An Objective Linked Data Quality Framework

Ahmad Assaf, Aline Senart [a] and Raphal Troncy [b]
[a] *SAP Research, SAP Labs France SAS,*
805 avenue du Dr. Maurice Donat, BP 1216, 06254 Mougins Cedex, France
e-mail: first.last@sap.com
[b] *EURECOM,*
2229 route des cretes, 06560 Sophia Antipolis, France
e-mail: raphael.troncy@eurecom.fr

Abstract. The standardization of Semantic Web technologies and specifications has resulted in a staggering volume of data being published. However, data should be of good quality to be integrated properly. In this paper, we propose an assessment framework for data quality that issues a certificate for a given dataset. This framework helps on one hand data owners to rate the quality of their datasets and get some hints on possible improvements, and on the other hand data consumers to choose their data sources from a ranked set. In a previous work, we identified potential quality issues of Linked Data and listed quality principles for all stages of data management. We refine this work here with a framework composed of objective quality indicators and associated metrics. For each indicator, we selected a set of tools and systems that can be used to rate datasets according to key quality principles. We show how the framework can be used to assess an existing dataset.

Keywords: Data Quality, Linked Data, Quality Framework, Semantic Web, Quality Framework

1. Introduction

In the last few years the Semantic Web gained a momentum supported by the introduction of many related initiatives like the Linked Open Data (LOD)[1]. From 12 datasets cataloged in 2007, the Linked Open Data has grown to almost 300 datasets containing almost 32 billion triples [6]. Data is being published by both public and private sectors and covers a diverse set of domains from life sciences to military. This success lies in the cooperation between data publishers and consumers. Users are empowered to find, share and combine information in their applications easily.

The Linked Open Data is a gold mine for organizations who are trying to leverage external data sources in order to produce more informed business decisions [10]. However, the heterogeneous nature of data sources reflects directly on the data quality as these sources often contain inconsistent as well as misinterpreted and incomplete information.

Traditional data quality is a thoroughly researched field with several benchmarks and frameworks to grasp its dimensions [32][44][50]. Data quality principles typically rely on many subjective indicators that

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

are complex to measure automatically. The quality of data in indeed realized when it is used [31], thus directly relating to the ability of satisfying users' continuous needs.

Web documents that are by nature unstructured and interlinked require different quality metrics and assessment techniques than traditional datasets. For example, the importance and quality of Web documents can be automatically calculated via algorithms like Page Rank [35].

Assuring data quality in Linked Open Data is another challenge. It consists of structured information supported by models, ontologies and vocabularies and contains query endpoints and links. This makes data quality assurance a challenge. Despite the fact that Linked Open Data quality is a trending and highly demanded topic, very few efforts are currently trying to standardize, track and formalize frameworks to issue scores or certificates that will help data consumers in their integration tasks.

In this paper, we propose a comprehensive objective framework to evaluate the quality of Linked Data sources. The framework helps on one hand data owners to rate the quality of their dataset and get some hints on possible improvements, and on the other hand data consumers to choose their data sources from a ranked set. The aim of this paper is to provide researchers and practitioners with a comprehensive understanding of the objective issues surrounding Linked Data quality.

The framework we propose is based on a refinement of the data quality principles proposed in our previous work [2]. Some attributes have been grouped for more detailed quality assessments. We have also extended the framework by adding for each attribute a set of objective indicators. These indicators are measures that provide users with quality performance. We finally propose when possible existing tools and frameworks that can be used to evaluate and improve each indicator. These tools and frameworks have been identified from an extensive survey that we conducted.

This paper is structured as follows: In Section 2, we present the related work, Section 3 explains the methodology we used to refine our previous framework with the findings of the related work survey and the classification of the new objective framework; Section 4 defines the existing tools and framework in the Linked Open Data quality landscape; Section 5 presents concluding remarks and identifies future work.

2. Related Work

We are entering an era where open is the new default. Governments, universities, organizations and even individuals are publicly publishing huge amounts of open data. This openness should be accompanied with a certain level of trust or guarantees about the quality of data. To our knowledge, only one certificate is available to data publishers to assess the quality level of their datasets, the ODI certificate[2].

This certificate provides a description of the published data quality in plain English. It aspire to act as a mark of approval that helps publishers understand how to publish good open data and users how to use it. It wants to give publishers the ability to provide assurance and support on their data while encouraging further improvements through an ascending scale.

ODI comes as an online and free questionnaire for data publishers focusing on certain characteristics about their data. The questions are classified into the following categories: general information (about dataset, publisher and type of release), legal information (e.g., rights to publish), licensing, privacy (e.g., whether individuals can be identified), practical information (e.g., how to reach the data), quality, reliability, technical information (e.g., format and type of data) and social information (e.g., contacts, communities, etc.). Based on the information provided by the data publisher, a certificate is created with one of four different ratings.

[2] https://certificates.theodi.org/

Although ODI is a great initiative, the issued certificates are self-certified. ODI does not verify or review submissions but retains the right to revoke a certificate at any time. The dynamicity of Linked Data makes it also very difficult to update the certificates manually, especially when these changes are frequent and affect multiple categories. There is clearly a need for automatic certification which can be supplemented with some manual input for categories that cannot be processed by machines.

The emerging critical need for large, distributed, heterogeneous, and complex structured datasets identified the necessity to establish industry cooperation between vendors of RDF and Graph database technologies in developing, endorsing, and publishing reliable and insightful benchmark results. The Linked Data Benchmark Council (LDBC)[3] aims to bridge the gap between the industry and the new trending stack of semantic technologies and their vendors.
LDBC more specifically aims at developing new benchmarks that will lead to significant progress in scalability, storage, indexing and query optimization techniques to become the de facto standard for publishing performance results. LDBC is promising initiative, but it is still work in progress with the final report expected on the first quarter of 2015.

