Towards An Objective Assessment Framework for Linked Data Quality

A Tool for Automatic Detection of Linked Data Quality Problems

Ahmad Assaf [a,b], Aline Senart [a] and Raphaël 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: first.last@eurecom.fr

Abstract. The standardization of Semantic Web technologies and specifications has resulted in a staggering volume of data being published. The Linked Open Data (LOD) is a gold mine for organizations trying to leverage external data sources in order to produce more informed business decisions. 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. Ensuring data quality in Linked Open Data is much more complex. It consists of structured information supported by models, ontologies and vocabularies and contains queryable endpoints and links. In this paper, we first propose an objective assessment framework for Linked Data quality. In a previous work, we identified potential quality issues of Linked Data and listed quality principles for all stages of data management. In this paper, we build upon this work but focus only on the objective quality indicators based on metrics that can be automatically measured. Secondly, we surveyed the landscape of Linked Data quality tools to discover that they only cover a subset of the proposed quality indicators. As a result, we present an extensible quality measurement tool that 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. We evaluate this tool by measuring the quality of the LOD cloud. The results demonstrate that the general quality of LOD cloud compared to various data portals needs more attention as most of the datasets suffer from various quality issues.

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

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 cloud has grown to nearly 1000 datasets containing more than 82 billion triples. Data is being

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

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 where users are empowered to find, share and combine information in their applications easily.

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. The Linked Open Data is a gold mine for those trying to leverage external data sources in order to produce more informed business decisions [9]. However, the heterogeneous nature of 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 [34][54][61]. Data quality principles typically rely on many subjective indicators that are complex to measure automatically. The quality of data in indeed realized when it is used [33], 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 subjectively calculated via algorithms like Page Rank [39]. Ensuring data quality in Linked Open Data is much more complex. It consists of structured information supported by models, ontologies and vocabularies and contains queryable 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.

Data quality assessment is the process of evaluating if a piece of data meets the consumers need in a specific use case [6]. The dimensionality of data quality makes it dependent on the task and users requirements. For example, DBpedia [7] is a knowledge base containing data extracted from structured and semi-structured sources. It is used in a variety of applications e.g. annotation systems [45], exploratory search [43] and recommendation engines [47]. However, DBpedia's data isn't integrated into critical systems e.g. life critical (medical applications) or safety critical (aviation applications) as its data quality is found to be insufficient. In this paper, we first propose a comprehensive objective framework to evaluate the quality of Linked Data sources. Secondly, we present an extensible quality measurement tool that 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 [3] and those surveyed in [22]. Some attributes have been grouped for more detailed quality assessments while we have also extended them by adding for each attribute a set of objective indicators. These indicators are measures that provide users with quality metrics measurable by tools regardless of the use case. For example, when measuring the quality of DBpedia dataset, an objective metric would be the availability of human or machine readable license information rather than the trustworthiness of the publishers.

Furthermore, we surveyed the landscape of Linked Data quality tools to discover that they only cover a subset of the proposed objective quality indicators. As a result, we present an extensible quality measurement tool and evaluate it by measuring the quality of the LOD cloud datasets. The results demonstrate that the general quality of LOD cloud needs more attention as most of the datasets suffer from various quality issues.

This paper is structured as follows: Section 2 presents the related work, Section 3 presents the framework (objective quality measures and indicators); Section 4 reviews the existing tools and framework in the Linked Open Data quality landscape; Section 5 presents our tool for evaluating objective Linked Data

quality indicators; Section 6 presents our evaluation and experiments running this tool on LOD; Section 7 presents concluding remarks and identifies future work.

2. Related Work

In [22], the authors present a comprehensive systematic review of data quality assessment methodologies applied to LOD. They have extracted 26 quality dimensions and a total of 110 objective and subjective quality indicators. However, some of those objective indicators are dependent on the use case thus there is no clear separation on what can be automatically measured. For example, data completeness is generally a subjective dimension. However, the authors specified that the detection of the degree on which all the real-world objects are represented, detection of number of missing values for specific property and detection of the degree to which instances in the dataset are interlinked are considered as objective indicators given the presence of a gold standard or the original data source to compare with. Moreover, lots of the defined performance dimensions like low latency, high throughput or scalability of a data source were defined as objective but are still dependent on multiple subjective factors like network congestion. In addition, there were some missing objective indicators vital to the quality of LOD e.g. indication of the openness of the dataset.

