SWIQA – A SEMANTIC WEB INFORMATION QUALITY ASSESSMENT FRAMEWORK

Christian Fürber

Martin Hepp

Follow this and additional works at: http://aisel.aisnet.org/ecis2011

Recommended Citation
http://aisel.aisnet.org/ecis2011/76

This material is brought to you by the European Conference on Information Systems (ECIS) at AIS Electronic Library (AISeL). It has been accepted for inclusion in ECIS 2011 Proceedings by an authorized administrator of AIS Electronic Library (AISeL). For more information, please contact elibrary@aisnet.org.
SWIQA – A SEMANTIC WEB INFORMATION QUALITY ASSESSMENT FRAMEWORK

Fürber, Christian, Bundeswehr University Munich, Werner-Heisenberg-Weg 39, 85577 Neubiberg, Germany, christian@fuerber.com

Hepp, Martin, Bundeswehr University Munich, Werner-Heisenberg-Weg 39, 85577 Neubiberg, Germany, mhepp@computer.org

Abstract

The internet is currently evolving from the "Web of Documents" into the "Web of Data" where data is available on web-scale in the so called Semantic Web (1) to retrieve information or (2) for data reuse, e.g. within applications for a higher degree of automation. At present, there is already a lot of data available on the Semantic Web, but unfortunately we do not know much about their quality due to missing techniques and methodologies for information quality assessment. In this paper, we provide a framework for information quality assessment of Semantic Web data called SWIQA by solely using Semantic Web technologies. Other than survey-based techniques for information quality assessment SWIQA employs data quality rule templates to express quality requirements which are automatically used to identify deficient data and calculate quality scores. Hence, using our approach minimizes manual effort while providing transparency about the quality of Semantic Web data. SWIQA may, therefore, be used by data consumers to find high quality data sources or by data owners to keep track of the quality of their own data.

Keywords: Semantic Web, Linked Data, Information Quality, Data Quality, Trust, SPARQL.
1 Introduction

One of the major Semantic Web visions is the supply of meaningful data on web scale that can be interpreted by humans and machines, e.g. to gain a higher degree of automation. While the traditional web was mainly designed to publish documents, the Semantic Web is designed to publish data including the semantics of data with the use of shared vocabularies and data annotations through ontologies (Berners-Lee, Hendler and Lassila, 2001). Ontologies are thereby a formal conceptualization of a domain of interest (Uschold and Gruninger, 1996). Hence, the meaning of data can also be understood by machines. At present, there are a lot of data sets available on the Semantic Web, e.g. containing geographical information, information about books, films, music, television and radio programs, census information, etc. (Bizer, Heath and Berners-Lee, 2009). Moreover, several companies such as BestBuy and O’Reilly have started to publish data about products and services using the GoodRelations ontology, a shared vocabulary for E-Commerce (Hepp, 2008). Hence, the growth of data on the Semantic Web is currently subject to significant progress, also from business-perspective. As the amount and thereby the usage of Semantic Web data grows, methodologies will be required to identify and manage the quality of the published information. Unfortunately, the Semantic Web has not been addressed much by information quality research, yet. Similar to data in information systems and databases, Semantic Web data suffers from quality problems, e.g. as described in (Fürber and Hepp, 2010a; Fürber and Hepp, 2010b; Hogan, Harth, Passant, Decker and Polleres, 2010).
Without the existence of methodologies for information quality management of Semantic Web data, we will only discover information quality problems when our Semantic Web applications fail or our queries return incorrect results. Particularly, we need to know about the quality state of a Semantic Web source (1) for the selection of appropriate data, (2) to decide how much we can trust the provided information, and (3) to detect data quality problems as a basis for data quality improvement. Moreover, by providing data quality management approaches based on ontologies, we facilitate the reuse of data quality rules for multiple different information systems which may be integrated using domain ontologies, e.g. via wrapping technologies such as D2RQ or converted to knowledge bases, for the centralized application of data quality management activities.