In addition to the initiatives mentioned above, there exist a number of data quality frameworks and tools that are either standalone or implemented as modules in data integration tools.

LODGRefine[4] is the Open Refine[5] of Linked Data. It does not act as a quality assessment tool, but it is powerful in cleaning and refining raw instance data. LODGRefine can help detect duplicates, empty values, spot inconsistencies, extract Named Entities, discover patterns and more. LODGRefine helps in improving the quality of the dataset by improving the quality of the data at the instance level.

PROLOD [9] is also not a quality assessment tool. It is a Linked Data profiling tool that provides clustering and labeling capabilities, schema discovery and statistics about data types and patterns. The statistics are about properties distribution, link-to-literal ratio, number of entities and RDF triples, average properties per entity and average error. PROLOD had been tested with DBpedia but the authors plan to improve its scalability to larger datasets.

Sieve [38] is framework for expressing quality assessment and fusion methods. It is implemented as a component of the Linked Data Integration Framework (LDIF)[6]. Sieve leverages the LDIF provenance metadata as quality indicators to produce quality assessment scores. However, despite its nice features, it is only targeted to perform data fusion based on user-configurable conflict resolution tasks. Moreover, since Sieve main input is provenance metadata, it is only limited to domains that can provide such metadata associated with their data.

Quality Assessment of Data Sources (Flemming's Data Quality Assessment Tool)[7] calculates data quality scores based on manual user input. The user should assign weights to the predefined quality metrics and answer a series of questions regarding the dataset. These include, for example, the use of obsolete classes and properties by defining the number of described entities that are assigned disjoint classes, the usage of stable URIs and whether the publisher provides a mailing list for the dataset. The main disadvantage for using this tool is the manual intervention which requires deep knowledge in the dataset examined. Moreover, the tool lacks support for several quality concerns like completeness or consistency.

[3] http://ldbc.eu/
[4] http://code.zemanta.com/sparkica/
[5] http://openrefine.org/
[6] http://ldif.wbsg.de/
[7] http://linkeddata.informatik.hu-berlin.de/LDSrcAss/datenquelle.php

SWIQA [20] is composed of three layers: data acquisition, query and ontology layers. It uses query templates based on the SPARQL Inferencing Notation (SPIN)[8] to express quality requirements. The queries are built to compute weighted and unweighted quality scores. At the end of the assessment, It uses vocabulary elements to annotate important values of properties and classes, assigning inferred quality scores to ontology elements and classifying the identified data quality problems.

Despite all the recent efforts in providing frameworks and tools for data quality in Linked Open Data, there is still no framework for the objective assessment of such quality taking into account all aspects of Linked Open Data.

3. Objective Linked Data Quality Classification

The basic idea behind Linked Data is that its usefulness increases when it is more interlinked with other datasets. Tim Berners-Lee defined four main principles for publishing data that can ensure a certain level of uniformity reflecting directly data's usability [48]:

- **Make the data available on the Web**: assign URIs to identify things.
- **Make the data machine readable**: use HTTP URIs so that looking up these names is easy.
- **Use publishing standards**: when the lookup is done provide useful information using standards like RDF.
- **Link your data**: include links to other resources to enable users to discover more things.

Building on these principles, we group the quality attributes into four main categories:

- **Quality of the entities** : Quality indicators that focus on the data at the instance level (i.e. syntactic checkers).
- **Quality of the dataset**: Quality indicators at the dataset level.
- **Quality of the semantic model**: Quality indicators that focus on the semantic models, vocabularies and ontologies.
- **Quality of the linking process**: Quality indicators that focus on the inbound and outbound links between datasets.

In our previous work [2] we have identified 24 different Linked Data quality attributes. In this paper, we refine these attributes into a condensed framework of 13 objective attributes. Since these attributes are rather abstract, we should rely on quality indicators that reflect the quality of a data source [17]. In this paper, we transform the quality indicators presented as a set of questions in [2] into more concrete quality indicator metrics. We extend them with the the objective quality indicators listed in the systematic review done in [19].

3.1. Completeness

data completeness can be judged in the presence of a task where the ideal set of attributes and objects are known. An entity is considered to be complete if it contains all the attributes needed for a given task, has complete language coverage [37] and has documentation properties [39][37].
A dataset is considered to be complete if it contains all the necessary objects for a given task [38], contains supporting structured metadata [28],contains links for external data providers, supports providing data

[8] http://spinrdf.org/

in multiple serializations [19], includes the correct MIME-type for the content [28] contains appropriate volume of data for a particular task [19], has different queryable endpoints to access the data (i.e. SPARQL endpoint, RDF Dump, REST API, etc.) [19], has been checked against syntactic errors [28] and if the publishers use datasets description vocabularies like DCAT[9] or VOID[10] to provide descriptions about the size (using void:statItem, void:numberOfTriples or void:numberOfDocuments) and categorization (using dcterms:subject) of the dataset.

Links are considered to be complete if all the in-bound and out-bound links are dereferencable [28][37][22] and have the linkage information represented in the metadata [28].

Models are considered to be complete if they have a complete set of values [37] and do not contain disconnected graph clusters [37]. Disconnected graphs are the result of incomplete data acquisition or accidental deletion of terms that leads to deprecated terms. In addition to that, models are considered to be complete if they do not contain omitted top concepts or unidirectional related concepts [28] and if there exists some metadata about the kind and number of used vocabularies [19].