The ODI certificate[2] provides a description of the published data quality in plain English. It aspires to act as a mark of approval that helps publishers understand how to publish good open data and users how to use it. It gives 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.

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 a promising initiative, but it is still work in progress with the final report expected on the first quarter of 2015.

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

[3] http://ldbc.eu/

The Data Hub LOD Validator[4] gives an overview of Linked Data sources cataloged on the Data Hub. It offers a step-by-step validator guidance to check a dataset completeness level for inclusion in the LOD cloud. The results are divided into four different compliance levels from basic to reviewed and included in the LOD cloud. Although it is an excellent tool to monitor LOD compliance, it still lacks the ability to give detailed insights about the completeness of the metadata and overview on the state of the whole LOD cloud group and is very specific to the LOD cloud group rules and regulations.

In [51], the authors propose a methodology for assessing Linked Data quality. It consists of three main steps: (1) requirement analysis, (2) quality assessment and (3) quality improvement. Considering the multidimensionality of data quality, the methodology requires users to provide the details of a use case or a scenario that describes the intended usage of the data. Moreover, quality issues identification is done with the help of a checklist. The user must have prior knowledge about the details of the data in order to fill this list.
Tools implementing the proposed methodology should be able to generate comprehensive quality measures. However, they will require heavy manual intervention and deep knowledge on the data examined. These issues highly affect detecting quality issue on large scale.

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

3. Objective Linked Data Quality Classification

In this section, we present the objective Linked Data quality framework based on the refinement of our previous work [3] and those surveyed in [22].

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 [58]:

- **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.
- **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 [3] we have identified 24 different Linked Data quality attributes. In this paper, we refine these attributes into a condensed framework of 9 objective measures. Since these measures are rather

[4] http://validator.lod-cloud.net/

abstract, we should rely on quality indicators that reflect data quality [20]. In this paper, we transform the quality indicators presented as a set of questions in [3] into more concrete quality indicator metrics. Our proposed tool supports data quality assessment for datasets hosted on CKAN[5] powered data portals. As a result, we support the suggestion of quality indicators with examples based on the CKAN standard dataset model[6].
Independent indicators for entity quality are mainly subjective e.g. the degree to which all the real-world objects are represented, the scope and level of details, etc. However, since entities are governed by the underlying model, we have grouped their indicators with those of the modeling quality. Table 1 lists the refined measures alongside their quality indicators. These attributes are presented in the following sections.

Table 1: Objective Linked Data Quality Framework

Quality Attribute	Quality Category	ID	Quality Indicator
Completeness	Dataset Level	QI.1	Existence of supporting structured metadata [30]
		QI.2	Supports multiple serializations [22]
		QI.3	Has different data access points
		QI.4	Uses datasets description vocabularies
		QI.5	Existence of descriptions about its size categorization
		QI.6	Existence of descriptions about its structure (MIME Type, Format)
		QI.7	Existence of descriptions about its organization and categorization
		QI.8	Existence of information about the kind and number of used vocabularies [22]
	Links Level	QI.9	Existence of dereferencable links for the dataset [30][42][24]
	Model Level	QI.10	Absence of disconnected graph clusters [42]
		QI.11	Absence of omitted top concept [30]
		QI.12	Has complete language coverage [42]
		QI.13	Absence of unidirectional related concepts [30]
		QI.14	Absence of missing labels [42]
		QI.15	Absence of missing equivalent properties [35]
		QI.16	Absence of missing inverse relationships [35]
		QI.17	Absence of missing domain or range values in properties [35]
Availability	Dataset Level	QI.18	Existence of an RDF dump that can be downloaded by users [20][30]
		QI.19	Existence of a queryable endpoint that responds to direct queries
		QI.20	Existence of valid dereferencable URLs (respond to HTTP request)
Licensing	Dataset Level	QI.21	Existence of human and machine readable license information [31]
		QI.22	Existence of de-referenceable links to the full license information [31]
		QI.23	Specifies permissions, copyrights and attributions [22]
Freshness	Dataset Level	QI.24	Existence of timestamps that can keep track of its modifications [21]
Correctness	Dataset Level	QI.25	Includes the correct MIME-type for the content [30]
		QI.26	Includes the correct size for the content
		QI.27	Absence of syntactic errors on the instance level [30]
	Links Level	QI.28	Absence of syntactic errors [56]
		QI.29	Use the HTTP URI scheme (avoid using URNs or DOIs) [42]
	Model Level	QI.30	Contains marked top concepts [42]
		QI.31	Absence of broader concepts for top concepts [42]
		QI.32	Absence of missing or empty labels [2][42]
		QI.33	Absence of unprintable characters [2][42] or extra white spaces in labels [55]
		QI.34	Absence of incorrect data type for typed literals [30][2]
		QI.35	Absence of omitted or invalid languages tags [55][42]
		QI.36	Absence of terms without any associative or hierarchical relationships [41]