In this paper, we provide a Semantic Web Information Quality Assessment Framework called SWIQA that classifies data quality problems and calculates information quality (IQ) scores for the IQ dimensions syntactic accuracy, semantic accuracy, completeness, uniqueness, and timeliness based on previously defined data quality rules. We thereby understand information quality assessment as the "process of assigning numerical and categorical values to IQ dimensions" (Ge and Helfert, 2008). Our proposed methodology attempts to raise the level of objectivity when judging the quality of Semantic Web sources. Therefore, data quality rule templates were designed to capture and structure quality-relevant domain knowledge, so that it can be used to calculate quality scores. We successfully applied SWIQA on real-world Semantic Web data from BestBuy. Throughout this paper, we use the terms “data quality” and “information quality” interchangeably unless specified otherwise.

2 Data Quality in the Semantic Web

2.1 Data in the Semantic Web

Semantic Web data is typically represented in RDF (Resource Description Framework) (cf. McBride, 2004). The core structure of RDF is represented by triples, which allow the definition of statements in

---

1 http://www4.wiwiss.fu-berlin.de/bizer/d2rq/
a subject, predicate, object format, e.g. John hasFather Jack. This combination of two entities (subject, object) and a relationship (predicate) is called a triple. Hence, with Semantic Web programming languages like RDF or OWL (Web Ontology Language) it is possible to define relationships between things, similar to an entity relationship model or a database schema, but with more expressive instruments, such as class relationships, disjointness axioms, or class restrictions. This formal conceptualization of a domain of interest (Gruber, 1993) is known as an ontology. The data values of Semantic Web resources are structured through an ontology and are typically represented in the object position as literal values of so called datatype properties. In our approach, we do not regard the quality of the schema, i.e. of the ontology. We rather focus on literal values and instances that are structured by ontologies.

2.2 The Role of Data Quality Rules for Defining Information Quality

<table>
<thead>
<tr>
<th>Data Quality Rule</th>
<th>Definition</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mandatory Property and Literal Rules</td>
<td>Properties and their literals become mandatory, if the data is required for the task at hand.</td>
<td>The properties indicating the geographical coordinates must exist and have values for all instances of the class foo:Location to be able to navigate to each location.</td>
</tr>
<tr>
<td>Syntactic Rules</td>
<td>Syntactic rules define the type of characters and/or the pattern of literal values.</td>
<td>Literal values for the property foo:country-name must only contain letters.</td>
</tr>
<tr>
<td>Functional Dependency Rules</td>
<td>Functional dependencies are dependencies between the values of two or more different properties.</td>
<td>The literal value for foo:city is always dependent to the literal value for foo:country, since certain city names only exist in certain countries.</td>
</tr>
<tr>
<td>Legal Value Rules</td>
<td>Legal value rules are the explicit definition of the allowed values for a certain property.</td>
<td>The property foo:gender must only contain the values “male”, “female”, “m”, or “f”.</td>
</tr>
<tr>
<td>Legal Value Range Rules</td>
<td>Legal value range rules are the explicit definition of the allowed value range for properties holding numerical values. A value range contains an upper and/or lower limit.</td>
<td>The property foo:population must only contain values greater than zero.</td>
</tr>
<tr>
<td>Illegal Value Rules</td>
<td>Illegal value rules are the explicit definition of the values that may not be assigned to a certain property.</td>
<td>The property foo:gender may never contain the value “mail”.</td>
</tr>
<tr>
<td>Illegal Value Range Rules</td>
<td>Illegal value range rules are the explicit definition of prohibited value ranges for properties holding numerical values. A value range contains an upper and/or lower limit.</td>
<td>The property foo:population must not contain values less than one.</td>
</tr>
<tr>
<td>Unique Value Rules</td>
<td>Unique value rules define properties that may contain each literal value not more than once within a defined collection of values.</td>
<td>Each value for property foo:ISBN in instances of class foo:Book may not occur more than once.</td>
</tr>
<tr>
<td>Outdated Value Rules</td>
<td>Outdated value rules are rules that identify instances that represent an outdated state of the corresponding real world entity.</td>
<td>Instances of the class foo:Offer are outdated, if its value for foo:validThrough is elder than the current date and time.</td>
</tr>
</tbody>
</table>