3.2. Availability

The dataset is considered to be available if the publishers provide an RDF dump that can be downloaded by users [17][28] and if its queryable endpoints respond to direct queries.

3.3. Correctness

Correctness of the data is related to the validity of its entities. An entity is considered to be correct if there no missing or empty labels [1][37], no incorrect data type for typed literals [28][1], no omitted or invalid languages tags [45][37] and does not contain terms without any associative or hierarchical relationships “orphan terms”[36].

Links are considered to be correct if they actually show related content to the subject of the RDF triple [45][1] and are syntactically correct.

Models are considered to be correct if the top concepts are marked and do not have broader concepts (for example having incoming hasTopConcept or outgoing topConceptOf relationships) [37].

3.4. Conciseness

Extensional conciseness measures the number of unique objects in relation to the overall number of objects representation in the data set. Intensional conciseness measures the number of unique attributes of a dataset in relation to the overall number of attributes in a target schema [8].

An entity is considered to be concise if it has intensional conciseness (it does not contain redundant attributes, which means that there is no equivalent attributes with different names) [38] and uses short URIs [19] that follow the HTTP URI scheme [29][46].

A dataset is considered to be concise if it has extensional conciseness (it does not contain redundant objects, which means that there is no equivalent objects with different identifiers) [38].

3.5. Consistency

An entity is considered to be consistent if it does not contain overlapping labels such as two concepts have the same preferred lexical label in a given language when they belong to the same schema [30][37]. Moreover, an entity is considered to be consistent if it does not contain disjoint labels [37], extra white spaces in labels [45] and does not contain inconsistent preferred labels per language tag and no more than one value of skos:prefLabel without a language tag [37][45].

[9] http://www.w3.org/TR/vocab-dcat/

[10] http://www.w3.org/TR/void/

A dataset is considered to be consistent if it is free of conflicting information. This can be measured by considering properties with cardinality 1 that contain more than one distinct value [38].
Models are considered to be consistent if they do not include atypical use of collections, containers and reification [28], overlapping usage of owl:sameAs and owl:differentFrom [28], overlapping usage of owl:AllDifferent and owl:distinctMembers [28], asserted members of owl:Nothing and membership violation for disjoint classes [28].

3.6. Coherence

Coherence is the ability to interpret data as expected by the publisher or vocabulary maintainer [28]. It is mainly associated with the modeling quality.

A model is considered to be coherent when it does not contain:

- Usage of undefined classes and properties [28]. Many errors that are due to spelling or syntactic mistakes are resolvable through minor fixes via ontology checkers tools. However, for new terms, [28] suggests to have them defined in a separate namespace in order to allow reuse [37].
- Usage of blank nodes as they affect merging data from different sources [29].
- Misplaced or deprecated classes or properties [28].
- Misuse of the owl:DataTypeProperty or owl:ObjectProperty [28].
- Relations and mappings clashes [45].
- Invalid inverse-functional values [28].
- Cyclic hierarchical relations [43][45][37].
- Incomplete literals with datatype range [28].
- Solely transitive related concepts [37].
- Redefinitions of existing vocabularies [28].
- Valueless associative relations [37].

3.7. Efficiency

Dataset efficiency is calculated by measuring how fast it can be identified [49]. A dataset is considered to be efficient if it satisfies the following performance metrics:

- No usage of slash-URIs where large amounts of data is provided [19].
- Acceptable delay between the request and its response [5].
- Low Latency HTTP requests (average answer time of one second) [19].
- Scalable such that the time to answer an amount of ten requests divided by ten is not longer than the time it takes to answer one request [19].

3.8. Freshness

Freshness is a measure for the recency of data. The basic assumption is that old information is more likely to be outdated and unreliable [18]. Entity freshness can be identified if it contains timestamps that can keep track of its modifications.

3.9. Accuracy

Accuracy describes the proximity of data value representations of an object related to their real world states [20]. A dataset is considered to be accurate when it does not contain outliers and attributes that do not contain useful values for data entries [19].

3.10. Provenance

Entity level provenance can be calculated by constructing decision networks informed by provenance graphs [21]. The accuracy of computing trust between two entities [19] can be computed by calculating an aggregate trust value based on the combination of the propagation and aggregation algorithms on weighted mechanism [41]. Provenance can be achieved at the dataset level by including metadata that describes its authoritative information (title, content and URI), ensuring the reliability and trustworthiness of the publisher [18], verifying if the dataset uses a provenance vocabulary like PROV [3] and digital signatures [19].
Models provenance can be achieved by ensuring the trustworthiness of RDF statements [25].

3.11. Security

Security is a quality attribute that is measured on the dataset level. It is identified if the publishers use login credentials, SSL or SSH to provide access to their dataset, or if they only grant access to specific users [19].

3.12. Licensing

Licensing is a quality attribute that is measured on the dataset level. It includes the availability of machine readable license information [29], human readable license information in the documentation of the dataset or its source [29] and the indication of permissions, copyrights and attributions specified by the author [19].

3.13. Comprehensibility

Comprehensibility is identified if the publisher indicates at least one exemplary URI and SPARQL query, regular expression pattern that matches the URIs of a dataset [19], provides a list of used vocabularies and an active mailing list or message board for the dataset [17].

