properties for some DBpedia categories). In the case of Freebase, the pattern is less expressive, as it is not greater than 12 (for considered categories in our experiments). So as a conclusion, we can say that, the more instances descriptions are homogeneous, the more completeness values are close (for any subset), and the more expressive are the inferred schemes.

Finally, once the completeness is measured regarding a reference schema for a given data source, the properties composing the schema could constitute good candidates for interlinking the dataset with external sources. Indeed, the completeness values obtained for the inferred schema will propagate through the links and assure a certain completeness for the external source.

6. RELATED WORKS

Information quality attracted many research works since two decades. It is an interesting theoretical as well as practical domain and several projects proposed methodologies to help dealing with quality assurance in traditional business information systems [21, 2, 4]. From the Web of Data perspective, the problem is a rather new issue. Researchers pointed out the fact that making data accessible is not sufficient especially when the users are companies and governmental agencies and especially when the target usage is business, research or countries' security. The credibility of data sources is thus bound to the quality of the content.

Several proposals could be classified using the 4 categories from the TDQM community [29] namely: *Intrinsic* (accuracy, reputation, believability and provenance), *Representational* (understandability, consistency and conciseness), *Accessibility* (accessibility and security), and *Contextual* (amount of data, relevance, completeness and timeliness). Intrinsic quality relies on internal characteristics of the data during evaluation. Most of the proposal concentrate on provenance data elicitation [25, 22, 15, 10]. From an external point of view, representational quality is concerned with factors influencing users interpretations and practices such as understandability [23] or data conciseness [30]. Concerning accessibility, a very important criteria in the context of web data, we could cite [18] where authors discussed common RDF publishers' errors that have a direct impact on data accessibility. Finally, contextual quality means that data could not be said "good" or "poor" without considering the context in which it is produced or used. To asses timeliness, the work presented in [16] relies on data provenance whereas authors in [5] propose metrics measuring the amount of data appropriate to meet associations mining requirements. The relevance dimension is considered from a ranking point of view in [8] and in the context of heterogeneous web data search in [17].

A quality criterion that remains critical and relies heavily on context is completeness. However, one can rarely find data having no missing entries. The incompleteness could have several reasons, ranging from human omission or misunderstanding during acquisition to data processing leading to lost data. The role of completeness evaluation is then to add information about data completeness to avoid biased analysis results and invalid conclusions. A first research direction focus on query answers completeness. We could cite [7] where authors introduced a framework to specify completeness statement on RDF data. To evaluate the completeness of data sources, Fürber at al. [9] distinguishes schema completeness from population or data completeness. The

Table 3: Spearman rank correlation ρ between ξ and $\mathcal{CP}$, and ξ and $|\hat{P}_w|$

$\mathcal{CP}$	$\lvert\hat{P}_w\rvert$	$\mathcal{CP}$	$\lvert\hat{P}_w\rvert$	$\mathcal{CP}$	$\lvert\hat{P}_w\rvert$	$\mathcal{CP}$	$\lvert\hat{P}_w\rvert$
Populated Place		Organisation		Actor		Athlete	
0.94	−0.82	0.9	−0.68	0.94	−0.81	0.94	−0.82
Citytown		Football Player		Museum		Organization	
0.81	−0.56	0.93	−0.73	0.82	−0.36	0.78	−0.81

former measures to which extent classes and properties are taken into account in data description. The second however, measures how complete is the population represented in the data compared to a complete population. A similar vision of completeness is adopted in [24] through extensional and intentional completeness. These definitions and the related metrics are very close to those used in the context of relational databases where the schema is the database schema and where the population refers to the database instances that is compared to an assumed "real world".

Note that the limitation of such a vision is that we should assume a closed-word assumption where a gold standard exists and could be used for both schema and population completeness assessment. However, our proposal, that addresses the schema completeness issue, do not rely on such an assumption, moreover difficult to meet. We propose to compute an intrinsic completeness value of the data source using both its extension (properties) and intention (instances). The underlying assumption is that, as data sources are populated by several persons and/or originate from several sources, the more frequent schema is likely to be a "consensus" schema.

7. CONCLUSION

We have presented in this article an approach for evaluating the completeness of RDF linked data sources. The approach is a two steps process that first computes a plausible schema as a set of properties and then computes the completeness of the related data source based on the discovered schema. Both schema and completeness computations use only the information from the data source without need of a gold standard.

The properties composing the schema are obtained by applying the well known FP-growth algorithm with an underlying assumption stating that a more frequent schema is likely to be more relevant. This solution is a compromise between the gold-standard approach difficult to apply in practice, and a brute force approach leading to irrelevant results.