Table 1. Generic data quality rules of SWIQA

Information quality may be defined from different perspectives (Ge and Helfert, 2007). (Wang and Strong, 1996) analyzed the consumer-centric perspective in detail and thereby identified the 15 most important information quality dimensions according to data consumers. In their empirical study, they
assumed information to be of high quality when they are “fit for use by data consumers”. Hence, this definition of information quality relies on the subjective requirements of data consumers. According to (Redman, 2001) “data are of high quality when they are fit for their intended uses in operations, decision making, and planning.” This latter definition centralizes the tasks that need to be performed with the data, rather than the subjective requirements of data consumers. The above definitions imply that requirements (of data consumers and tasks) serve as a benchmark to determine data quality. Thus, we understand data quality in a perspective-neutral sense as the degree to which data fulfils quality requirements. According to (Pipino, Lee and Wang, 2002), information quality may be assessed based on task-dependent and task-independent metrics. Task-independent metrics “reflect states of the data without the contextual knowledge of the application and can be applied to any data set” (Pipino et al., 2002). Thus, task-independent metrics are valid for every task the data is used for, while task-dependent metrics are only valid for certain tasks. Through the identification of task-dependent and task-independent metrics, we can define the required quality of data in general and for performing certain tasks. In this paper we regard such metrics as quality requirements which can be used to determine the level of data quality of a specific data source. To facilitate a structured acquisition of quality requirements, we defined generic data quality rules focused on the quality of literal values and instances, as listed in table 1. The data quality rules are the foundation for the assessment metrics of SWIQA and are based on an analysis of data quality problem typologies from database-oriented research (Leser and Naumann, 2007; Oliveira, Rodrigues and Henriques, 2005; Oliveira, Rodrigues, Henriques and Galhardas, 2005; Rahm and Do, 2000).

3 SWIQA – A Semantic Web Information Quality Assessment Framework

3.1 Calculating Information Quality Scores

Based on the data quality rules, we attempt to compose metrics to assess the quality of Semantic Web data sources for information quality dimensions that are measurable via data analysis without further metadata annotations. Thereby, we aim to provide an assessment methodology that minimizes subjectivity and thereby allows a task-oriented judgment of Semantic Web data sources. Since not all information quality dimensions can be measured solely based on data quality rules, we do not regard dimensions in our model that can only be assessed through surveys. Based on the “dimensions and metrics” summary in (Batini, Cappiello, Francalanci and Maurino, 2009) we identified five dimensions that can be measured by applying our data quality rules, namely syntactic and semantic accuracy, completeness, timeliness2, and uniqueness. Moreover, our information quality score metrics are based on the simple ratio calculation as described by (Pipino et al., 2002). The simple ratio is measured by subtracting the ratio between the total number of instances that violate a data quality rule (DQRV) and the total number of relevant instances (T) from one. The results of formulas (1), (3), and (4) lie between zero representing the worst state of data quality and one representing the perfect state of data quality. Formula (1) below shows the calculation of individual IQ scores that can be applied for each property of the underlying data set. Formula (2) shows an adjusted version of the individual IQ score which additionally multiplies the ratio by a weighting factor ω representing the importance of the property for the intended task. This latter calculation is computed as an intermediate step to calculate an aggregated IQ score for each IQ dimension with formula (3) that accounts for the different importance levels of data for the performed task. We, therefore, calculate the weighted information quality scores of each property and divide their sum by the sum of all weighting factors of the regarded properties (W). This weighted IQ score requires importance annotations of all tested

2 We do not differ between currency and timeliness in this paper
properties and classes with SWIQA’s annotation property `dqm:importance`. We suggest assigning integer values ranging from one meaning "slightly important" to five meaning "task critical".

\[
\text{IQ-Score} = (1 - (DQRV / T)) \\
\text{IQ-Score}_w = (1 - (DQRV / T)) \times \omega \\
\text{Aggregated-IQ-Dim-Score}_w = \sum (\text{IQ-Score}_w) / W \\
\text{Aggregated-IQ-Dim-Score} = \sum (\text{IQ-Score}) / P
\]

In cases with equal importance of the property data for the task at hand or in cases where it is not possible to annotate importance values, formula (4) may be used to assess an unweighted score for each IQ dimension. Instead of using the sum of all weighting factors in the denominator, we divide the sum of all individual IQ scores by the number of all tested properties P to calculate the unweighted IQ dimensional score. In the following sections, we explain the semantics and composition of the IQ assessment metrics for each IQ dimension of SWIQA.