Table 1: Objective Linked Data Quality Framework

Quality Attribute	Quality Category	ID	Quality Indicator
Completeness	Entity Level	QI.1	Covers of all the attributes needed for a given task [38]
		QI.2	Has Complete language coverage [37]
		QI.3	Existence of documentation properties [39][37]
	Dataset Level	QI.4	Existence of all the necessary objects for a given task [38]
		QI.5	Existence of supporting structured metadata [28]
		QI.6	Supports multiple serializations [19]
		QI.7	Includes the correct MIME-type for the content [28]
		QI.8	Contains appropriate volume of data for a particular task [19]
		QI.9	Has different queryable endpoints to access the data [19]
		QI.10	Checked against syntactic errors [28]
		QI.11	Usage of datasets description vocabularies
		QI.12	Existence of descriptions about its size and categorization
	Links Level	QI.13	Existence of complete dereferencable in-bound and out-bound links [28][37][22]
		QI.14	Existence of supporting linkage metadata [28]
	Model Level	QI.15	Covers the complete set of values [37]
		QI.16	Absence of disconnected graph clusters [37]
		QI.17	Absence of omitted top concept [28]
		QI.18	Absence of unidirectional related concepts [28]
		QI.19	Existence of supporting metadata about the kind and number of used vocabularies [19]
Availability	Dataset Level	QI.20	Existence of an RDF dump that can be downloaded by users [17][28]
		QI.21	Existence of queryable endpoints that respond to direct queries
Correctness	Entity Level	QI.22	Absence of missing or empty labels [1][37]
		QI.23	Absence of incorrect data type for typed literals [28][1]
		QI.24	Absence of omitted or invalid languages tags [45][37]
		QI.25	Absence of terms without any associative or hierarchical relationships [36]
	Links Level	QI.26	Existence of content related to the subject of the RDF triple [45][1]
		QI.27	Absence of syntactic errors [46]
	Model Level	QI.28	Contains marked top concepts [37]
		QI.29	Absence of broader concepts for top concepts [37]
Conciseness	Entity Level	QI.30	Absence of redundant attributes [38]
		QI.31	Existence of short URIs [19]
	Dataset Level	QI.32	Absence of redundant objects [38]
		QI.33	Follows the HTTP URI scheme [29][46]
Security	Dataset Level	QI.34	Uses login credentials to restrict access [19]
		QI.35	Uses SSL or SSH to provide access to their dataset [19]
		QI.36	Grants access to specific users [19]
Freshness	Entity Level	QI.37	Existence of timestamps that can keep track of its modifications [18]
Licensing	Dataset Level	QI.38	Existence of machine readable license information [29]
		QI.39	Existence of human readable license information [29]
		QI.40	Specifies permissions, copyrights and attributions [19]
			Continued on next page

Table 1 Objective Linked Data Quality Framework

Quality Attribute	Quality Category	ID	Quality Indicator
Comprehensibility	Dataset Level	QI.41	Existence of at least one exemplary URI [19]
		QI.42	Existence of at least one exemplary SPARQL query [19]
		QI.43	Existence of regular expression pattern that matches the URIs of a datasets [19]
		QI.44	Existence a list of used vocabularies
		QI.45	Existence of a mailing list or message board [17]
Consistency	Entity Level	QI.46	Existence of consistent preferred labels per language tag [30][37]
		QI.47	Absence of overlapping labels
		QI.48	Absence of disjoint labels [37]
		QI.49	Absence of extra white spaces in labels [45]
		QI.50	Existence of only one value of skos:prefLabel without a language tag [37][45]
	Dataset Level	QI.51	Absence of conflicting information [38]
	Model Level	QI.52	Absence of atypical use of collections, containers and reification [28]
		QI.53	Absence of overlapping usage of owl:sameAs and owl:differentFrom [28]
		QI.54	Absence of overlapping usage of owl:AllDifferent and owl:distinctMembers [28]
		QI.55	Absence of asserted members of owl:Nothing [28]
		QI.56	Absence of membership violation for disjoint classes [28]
Coherence	Model Level	QI.57	Absence of misplaced or deprecated classes or properties [28]
		QI.58	Absence of misused owl:DataTypeProperty or owl:ObjectProperty [28]
		QI.59	Absence of relations and mappings clashes [45]
		QI.60	Absence or minimal usage of blank nodes [29]
		QI.61	Absence of invalid inverse-functional values [28]
		QI.62	Absence of cyclic hierarchical relations [43][45][37]
		QI.63	Absence of undefined classes and properties usage [28]
		QI.64	Absence of solely transitive related concepts [37]
		QI.65	Absence of redefinitions of existing vocabularies [28]
		QI.66	Absence of valueless associative relations [37]
		QI.67	Absence of incomplete literals with datatype range [28]
Efficiency	Dataset Level	QI.68	Absence of slash-URIs [19]
		QI.69	Provides acceptable delay between the request and its response [5]
		QI.70	Serves low Latency HTTP requests [19]
		QI.71	Has the ability to scale [19]
Accuracy	Dataset Level	QI.72	Absence of outliers [19]
		QI.73	Absence of attributes that do not contain useful values for data entries [19]
Provenance	Entity Level	QI.74	Able to construct decision networks informed by provenance graphs [21]
	Dataset Level	QI.75	Existence of metadata that describes its authoritative information [18]
		QI.76	Ensures the reliability and trustworthiness of the publisher [18]
		QI.77	Uses a provenance vocabulary
		QI.78	Uses digital signatures [19]
	Model Level	QI.79	Ensures the trustworthiness of RDF statements [25]

4. Linked Data Quality Tools

The literature contains several tools that reflect the different aspects of LOD: modeling, ontologies and vocabularies, dataset and SPARQL end-points. In this section we present the results of our survey on these tools.

4.1. Information Modeling Quality

RDF is the standard to model information in the Semantic Web. Linked Data publishers can pick from a plethora of tools that can check their RDF files for quality problems[11]. Syntactic RDF checkers are able to detect errors in RDF documents like the W3C RDF Validator[12], RDF:about validator and Converter[13] and The Validating RDF Parser (VRP)[14]. The RDF Triple-Checker[15] is an online tool that helps find typos and common errors in RDF data. Vapour[16] [4] is a validation service to check whether semantic Web data is correctly published according to the current best practices[48].