The approach has been evaluated on two datasets that are DBpedia and Freebase. Several experiments have been conducted to evaluate the completeness measure robustness regarding the number of instances and the user-specified threshold. We have also analyzed how the approach behaves when varying several parameters factors. We have, for instance, studies the impact of the threshold and the number of transactions on the number of properties composing the mined schema.

Our analysis revealed some interesting characteristics allowing the characterization of the sources and the behavior of the community that maintains each of the data sources. The results show a low size of the computed schema for some categories such as *Organization* indicating the high heterogeneity of properties in the schema of data sources.

In the future, we will study how to improve the completeness of data sources by concentrating the effort on the more relevant properties. We will also work on how to use the completeness values to improve query expressions to increase the completeness of returned results.

References

[1] D. P. Ballou and H. L. Pazer. Modeling completeness versus consistency tradeoffs in information decision contexts. *Knowledge and Data Engineering, IEEE Transactions on*, 15(1):240–243, 2003.

[2] C. Batini, C. Cappiello, C. Francalanci, and A. Maurino. Methodologies for data quality assessment and improvement. *ACM Computing Surveys (CSUR)*, 41(3):16, 2009.

[3] S. Bechhofer, I. Buchan, D. De Roure, P. Missier, J. Ainsworth, J. Bhagat, P. Couch, D. Cruickshank, M. Delderfield, I. Dunlop, et al. Why linked data is not enough for scientists. *Future Generation Computer Systems*, 29(2):599–611, 2013.

[4] L. Berti-Equille, I. Comyn-Wattiau, M. Cosquer, Z. Kedad, S. Nugier, V. Peralta, S. Si-said Cherfi, and V. Thion-Goasdoué. Assessment and analysis of information quality: a multidimensional model and case studies. *IJIQ*, 2(4):300–323, 2011.

[5] P. Chen and W. Garcia. Hypothesis generation and data quality assessment through association mining. In F. Sun, Y. Wang, J. Lu, B. Zhang, W. Kinsner, and L. A. Zadeh, editors, *Proceedings of the 9th IEEE International Conference on Cognitive Informatics, ICCI 2010, July 7-9, 2010, Beijing, China*, pages 659–666. IEEE, 2010.

[6] E. F. Codd. Missing information (applicable and inapplicable) in relational databases. *SIGMOD Record*, 15(4):53–78, 1986.

[7] F. Darari, W. Nutt, G. Pirrò, and S. Razniewski. Completeness statements about RDF data sources and their use for query answering. In H. Alani, L. Kagal, A. Fokoue, P. T. Groth, C. Biemann, J. X. Parreira, L. Aroyo, N. F. Noy, C. Welty, and K. Janowicz, editors, *The Semantic Web - ISWC 2013 - 12th International Semantic Web Conference, Sydney, NSW, Australia, October 21-25, 2013, Proceedings, Part I*, volume 8218 of *Lecture Notes in Computer Science*, pages 66–83. Springer, 2013.

[8] C. M. Eastman and B. J. Jansen. Coverage, relevance, and ranking: The impact of query operators on web search engine results. *ACM Transactions on Information Systems (TOIS)*, 21(4):383–411, 2003.

[9] C. Fürber and M. Hepp. Swiqa-a semantic web information quality assessment framework. In *ECIS*, volume 15, page 19, 2011.

[10] J. Golbeck. Combining provenance with trust in social networks for semantic web content filtering. In L. Moreau and I. T. Foster, editors, *Provenance and Annotation of Data, International Provenance and Annotation Workshop, IPAW 2006, Chicago, IL, USA, May 3-5, 2006, Revised Selected Papers*, volume 4145 of *Lecture Notes in Computer Science*, pages 101–108. Springer, 2006.

[11] K. Gouda and M. J. Zaki. Efficiently mining maximal frequent itemsets. In *Proceedings of the 2001 IEEE International Conference on Data Mining*, ICDM '01, pages 163–170, Washington, DC, USA, 2001. IEEE Computer Society.

[12] G. Grahne and J. Zhu. Efficiently using prefix-trees in mining frequent itemsets. In B. Goethals and M. J. Zaki, editors, *FIMI '03, Frequent Itemset Mining Implementations, Proceedings of the ICDM 2003 Workshop on Frequent Itemset Mining Implementations, 19 December 2003, Melbourne, Florida, USA*, volume 90 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2003.

[13] J. Han, J. Pei, and Y. Yin. Mining frequent patterns without candidate generation. In W. Chen, J. F. Naughton, and P. A. Bernstein, editors, *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data, May 16-18, 2000, Dallas, Texas, USA.*, pages 1–12. ACM, 2000.