### 3.1.1 Accuracy

In general, there is no common agreement on the definition of accuracy in data quality research. According to (Batini and Scannapieco, 2006), the accuracy dimension describes the proximity of data value representations of an object related to their real-world states. Furthermore, they distinguish between (1) syntactic accuracy and (2) semantic accuracy. A value is syntactically accurate, when it is part of a legal value set for the represented domain or it does not violate syntactical rules defined for the domain (cf. Batini and Scannapieco, 2006). E.g. if we have an instance representing a car that has the value “red” describing its color, but the car we have attempted to describe is really green, then the value “red” will still be syntactically accurate when “red” is part of the legal value set. Semantic accuracy also regards whether the data value represents the correct state of an object (Batini and Scannapieco, 2006). Thus, in our example the instance will also be semantically accurate, if the literal value that indicates the color of the car represents its correct real-world state “green”. For the assessment of syntactic accuracy, we use (1) legal value rules, (2) legal value range rules, or (3) syntactic rules. The definition of legal value rules is the most exact, but also the most restrictive option, since it defines the exact value patterns. In our understanding, only one of these data quality rules should be chosen to identify incorrect values for each property, since they provide different levels of precision. Hence, we integrated all three rules into separate queries of SWIQA for IQ assessment of syntactic accuracy with formulas (1) – (4). Although it is very difficult to assess semantic accuracy, we may identify many semantically incorrect values through the definition of functional dependency rules. With SWIQA we integrated queries for identification of functional dependency violations in formulas (1) – (4) to assess semantic accuracy. The parameter T thereby represents the total number of relevant instances of the class in which the functional dependency violation may occur. This may only be a subset of all instances within a class. The aggregated IQ dimension score for semantic accuracy calculated with formula (4) represents the percentage of data in the tested data set that does not violate any known functional dependencies. Formula (3) calculates an adjusted score in which functional dependency violations for task-critical data have a stronger influence on the quality score than violations of less important data.

### 3.1.2 Completeness

Completeness is defined by (Wang and Strong, 1996) as “The extent to which data are of sufficient breadth, depth, and scope for the task at hand”. Hence, we can differentiate between schema completeness, population completeness, and column completeness (cf. Batini and Scannapieco, 2006; Pipino et al., 2002). In terms of the Semantic Web, schema completeness can be defined as the degree
to which elements of the ontology, i.e. classes and properties, are represented. Thus, it could be called “ontology completeness”. Population completeness refers to whether all objects of the real-world reference are represented (Batini and Scannapieco, 2006; Pipino et al., 2002). Column completeness is “a measure of the missing values for a specific property or column” (Batini and Scannapieco, 2006). In terms of the Semantic Web, we can call it “property completeness”. Property completeness may be assessed by the definition of mandatory property and literal value rules. We thereby can also integrate checks for conditional mandatory properties and literal values, i.e. properties and literal values that are required for a certain subset of the instances. We calculate the property completeness scores by using formulas (1) – (4) from section 3.1. The number of instances that violate a data quality rule (DQRV) is thereby composed by the number of missing literal values and the number of missing properties. In cases with conditional mandatory properties, we need to define the total number of instances \( T \) only on the relevant subset to receive representative completeness scores. The aggregated IQ dimension score for completeness in (3) represents the degree to which the required properties are completely available accounting for their different levels of importance for the task at hand. Formula (4) represents an unweighted score for property completeness. Since we currently focus on data values, we do not regard schema completeness, nor do we integrate population completeness in our model at present.