4.2. Ontologies and Vocabularies Quality

Reusing existing ontologies is a common practice that Linked Data publishers are always trying to adopt. However, ontologies and vocabularies development is often a long error-prone process especially when many contributors are working consecutively or collaboratively [46]. This can introduce deficiencies such as redundant concepts or conflicting relationships [23]. Getting to choose the right ontology or vocabulary is vital to ensure modeling correctness and consistency.

qSKOS[17] [37] scans SKOS vocabularies to provide reports on vocabulary resources and relations that are problematic. Skosify [45] supports OWL and RDFS ontologies by converting them into well-structured SKOS vocabularies. Skosify includes automatic correction abilities for quality issues that have been observed by reviewing vocabularies on the Web. The OOPS! pitfall scanner [40] evaluates OWL ontologies against a rules catalog and provides the user with a set of guidelines to solve them. PoolParty checker[18] highlights quality issues in OWL, RDFS and SKOS ontologies and vocabularies, the latest version supports qSKOS to indicate the quality of controlled vocabularies on the Web. ASKOSI[19] retrieves vocabularies from different sources, stores and displays the usage frequency of the different concepts used by different applications. It promotes reusing existing information systems by providing better management and presentation tools.

Some errors in RDF will only appear after reasoning (incorrect inferences). In [42][47] the authors perform quality checking on OWL ontologies using integrity constraints involving the Unique Name Assumption (UNA) and the Closed World Assumption (CWA). Pellet[20] provides reasoning services for OWL ontologies. It incorporates a number of heuristics to detect and repair quality issues among disjoint properties, negative property assertions and reflexive, irreflexive, symmetric, and anti-symmetric properties. Eyeball[21] provides quality inspection for RDF models (including OWL). It provides checks for a variety of problems including the usage of unknown predicates, classes, poorly formed namespaces, literal syntax validation, type consistency and other heuristics. RDF:Alerts[22] provides validation for many issues highlighted in [28] like misplaced, undefined or deprecated classes or properties.

[11] http://www.w3.org/2001/sw/wiki/SWValidators

[12] http://www.w3.org/RDF/Validator/

[13] http://rdfabout.com/demo/validator/

[14] http://139.91.183.30:9090/RDF/VRP/index.html

[15] http://graphite.ecs.soton.ac.uk/checker/

[16] http://validator.linkeddata.org/vapour

[17] https://github.com/cmader/qSKOS

[18] http://www.poolparty.biz/

[19] http://www.w3.org/2001/sw/wiki/ASKOSI

[20] http://clarkparsia.com/pellet

[21] http://jena.sourceforge.net/Eyeball/

[22] http://swse.deri.org/RDFAlerts/

4.3. Dataset Quality

Considering the large amount of available datasets in the Linked Open Data, users have a hard time trying to identify appropriate datasets that suit certain tasks. There are two approaches to rank datasets, a manual and an automatic approach. The manual approach depends on the wisdom of the crowd to highlight specific quality issues. The automatic approach has several implementations. The most adopted ones are based on link assessment. Provenance-based approaches and entity-based approaches are also used to compute not only dataset rankings, but also rankings on the entity level.

4.3.1. Manual Ranking Tools

There are several quality issues that can be difficult to spot and fix automatically. In [1] the authors highlight the fact that the RDFification process of some data can be more challenging than others, leading to errors in the Linked Data provisioning process that needs manual intervention. This can be more visible in datasets that have been semi-automatically translated to RDF from their primary source (the best example for this case is DBpedia [7]). The authors introduce a methodology to adjust crowdsourcing input from two types of audience: 1) Linked Data experts, researchers and enthusiasts through a contest to find and classify erroneous RDF triples and 2) Crowd-sourcing through the Amazon Mechanical Turk[23].

TripleCheckMate [34] is the tool used by the authors to run out their assessment. The tool allows users to select resources, identify and classify possible issues according to a pre-defined taxonomy of quality problems. It measures inter-rater agreements, meaning that the resources defined are checked multiple times. This features turn out to be extremely useful to analyze the performance of users and allows better identification of potential quality problems. TripleCheckMate is used to identify accuracy issues in the object extraction (completeness of the extraction value for object values and datatypes), relevancy of the extracted information, representational consistency and the interlinking with other datasets.

4.3.2. Automatic Ranking Tools

Links Based Approach

The basic idea behind link assessment tools is to provide rankings for datasets based on the cardinality and types of the relationships with other datasets. Traditional link analysis has proven to be an effective way to measure the quality of Web documents search. Algorithms like PageRank [35] and HITS [33] became successful based on the assumption that a certain Web document is considered to have higher importance or rank if it has more incoming links that other Web documents [11][13].

However, the basic assumption that links are equivalent does not suit the heterogeneous nature of links in the Linked Open Data. Thus, the previous approaches fall short to provide reliable rankings as the types of the links can have a direct impact on the ranking computation [49].

The first adaption of PageRank for Semantic Web resources was the Ontology Rank algorithm implemented in the Swoogle search engine [16]. They use a rational random surfing model that takes into account the different types of links between discovered sets and compute rankings based on three levels of granularity: documents, terms and RDF graphs. ReConRank [27] rankings are computed at query time based on two levels of granularity: resources and context graphs. DING [49] adapted the PageRank to rank datasets based on their interconnections. DING can also automatically assign weights to different link types based on the nature of the predicate involved in the link.