Continued on next page

[5]http://ckan.org
[6]http://demo.ckan.org/api/3/action/package_show?id=adur_district_spending

Table 1 Objective Linked Data Quality Framework

Quality Attribute	Quality Category	ID	Quality Indicator
Comprehensibility	Dataset Level	QI.37	Existence of at least one exemplary URI [22]
		QI.38	Existence of at least one exemplary SPARQL query [22]
		QI.39	Existence of general information (title, URL, description) for the dataset
		QI.40	Existence of a mailing list, message board or point of contact [20]
	Model Level	QI.41	Absence of misuse of ontology annotations [42][35]
		QI.42	Existence of annotations for concepts [35]
		QI.43	Existence of documentation for concepts [42][35]
Provenance	Dataset Level	QI.44	Existence of metadata that describes its authoritative information [21]
		QI.45	Usage of a provenance vocabulary
		QI.46	Usage of a provenance timestamps
Coherence	Model Level	QI.47	Absence of misplaced or deprecated classes or properties [30]
		QI.48	Absence of relation and mappings clashes [55]
		QI.49	Absence of blank nodes [31]
		QI.50	Absence of invalid inverse-functional values [30]
		QI.51	Absence of cyclic hierarchical relations [53][55][42]
		QI.52	Absence of undefined classes and properties usage [30]
		QI.53	Absence of solely transitive related concepts [42]
		QI.54	Absence of redefinitions of existing vocabularies [30]
		QI.55	Absence of valueless associative relations [42]
Consistency	Model Level	QI.56	Consistent usage of preferred labels per language tag [32][42]
		QI.57	Consistent usage of naming criteria for concepts [35]
		QI.58	Absence of overlapping labels
		QI.59	Absence of disjoint labels [42]
		QI.60	Absence of atypical use of collections, containers and reification [30]
		QI.61	Absence of wrong equivalent, symmetric or transitive relationships [35]
		QI.62	Absence of membership violations for disjoint classes [30]
Security	Dataset Level	QI.63	Uses login credentials to restrict access [22]
		QI.64	Uses SSL or SSH to provide access to their dataset [22]

3.1. Completeness

Data completeness can be judged in the presence of a task where the ideal set of attributes and objects are known. It is generally a subjective measure depending highly on the scenario and use-case in hand. For example, an entity is considered to be complete if it contains all the attributes needed for a given task, has complete language coverage [42] and has documentation properties [46][42].
Dataset completeness has some objective measures which we include in our framework. A dataset is considered to be complete if it contains supporting structured metadata [30], provides data in multiple serializations [22], has different queryable endpoints to access the data (i.e. SPARQL endpoint, REST API, etc.) [22], uses datasets description vocabularies like DCAT[7] or VOID[8], and if the publishers provide descriptions about the size (e.g. `void:statItem`, `void:numberOfTriples` or `void:numberOfDocuments`) and categorization (e.g. `dcterms:subject`) of the dataset and if there exists metadata information about the kind and number of used vocabularies [22].
Links are considered to be complete if all the inbound and outbound links are de-referenceable [30][42][24]. Models are considered to be complete if they do not contain disconnected graph clusters [42]. 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 have complete language coverage

[7] http://www.w3.org/TR/vocab-dcat/
[8] http://www.w3.org/TR/void/

(each concept labeled in each of the languages that are also used on the other concepts) [42], do not contain omitted top concepts or unidirectional related concepts [30] and if they are not missing labels [42], equivalent properties, inverse relationships, domain or range values in properties [35].