3.1.3 Timeliness

According to (Pipino et al., 2002) timeliness “reflects how up-to-date the data is with respect to the task it’s used for.” In most current scenarios of the Semantic Web the data is usually being created and modified by other source systems, e.g. relational databases. In other scenarios, Semantic Web data already provides expiration dates of data instances, e.g. data published via the GoodRelations ontology. In both cases, we can use data quality rules for the identification of outdated literal values to approximate the timeliness of data. Assuming that at least timestamps about the last modification of data are available in source and target data source, we can compare the two timestamps to identify potential obsolescence of data. Thus, the outdated value rule checks whether there are instances in the Semantic Web source that have a date of modification \( ModTime_{SemWeb} \) elder than the last modification time of the source instance \( ModTime_{Source} \) as proposed in formula (5). Although we do not calculate whether the represented values in the source represent the current real-world state of an object, we can at least identify all instances that represent an outdated state in comparison to its data source and thereby approximate timeliness assuming the data source has the closest data representation compared to its real-world state. In cases where an expiry date of an instance is available we may instead use equation (6) to identify outdated instances:

\[
OV_i = ModTime_{SemWeb} < ModTime_{Source} \quad (5)
\]

\[
OV_i = ExpiryTime < NOW \quad (6)
\]

Other than for the previous dimensions, the timeliness queries of SWIQA do not calculate individual scores for each property. Instead the scores refer to whole classes, since it will usually be the whole instance that holds the time data. Hence, our timeliness scores calculated with formulas (1) and (2) represent the timeliness of the class which holds the tested instances. Likewise to the previous dimensions, SWIQA also calculates aggregated IQ-Dimension-Scores for timeliness using formulas (3) and (4).

3.1.4 Uniqueness

It is not only possible that data or schema elements are missing in the source, but also that these elements are represented multiple times. Thus, we understand uniqueness as the degree to which data
is free of redundancies in breadth, depth, and scope (cf. Batini et al., 2009). Uniqueness in breadth is the degree to which an ontology is free of redundancies regarding its represented classes and properties. Uniqueness in scope in our understanding may be defined as the degree to which a knowledge base has multiple different instances to represent the same object. Regarding this latter definition it must be noted that redundancies in the form of synonyms are explicitly allowed on the Semantic Web and may be flagged by elements such as the \texttt{owl:sameAs} property or functional properties. Hence, assessment methodologies for uniqueness in scope should only be used in settings where synonymous instances are not wanted. With SWIQA we concentrate on assessing the uniqueness in depth which represents the degree to which values of a property are unique. Uniqueness in depth is only suitable for properties that must assign each value of their domain only to one instance. It may be assessed by integration of uniqueness rules into equations (1) - (4) from section 3.1 resulting in weighted and unweighted scores for each property and the whole dimension. The results of formula (3) and (4) thereby represent a task-dependent and task-independent degree to which the values of the tested data conform to the defined unique value rules.

![Figure 1. Proposed configuration of data quality rules for information quality assessment](image)

### 3.2 Architecture of SWIQA

The current architecture of SWIQA is composed by three basic layers: (1) the data acquisition layer, (2) the query layer, and (3) the ontology layer. In the data acquisition layer, we have to retrieve all data sets that are of interest for information quality assessment. The retrieved data sets can be subject to information quality assessment or they can be used as trusted references, e.g. for the definition of legal values or legal functional dependencies as explained in (Fürber and Hepp, 2010a). Through techniques for wrapping relational databases in RDF, such as D2RQ, it is, moreover, possible to retrieve relational data sources for these purposes. The query layer of SWIQA contains a library of query templates which can be used to express data quality requirements to be processed for data quality problem classification and IQ assessment. The query templates were designed based on the SPARQL\(^3\) Inferencing Notation (SPIN)\(^4\) and hold parameters for Uniform Resource Identifiers (URIs) of classes and properties involved in the assessment. Before starting the quality assessment, the parameters have to be adjusted to the actual data set. This can be performed via forms, so that programming knowledge.