Broken links are a major threat to Linked Data. They occur when resources are removed, moved or updated. DSNotify[24][26] is a framework that informs data consumers about the various types of events that occur on data sources. Their approach is based on an indexing infrastructure that extracts feature vectors and stores them to an index. A monitoring module detects events on sources and write them to a central

[23] https://www.mturk.com/

[24] http://www.cibiv.at/ niko/dsnotify/

event log which pushes notifications to registered applications.

Provenance-based Approach

Provenance based assessment methods are an important step towards transparency of data quality in the Semantic Web. In [25] the authors use a provenance model as an assessment method to evaluate the timeliness of Web data. Their model identifies types of "provenance elements" and the relationships between them. Provenance elements are classified into three types: actors, executions and artifacts. The assessment procedure is divided into three steps: 1) Creating provenance graph based on the defined model 2) Annotating the graph with impact values 3) Calculating the information quality score. In [18] the authors describe a set of provenance-based assessment metrics to support quality assessment and repair in Linked Open Data. They rely on both data and metadata and use indicators like the source reputation, freshness and plausibility. In [24] the authors introduce the notion of naming authority which connects an identifier with the source to establish a connection to it's provenance. They construct a naming authority graph that acts as input to derive PageRank scores for the data sources.

Entity-based Approach

Sindice [14] uses a set of techniques to rank Web data. They use a combination of query dependent and query independent rankings implemented in the Semantic Information Retrieval Engine (SIREn)[25] to produce a final entity rank. Their query dependent approach rates individual entities by aggregating the the score of the matching terms with a term frequency - inverse subject frequency (tf-isf) algorithm. Their query independent ranking is done using hierarchical links analysis algorithms [15]. The combination of these two approaches is used to generate a global weighted rank based on the dataset, entities and links ranks.

4.4. Queryable End-point Quality

The availability of Linked Data is highly dependent on the performance qualities of its queryable endpoints. The standard query language for Semantic Web resources is SPARQL, thus SPARQL endpoints are the main focus. In [12] the authors present their findings to measure the discoverability of SPARQL endpoints by analyzing how they are located and the metadata used to describe them. In addition to that, they also analyze endpoints interoperability by identifying features of SPARQL 1.0 and SPARQL 1.1 that are supported. The authors tackled the endpoints efficiency by testing the time taken to answer generic, content-agnostic SPARQL queries over HTTP. Finally, the authors measured endpoints reliability by monitoring the uptime of public SPARQL endpoints on a course of 27 months. The results for this work can be accessed online via the SPARQL Endpoints Status tool [26] and is queryable using their public SPARQL endpoint[27].

Table(2) shows the result for our evaluation of the various Linked Data tools mentioned in this paper. Due to space limitation, we have only shown the tools that can directly or indirectly assess one or more of the quality indicators listed in Table(1). Moreover, we have also grouped the Syntax Validation tools (W3C RDF Validator, RDF:about, VRP, The RDF Triple-Checker and Vapour) under one approach. We should highlight that the PoolParty tool includes qSKOS in its implementation, thus they are grouped in one tool which is the PoolParty.

[25] http://siren.sindice.com/

[26] http://labs.mondeca.com/sparqlEndpointsStatus/

[27] http://labs.mondeca.com/sparqlEndpointsStatus/endpoint/endpoint.html

Table 2: Linked Data quality tools evaluation

ID / Tool	PROLOD	Sieve	Flemming's Tool	SWIQA	Syntax Validators	Skosify	OOPS!	PoolParty	ASKOSI	Pellet	Eyeball	RDF:Alerts	TripleCheckmate	DING	DSNotify	SPARQL Endpoints-Status
QI.1		✓		✓									✓			
QI.2			✓			✓		✓	✓							
QI.3			✓					✓								
QI.4		✓		✓									✓			
QI.5			✓													
QI.6			✓											✓		
QI.7			✓													
QI.8	✓		✓													
QI.9			✓													
QI.10			✓		✓							✓				
QI.11	✓					✓	✓	✓	✓		✓			✓		✓
QI.12	✓		✓											✓		
QI.13			✓					✓				✓			✓	
QI.14															✓	✓
QI.15		✓		✓												
QI.16								✓								
QI.17																
QI.18						✓		✓								
QI.19			✓				✓									
QI.20			✓									✓				
QI.21			✓									✓				✓
QI.22						✓		✓	✓				✓			
QI.23							✓			✓	✓		✓			
QI.24																
QI.25								✓	✓							
QI.26																
QI.27			✓								✓	✓			✓	
QI.28						✓		✓								
QI.29			✓			✓		✓	✓		✓	✓				
QI.30				✓			✓	✓	✓							
QI.31			✓													
QI.32		✓	✓	✓												
QI.33			✓									✓				
QI.34																
QI.35																
QI.36																
QI.37		✓		✓												
QI.38			✓													
QI.39			✓													
QI.40			✓													
QI.41			✓													
QI.42			✓													
QI.43			✓													

Continued on next page

Table 2 Linked Data quality tools evaluation

ID / Tool	PROLOD	Sieve	Flemming's Tool	SWIQA	Syntax Validators	Skosify	OOPS!	PoolParty	ASKOSI	Pellet	Eyeball	RDF:Alerts	TripleCheckmate	DING	DSNotify	SPARQL Endpoints-Status
QI.44			✓													
QI.45			✓													
QI.46						✓		✓	✓							
QI.47							✓	✓			✓					
QI.48						✓	✓	✓			✓					
QI.49						✓					✓					
QI.50								✓	✓							
QI.51										✓	✓		✓			
QI.52											✓					
QI.53							✓			✓						
QI.54										✓						
QI.55										✓						
QI.56						✓	✓	✓		✓	✓					
QI.57				✓	✓		✓	✓			✓					
QI.58							✓	✓		✓	✓					
QI.59								✓				✓				
QI.60																
QI.61							✓			✓						
QI.62						✓	✓	✓	✓							
QI.63							✓	✓		✓	✓					
QI.64						✓		✓								
QI.65			✓							✓	✓	✓				
QI.66								✓								
QI.67				✓			✓	✓		✓	✓					
QI.68																
QI.69																✓
QI.70																✓
QI.71																✓
QI.72	✓			✓									✓			
QI.73				✓									✓			
QI.74																
QI.75			✓													
QI.76																
QI.77	✓		✓													
QI.78			✓													
QI.79																