3.2. Availability

A dataset is considered to be available if the publishers provide data dumps e.g. RDF dump, that can be downloaded by users [20][30] and if its queryable endpoints e.g. SPARQL endpoint, are reachable and respond to direct queries.

3.3. Correctness

A dataset is considered to be correct if it includes the correct MIME-type for the content [30] and doesn't contain syntactic errors [30].
Links are considered to be correct if they lack syntactic errors and use the HTTP URI scheme (avoid using URNs or DOIs) [42].
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) [42]. Moreover, if they don't contain incorrect data type for typed literals [30][2], no omitted or invalid languages tags [55][42], does not contain "orphan terms" (orphan terms are terms without any associative or hierarchical relationships [41]) and if the labels are not empty, do not contain unprintable characters [2][42] or extra white spaces [55].

3.4. Consistency

Consistency implies lack of contradictions and conflicts. The objective indicators are mainly associated with the modeling quality. A model 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 [32][42].
- Consistent preferred labels per language tag [42][55].
- Atypical use of collections, containers and reification [30].
- Wrong equivalent, symmetric or transitive relationships [35].
- Consistent naming criteria in the model [42][35]
- Overlapping labels in a given language for concepts in the same scheme [42].
- Membership violations for disjoint classes [30][35].

3.5. 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 [21]. Dataset freshness can be identified if the dataset contains timestamps that can keep track of its modifications. Data freshness could be considered as a subjective measure. However, our concern is the existence of temporal information allowing dataset consumers to subjectively decide its freshness for their scenario.

3.6. Provenance

Provenance can be achieved at the dataset level by including metadata that describes its authoritative information (author, maintainer, creation date, etc.), verifying if the dataset uses a provenance vocabulary like PROV [4] and uses digital signatures [22].

3.7. Licensing

Licensing is a quality attribute that is measured on the dataset level. It includes the availability of machine readable license information [31], human readable license information in the documentation of the dataset or its source [31] and the indication of permissions, copyrights and attributions specified by the author [22].

3.8. Comprehensibility

Dataset comprehensibility is identified if the publisher provides a title, description and URI for the dataset. Moreover, if he indicates at least one exemplary URI and SPARQL query, RDF file and if he provides a list of used vocabularies and an active mailing list or message board for the dataset [20].
A model is considered to be comprehensible if there is no misuse of ontology annotations and that all the concepts are documented and annotated [42][35].

3.9. Coherence

Coherence is the ability to interpret data as expected by the publisher or vocabulary maintainer [30]. The objective coherence measures are mainly associated with the modeling quality. A model is considered to be coherent when it does not contain:

- Usage of undefined classes and properties [30]. Many errors that are due to spelling or syntactic mistakes are resolvable through minor fixes via ontology checkers tools. However, for new terms, [30] suggests to have them defined in a separate namespace in order to allow reuse [42].
- Usage of blank nodes [31].
- Deprecated classes or properties [30].
- Relations and mappings clashes [55].
- Invalid inverse-functional values [30].
- Cyclic hierarchical relations [53][55][42].
- Solely transitive related concepts [42].
- Redefinitions of existing vocabularies [30].
- Valueless associative relations [42].

3.10. 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 [22].

4. Linked Data Quality Tools

In this section, we present the results of our survey on the Linked Data quality tools. There exists a number of data quality frameworks and tools that are either standalone or implemented as modules in data integration tools. These approaches can be classified into automatic, semi-automatic, manual or crowd sourced approaches.