---

\(^3\) SPARQL is the Semantic Web query language for RDF, specification available at [http://www.w3.org/TR/2008/REC-rdf-sparql-query-20080115/](http://www.w3.org/TR/2008/REC-rdf-sparql-query-20080115/)

\(^4\) [http://spinrdf.org/spin.html](http://spinrdf.org/spin.html)
is not required. Furthermore SPIN allows the usage of SPARQL queries as constraints and inferencing rules. We used the latter option to automate our problem classification and IQ assessment. The assessment queries were built to compute weighted and unweighted IQ scores as described in the previous sections and to classify the identified problems into a data quality problem ontology. Finally, the ontology layer consists of vocabulary elements defined in the dqm-namespace for (1) the annotation of importance values to properties and classes of the tested data sets for the calculation of task-oriented IQ scores, (2) the assignment of inferred IQ scores to ontology elements of the tested data set, and (3) the classification of identified data quality problems.

3.3 Closing the World for Data Quality Management

Typically the Semantic Web assumes an open world which does not allow inferring the truth of a statement simply by checking whether the statement is known. This so called Open World Assumption (OWA) assumes that everything we do not know is not defined, yet. For data quality problem identification and the calculation of information quality scores we need to define metrics that close the world, i.e. assume that everything that is not known can be assumed as false (cf. Hebeler, Fisher, Blace and Perez-Lopez, 2009). In other words, SWIQA’s metrics assume that we have complete and correct knowledge about what data is correct and what data is wrong. Of course, this assumption will most likely not hold in many cases since we often suffer from incomplete knowledge and high uncertainty especially when starting to perform data quality management. Thus, the interpretation of our information quality scores and data quality problem classification must be careful and requires the consideration of the possibility of incompleteness of our metrics. Consequently, when performing data quality management the metrics and rules have to be validated and refined continuously. Moreover, it is necessary to define the data quality rules as complete and precise as possible before its application with SWIQA in order to receive accurate IQ scores.

4 Evaluation

The evaluation of SWIQA can be separated into two parts. We first evaluated whether the algorithms of SWIQA correctly identify all known data quality rule violations as described in section 2.2 based on a test data set containing different types of data quality problems. On the other hand, we evaluated the applicability of SWIQA on real-world Semantic Web data. For the evaluation of SWIQA’s algorithms, we used the precision/recall-methodology as explained in (Batini and Scannapieco, 2006). When applying data quality problem identification queries on data it is possible that data quality problems are not identified as such and, hence, are missing in the assessment metrics. Such unidentified data quality problems are called false negatives (FN). In contrast, it is also possible that clean data is incorrectly identified as a data quality problem and, therefore, wrongly integrated into our assessment results. In this case, we use the term false positives (FP). Precision measures how many real data quality problems (TP) are identified among the result set of our data quality problem classification queries including false positives (TP + FP). Recall measures how many data quality problems (TP) are identified out of the total number of data quality problems (TP + FN).

\[
\text{Precision} = \frac{TP}{TP + FP} \tag{7}
\]
\[
\text{Recall} = \frac{TP}{TP + FN} \tag{8}
\]

The test data set consisted of 220 triples with product and location data. The data set contained missing literals and properties, syntax violations, uniqueness violations, outdated instances, functional
dependency violations, out of range values, and illegal literal values. The locations of the errors were all known. Since we iteratively fixed the queries during the evaluation, the final queries achieved perfect results for precision and recall. Besides testing the effectiveness of our queries we also tested the applicability of SWIQA on real-world data sets. Therefore, we applied SWIQA on GoodRelations data from BestBuy\(^5\). We then analyzed the data set to identify the elements of the ontology that contain most of the data. Thereby we discovered that some instances were not directly classified, i.e. direct instances of a class. This was caused due to a design option in RDF. Thus, we had to develop a different variant of our assessment metrics that generate paths to blank node instances that shall be tested by SWIQA. After adjusting our assessment queries, we defined data quality rules for the BestBuy data set based on our generic assessment queries. We thereby focused on task-independent or obvious task-dependent rules, such as valid city country combinations, syntactic rules for email addresses and phone numbers, expired gr:validThrough dates, and the availability of address, contact, and opening hours data. Finally, we applied the rules on the data set. The data quality rules were represented by 33 different SPARQL queries which returned the result set within 18.31 minutes. The data set thereby contained in total 1,341,177 triples. Most of the triples (1,223,912) were retrieved from Geonames\(^6\), a publicly available data source for geographic data, as a trusted reference for valid city country combinations in the semantic accuracy assessment. Therefore, we saved the effort for defining legal city country combinations. It must be noted that we only integrated a relevant subset of Geonames data to reduce the overall size of the data set. The evaluation of SWIQA was performed on an AMD Athlon II X4 630 CPU with 6 GB RAM available for querying. The aggregated scores for the IQ dimensions of SWIQA are shown in table 2 below.