5. Conclusions and Future Work

In this paper, we have presented a comprehensive objective quality framework applied to the Linked Open Data. We have refined our previous work and presented 13 different quality attributes. To measure these abstract attributes, we have identified a total of 79 quality indicators that were mapped when suitable to four main categories (entity, dataset, links, models). We have also surveyed more than 25 different tools that measure different quality aspects of Linked Open Data. We evaluated these tools against our quality indicators. As a result, we have identified the need for a complete quality framework that can assess all the quality indicators. Most of the tools were designed with limited coverage to certain aspects, for example, ontology and vocabulary checkers focus mainly on the coherence, completeness and correctness at the modeling level. Flemming's tool covers several attributes like the completeness, correctness, conciseness, security, licensing and comprehensibility, but it falls short in measuring the consistency, coherence, efficiency and provenance.

We have also noticed the lack of tools to measure certain quality indicators like the dataset's security. Moreover, to our knowledge, there are no tools that can measure all the provenance quality indicators (except for Flemming's tool that is able to check for the use of digital signatures) although the literature covers several approaches to achieve that [25][18][24].

We plan to develop a comprehensive objective Linked Data quality evaluation tool. The tool will be able to automatically measure the various quality indicators listed in this paper, introduce a scoring function with different weights for the various quality attributes and issue a quality certificate.

References

[1] M. Acosta, A. Zaveri, E. Simperl, and D. Kontokostas. Crowdsourcing Linked Data quality assessment. *ISWC 2013*, 2013.

[2] A. Assaf and A. Senart. Data quality principles in the semantic web. *CoRR*, abs/1305.4054, 2013.

[3] K. Belhajjame, J. Cheney, D. Corsar, D. Garijo, S. Soiland-Reyes, S. Zednik, and J. Zhao. Prov-o: The prov ontology. Technical report, 2012.

[4] D. Berrueta, S. Fernndez, and I. Frade. Cooking http content negotiation with vapour. In *In Proceedings of 4th workshop on Scripting for the Semantic Web 2008 (SFSW2008). co-located with ESWC2008*, 2008.

[5] C. Bizer. *Quality-Driven Information Filtering in the Context of Web-Based Information Systems*. PhD thesis, Mar. 2007.

[6] C. Bizer, A. Jentzsch, and R. Cyganiak. State of the lod cloud, 2011.

[7] C. Bizer, J. Lehmann, G. Kobilarov, S. Auer, C. Becker, R. Cyganiak, and S. Hellmann. DBpedia - a crystallization point for the web of data. *Web Semantics: Science, Services and Agents on the World Wide Web*, 7(3):154–165, Sept. 2009.

[8] J. Bleiholder and F. Naumann. Data fusion. *ACM Comput. Surv.*, 41(1):1:1–1:41, Jan. 2009.

[9] C. Bohm, F. Naumann, Z. Abedjan, F. Dandy, T. Grutze, D. Hefenbrock, M. Pohl, and D. Sonnabend. Proling Linked Open Data with ProLOD.PDF. *ICDE 2010*, 2010.

[10] D. Boyd and K. Crawford. Six provocations for big data. *A Decade in Internet Time: Symposium on the Dynamics of the Internet and Society*, pages 1–17, 2011.

[11] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. In *Proceedings of the seventh international conference on World Wide Web 7*, WWW7, pages 107–117, Amsterdam, The Netherlands, The Netherlands, 1998. Elsevier Science Publishers B. V.

[12] C. Buil-Aranda and A. Hogan. SPARQL Web-Querying Infrastructure: Ready for Action? *International ...*, 2013.

[13] S. Chakrabarti, B. E. Dom, S. R. Kumar, P. Raghavan, S. Rajagopalan, A. Tomkins, D. Gibson, and J. Kleinberg. Mining the web's link structure, 1999.

[14] R. Delbru. Sindice at SemSearch 2010. *WWW10*, 2010.

[15] R. Delbru, N. Toupikov, and M. Catasta. Hierarchical link analysis for ranking web data. *The Semantic Web: Research and Applications*, 2010.

[16] L. Ding, T. Finin, A. Joshi, R. Pan, and R. Cost. Swoogle: A semantic web search and metadata engine. *CIKM04*, 2004.

[17] A. Flemming. Quality characteristics of linked data publishing datasources, 2010.

[18] G. Flouris, Y. Roussakis, and M. Poveda-Villalón. Using provenance for quality assessment and repair in linked open data. pages 1–12, 2012.

[19] C. Framework, A. Zaveri, A. Rula, A. Maurino, R. Pietrobon, and J. Lehmann. Quality Assessment Methodologies for Linked Open Data. *Under review, Semantic Web Journal*, 1:1–5, 2012.

[20] C. Fürber and M. Hepp. SWIQAA Semantic Web information quality assessment framework. *ECIS 2011*, 2011.

[21] M. Gamble. Quality, Trust, and Utility of Scientic Data on the Web: Towards a Joint Model.pdf. *WebSci'11*, 2011.