[14] J. Han, J. Pei, Y. Yin, and R. Mao. Mining frequent patterns without candidate generation: A frequent-pattern tree approach. *Data Min. Knowl. Discov.*, 8(1):53–87, Jan. 2004.

[15] O. Hartig. Trustworthiness of data on the web. In *Proceedings of the STI Berlin & CSW PhD Workshop*. Citeseer, 2008.

[16] O. Hartig and J. Zhao. Using web data provenance for quality assessment. In J. Freire, P. Missier, and S. S. Sahoo, editors, *Proceedings of the First International Workshop on the role of Semantic Web in Provenance Management (SWPM 2009), collocated with the 8th International Semantic Web Conference (ISWC-2009), Washington DC, USA, October 25, 2009*, volume 526 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2009.

[17] D. M. Herzig and T. Tran. Heterogeneous web data search using relevance-based on the fly data integration. In A. Mille, F. L. Gandon, J. Misselis, M. Rabinovich, and S. Staab, editors, *Proceedings of the 21st World Wide Web Conference 2012, WWW 2012, Lyon, France, April 16-20, 2012*, pages 141–150. ACM, 2012.

[18] A. Hogan, A. Harth, A. Passant, S. Decker, and A. Polleres. Weaving the pedantic web. In C. Bizer, T. Heath, T. Berners-Lee, and M. Hausenblas, editors, *Proceedings of the WWW2010 Workshop on Linked Data on the Web, LDOW 2010, Raleigh, USA, April 27, 2010*, volume 628 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2010.

[19] M. G. Institute, M. Chui, J. Manyika, J. Bughin, R. Dobbs, C. Roxburgh, H. Sarrazin, G. Sands, and M. Westergren. *The social economy: Unlocking value and productivity through social technologies*. McKinsey Global Institute, July 2012.

[20] R. J. B. Jr. Efficiently mining long patterns from databases. In L. M. Haas and A. Tiwary, editors, *SIGMOD 1998, Proceedings ACM SIGMOD International Conference on Management of Data, June 2-4, 1998, Seattle, Washington, USA.*, pages 85–93. ACM Press, 1998.

[21] Y. W. Lee, D. M. Strong, B. K. Kahn, and R. Y. Wang. Aimq: a methodology for information quality assessment. *Information & management*, 40(2):133–146, 2002.

[22] M. Markovic, P. Edwards, D. Corsar, and J. Z. Pan. The crowd and the web of linked data: A provenance perspective. In *Wisdom of the Crowd, Papers from the 2012 AAAI Spring Symposium, Palo Alto, California, USA, March 26-28, 2012*, 2012.

[23] P. Mendes, C. Bizer, Y. Young, Z. Miklos, J. Calbimonte, and A. Moraru. Conceptual model and best practices for high-quality metadata.

[24] P. N. Mendes, H. Mühleisen, and C. Bizer. Sieve: linked data quality assessment and fusion. In *Proceedings of the 2012 Joint EDBT/ICDT Workshops*, pages 116–123. ACM, 2012.

[25] T. Omitola, N. Gibbins, and N. Shadbolt. Provenance in Linked Data Integration. In S. Auer, S. Decker, and M. Hauswirth, editors, *Proc. of Linked Data in the Future Internet at the Future Internet Assembly, Ghent 16/17 Dec 2010*, volume 700 of *CEUR Workshop Proceedings ISSN 1613-0073*, February 2010.

[26] L. L. Pipino, Y. W. Lee, and R. Y. Wang. Data quality assessment. *Communications of the ACM*, 45(4):211–218, 2002.

[27] M. Samwald, A. Jentzsch, C. Bouton, C. S. Kallesøe, E. Willighagen, J. Hajagos, M. S. Marshall, E. Prud'hommeaux, O. Hassanzadeh, E. Pichler, et al. Linked open drug data for pharmaceutical research and development. *Journal of cheminformatics*, 3(1):19, 2011.

[28] Y. Wand and R. Y. Wang. Anchoring data quality dimensions in ontological foundations. *Communications of the ACM*, 39(11):86–95, 1996.

[29] R. Y. Wang and D. M. Strong. Beyond accuracy: What data quality means to data consumers. *Journal of management information systems*, pages 5–33, 1996.

[30] A. Zaveri, A. Rula, A. Maurino, R. Pietrobon, J. Lehmann, S. Auer, and P. Hitzler. Quality assessment methodologies for linked open data. *Submitted to Semantic Web Journal*, 2013.