<table>
<thead>
<tr>
<th>IQ Dimension</th>
<th>Number of Problems</th>
<th>IQ-Dim Score</th>
<th>Number of tested properties</th>
</tr>
</thead>
<tbody>
<tr>
<td>Syntactic Accuracy</td>
<td>3</td>
<td>99 %</td>
<td>2</td>
</tr>
<tr>
<td>Semantic Accuracy</td>
<td>55</td>
<td>95 %</td>
<td>2</td>
</tr>
<tr>
<td>Completeness</td>
<td>55</td>
<td>99 %</td>
<td>25</td>
</tr>
<tr>
<td>Uniqueness</td>
<td>0</td>
<td>100 %</td>
<td>1</td>
</tr>
<tr>
<td>Timeliness</td>
<td>1,071</td>
<td>0 %</td>
<td>1 class</td>
</tr>
</tbody>
</table>

Table 2. SWIQA assessment results in BestBuy store data

Since we did not know the task importance of the ontology elements, we only assessed unweighted quality scores with formulas (1) and (4) from section 3.1. Since the queries of SWIQA also classify data quality problems, we were able to manually analyze the causes of the identified problems. Thereby, we discovered that most of the problems in the semantic accuracy dimension stem from incorrectly set white spaces at the end of the literals of the property v:locality. Since Geonames does not contain these white spaces at the end of the city name, our query could not find a matching literal in the reference data set. Thus, most of the identified problems which were integrated into the assessment of semantic accuracy were actually of syntactical nature. Besides the white space errors, we also identified three instances with cities that are actually located in Puerto Rico according to Geonames, but had a literal value "US" for the property v:country-name in the BestBuy data set. In this latter case, we assumed that cities located in Puerto Rico should have the value “PR”. Therefore, the three instances represent an inaccurate state of the country literals and belong to the semantic accuracy dimension. The low timeliness score of the BestBuy data set seems to be caused by an outdated expiry date indicated at the property gr:validThrough. A simple update of the literals of gr:validThrough would probably eliminate the problems and create a perfect timeliness score.

\(^5\) We retrieved the data on August 11th 2010 from http://stores.bestbuy.com/sitemap.xml

\(^6\) http://www.geonames.org/
However, the assessment queries of SWIQA sufficiently identified data quality problems and supplied scores to objectively measure the quality of the BestBuy data set. With the support of SWIQA we were able to quickly define quality requirements and assess the quality of Semantic Web data with minimal effort. Moreover, the classification of identified data quality problems allowed an analysis of the causes. During the evaluation of SWIQA we have learned that due to the different design decisions when modelling ontologies with RDF, it is difficult to predefine all generic variants of data quality rules. Thus, our library with generic information quality assessment queries will most likely be extended for other design variants with growing application of SWIQA. Consequently, one type of data quality rule will have different variants of queries in SWIQA. Regarding the performance of SWIQA we achieved sufficient results. With the use of scalable triple stores or simple random sampling techniques SWIQA may also be applicable on even larger data sets. The application of SWIQA on large data sets will be part of our future work.