[22] C. Guéret, P. T. Groth, C. Stadler, and J. Lehmann. Assessing linked data mappings using network measures. In *Proceedings of the 9th Extended Semantic Web Conference*, volume 7295 of *Lecture Notes in Computer Science*, pages 87–102. Springer, 2012.

[23] P. Harpring. *Introduction to Controlled Vocabularies: Terminology for Art, Architecture, and Other Cultural Works.* Getty Research Institute, Los Angeles, 2010.

[24] A. Harth, S. Kinsella, and S. Decker. Using naming authority to rank data and ontologies for web search. *ISWC 2009*, 2, 2009.

[25] O. Hartig and J. Zhao. Using web data provenance for quality assessment. In *In: Proc. of the Workshop on Semantic Web and Provenance Management at ISWC*, 2009.

[26] B. Haslhofer and N. Popitsch. Dsnotify: Detecting and fixing broken links in linked data sets. In *8th International Workshop on Web Semantics (WebS &#8217;09), co-located with DEXA 2009*, Berlin, Heidelberg, August 2009. Springer.

[27] A. Hogan, A. Harth, and S. Decker. Reconrank: A scalable ranking method for semantic web data with context. In *In 2nd Workshop on Scalable Semantic Web Knowledge Base Systems*, 2006.

[28] A. Hogan, A. Harth, A. Passant, S. Decker, and A. Polleres. Weaving the pedantic web. *LDOW 2010*, 2010.

[29] A. Hogan, J. Umbrich, A. Harth, R. Cyganiak, A. Polleres, and S. Decker. An empirical survey of linked data conformance. *Web Semant.*, 14:14–44, July 2012.

[30] A. Isaac and E. Summers. Skos simple knowledge organization system primer. World Wide Web Consortium, Working Draft WD-skos-primer-20080829, August 2008.

[31] J. M. Juran and A. B. Godfrey. *Juran's quality handbook.* Juran's quality handbook, 5e. McGraw Hill, 1999.

[32] B. K. Kahn, D. M. Strong, and R. Y. Wang. Information quality benchmarks: product and service performance. *Communications of the ACM*, 45(4ve):184–192, Apr. 2002.

[33] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. *J. ACM*, 46(5):604–632, Sept. 1999.

[34] D. Kontokostas, A. Zaveri, S. Auer, and J. Lehmann. TripleCheckMate: A Tool for Crowdsourcing the Quality Assessment of Linked Data. *4th Conference on Knowledge Engineering and Semantic Web*, pages 1–8, 2013.

[35] P. Lawrence, B. Sergey, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford University, 1998.

[36] H. Living. Review of: Hedden, heather. the accidental taxonomist medford, nj: Information today, inc., 2010. *Inf. Res.*, 15(2), 2010.

[37] C. Mader, B. Haslhofer, and A. Isaac. Finding quality issues in SKOS vocabularies. *Theory and Practice of Digital Libraries*, 2012.

[38] P. Mendes, H. Mühleisen, and C. Bizer. Sieve: linked data quality assessment and fusion. *LWDM2012 - Proceedings of the 2012 Joint EDBT*, 2012.

[39] A. Miles and S. Bechhofer. Skos simple knowledge organization system reference. w3c recommendation 18 august 2009., 2009.

[40] M. Poveda-Villaln, M. Surez-Figueroa, and A. Gmez-Prez. Validating ontologies with oops! In A. Teije, J. Vlker, S. Handschuh, H. Stuckenschmidt, M. dAcquin, A. Nikolov, N. Aussenac-Gilles, and N. Hernandez, editors, *Knowledge Engineering and Knowledge Management*, volume 7603 of *Lecture Notes in Computer Science*, pages 267–281. Springer Berlin Heidelberg, 2012.

[41] S. Shekarpour and S. Katebi. Modeling and evaluation of trust with an extension in semantic web. *Web Semantics: Science, Services and Agents on the World Wide Web*, 8(1), 2010.

[42] E. Sirin, M. Smith, and E. Wallace. Opening, closing worlds - on integrity constraints. In C. Dolbear, A. Ruttenberg, and U. Sattler, editors, *OWLED*, volume 432 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2008.

[43] D. Soergel. Thesauri and ontologies in digital libraries. In M. Marlino, T. Sumner, and F. M. S. III, editors, *JCDL*, page 421. ACM, 2005.

[44] B. Stvilia, L. Gasser, M. B. Twidale, and L. C. Smith. A framework for information quality assessment. *Journal of the American Society for Information Science and Technology*, 58(12):1720–1733, Oct. 2007.

[45] O. Suominen and E. Hyvönen. Improving the quality of skos vocabularies with skosify. In *Proceedings of the 18th international conference on Knowledge Engineering and Knowledge Management*, EKAW'12, pages 383–397, Berlin, Heidelberg, 2012. Springer-Verlag.

[46] O. Suominen and C. Mader. Assessing and Improving the Quality of SKOS Vocabularies. *Journal on Data Semantics*, June 2013.

[47] J. Tao, L. Ding, and D. L. McGuinness. Instance data evaluation for semantic web-based knowledge management systems. In *HICSS*, pages 1–10. IEEE Computer Society, 2009.

[48] B.-L. Tim. Linked data. Technical report, W3C, July 2006. http://www.w3.org/DesignIssues/LinkedData.html.

[49] N. Toupikov, J. Umbrich, and R. Delbru. DING! Dataset ranking using formal descriptions. *WWW09*, 2009.

[50] R. Wang and D. Strong. Beyond accuracy: What data quality means to data consumers. *Journal of management information systems*, 1996.