5 Related Work

On high level, we distinguish between (1) information systems oriented approaches, (2) "Web of Documents" oriented approaches, and (3) Semantic Web oriented approaches for data quality assessment. The few approaches that focus on information quality assessment on the Semantic Web may be further subdivided into (1) provenance-based assessment, (2) rating-based assessment, and (3) rule-based information quality assessment. (Hartig and Zhao, 2009) proposed a provenance-based methodology to assess information quality of Semantic Web data. They exemplify their approach by calculating timeliness based on provenance data adapting the formulas of (Ballou, Wang, Pazer and Tayi, 1998). Although this approach is a major step towards transparency of data quality in the Web of Data, we can currently not rely on the availability of previously created provenance information. Moreover, (Hartig, 2009) provided a rating-based approach for the integration of trust values into SPARQL queries which is based on user-ratings about the trustworthiness of Semantic Web data. Since users may not only be focused on the judgment of the quality of data values, user-ratings may occasionally be misleading or even incorrect. Furthermore, (Bizer and Cyganiak, 2009) proposed a rule-based approach for filtering high quality information of web-based information systems according to user-defined policies. While this approach may be suitable to quickly fulfil the information needs of an information consumer in the World Wide Web, it neither provides help for the improvement of data quality, nor to gain transparency about the quality state of data. (Grüning, 2009) utilizes Semantic Web technologies for the annotation of data quality related information to data of information systems from the energy domain. The annotations have to be performed by domain experts and are used to train identification algorithms for the facilitation of automated annotation of data quality problems. Moreover, Grüning provides an ontology-based data quality assessment model. Unfortunately, this information systems oriented approach only considers a limited set of data quality problems and does not use the potential of publicly available Semantic Web data. Finally, there are many other data quality assessment approaches in information systems research that we cannot all describe in this paper due to the limited space. Therefore, we refer to (Batini et al., 2009) for an excellent overview about data quality assessment approaches predominantly focusing on data quality management of information systems. Although the achievement of proof and trust is explicitly stated in the well know Semantic Web layer cake, "which describes the main layers of the Semantic Web design and vision" (Antoniou and Harmelen, 2008), there is currently not much research work on Semantic Web oriented approaches. SWIQA tries to fill this gap by providing a rule-based methodology for information quality assessment of Semantic Web data that does not require additional provenance annotations.

6 Conclusion and Future Work

In this paper, we have introduced a rule-based information quality assessment framework for the Semantic Web called SWIQA. With SWIQA it is possible (1) to identify and classify data quality
problems, and (2) to calculate task-dependent and task-independent information quality scores for the IQ dimensions syntactic accuracy, semantic accuracy, completeness, uniqueness, and timeliness based on previously defined quality requirements via data quality rule templates. SWIQA is applicable for Semantic Web resources as well as for relational databases with the support of wrapping technologies, such as D2RQ. In contrast to rating-based approaches and provenance-based approaches, SWIQA does not require additional data annotations besides optional importance ratings for the assessment of task-dependent IQ-Scores. The optional use of importance annotations, allows assessing the fitness of data for the task at hand. Through the use of data quality rules based on previously captured quality relevant knowledge, SWIQA reduces subjectivity when assessing information quality. Moreover, the generalized SPARQL query templates allow an easy application of SWIQA for any Semantic Web data set disregarding its domain and without the need for programming knowledge. On the other hand, SWIQA is still a prototype implementation that requires the extension to a broader set of data quality rules and its variants, e.g. in the area of syntactical rules. Additionally, we have to stress that each data quality rule is currently assigned to exactly one IQ dimension in SWIQA for simplicity reasons. Although of the pragmatism of this configuration, it must be stated that certain data quality rules may impact more than one dimension in reality, e.g. outdated values influence at least timeliness and accuracy at the same time (Ge and Helfert, 2007). Our future work will investigate possibilities to correctly integrate data quality problems into IQ dimensions according to their semantics. Furthermore, the generalization of our metrics is rather costly regarding computational performance which may become a problem with large data sets. Thus, as part of our future work we will optimize and extend SWIQA's queries. Moreover, we will examine possibilities for the extension of SWIQA by additional data quality rules and IQ dimensions such as population completeness. However, SWIQA may support data quality improvement of Semantic Web data by providing transparency about data quality. We will continue to apply SWIQA on real-world data sets from the Semantic Web as well as from information systems for further improvements of our framework.

7 References