4.1. Information Quality

RDF is the standard to model information in the Semantic Web. Linked Data publishers can pick from a plethora of tools that can automatically check their RDF files for quality problems[9]. Syntactic RDF checkers are able to detect errors in RDF documents like the W3C RDF Validator[10], RDF:about validator and Converter[11] and The Validating RDF Parser (VRP)[12]. The RDF Triple-Checker[13] is an online tool that helps find typos and common errors in RDF data. Vapour[14] [5] is a validation service to check whether semantic Web data is correctly published according to the current best practices [58].
ProLOD [8], ProLOD++ [1], Aether [48] and LODStats [18] are not purely quality assessment tools. They are Linked Data profiling tools providing 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.

4.2. Modeling 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 [56]. This can introduce deficiencies such as redundant concepts or conflicting relationships [25]. Getting to choose the right ontology or vocabulary is vital to ensure modeling correctness and consistency.

4.2.1. Semi-automatic Approaches

DL-Learner [40] uses supervised machine learning techniques to learn concepts from user-provided examples. CROCUS [13] applies a cluster-based approach for instance-level error detection. It validates identified errors by non-expert users and iterate to reach higher quality ontologies that can be safely used in industrial environments.

4.2.2. Automatic Approaches

qSKOS[15] [42] scans SKOS vocabularies to provide reports on vocabulary resources and relations that are problematic. PoolParty checker[16] is an online service based on qSKOS. Skosify [55] supports OWL and RDFS ontologies by converting them into well-structured SKOS vocabularies. It includes automatic correction abilities for quality issues that have been observed by reviewing vocabularies on the Web. The OOPS! pitfall scanner [49] evaluates OWL ontologies against a rules catalog and provides the user with a set of guidelines to solve them. ASKOSI[17] 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 [52][57] the authors perform quality checking on OWL ontologies using integrity constraints involving the Unique Name Assumption (UNA) and the Closed World Assumption (CWA). Pellet[18] 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[19]

[9] http://www.w3.org/2001/sw/wiki/SWValidators
[10] http://www.w3.org/RDF/Validator/
[11] http://rdfabout.com/demo/validator/
[12] http://139.91.183.30:9090/RDF/VRP/index.html
[13] http://graphite.ecs.soton.ac.uk/checker/
[14] http://validator.linkeddata.org/vapour
[15] https://github.com/cmader/qSKOS
[16] http://www.poolparty.biz/
[17] http://www.w3.org/2001/sw/wiki/ASKOSI
[18] http://clarkparsia.com/pellet
[19] http://jena.sourceforge.net/Eyeball/

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[20] provides validation for many issues highlighted in [30] like misplaced, undefined or deprecated classes or properties.

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. The most adopted approaches 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 Approaches

Sieve [44] is a framework for expressing quality assessment and fusion methods. It is implemented as a component of the Linked Data Integration Framework (LDIF)[21]. 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.

SWIQA [23] is a framework providing policies or formulas controlling information quality assessment. It is composed of three layers: data acquisition, query and ontology layers. It uses query templates based on the SPARQL Inferencing Notation (SPIN)[22] 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.

4.3.2. Crowd-sourcing Approaches

There are several quality issues that can be difficult to spot and fix automatically. In [2] 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 [38] is a crowd-sourcing tool used by the authors to run out their assessment supported by a semi-automatic quality verification metrics. 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. These features turn out to be extremely useful to analyze the performance of users and allow 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 interlinking with other datasets.

[20] http://swse.deri.org/RDFAlerts/
[21] http://ldif.wbsg.de/
[22] http://spinrdf.org/
[23] https://www.mturk.com/

4.3.3. Semi-automatic Approaches

Luzzu [15] is a generic Linked Data quality assessment framework. It can be easily extended through a declarative interface to integrate domain specific quality measures. The framework consists of three stages closely corresponding to the methodology in [51]. They believe that data quality cannot be tackled in isolation. As a result, they require domain experts to identify quality assessment metrics in a schema layer. Luzzu is ontology driven. The core vocabulary for the schema layer is the Dataset Quality Ontology (daQ) [14]. Any additional quality metrics added to the framework should extend it.

RDFUnit[24] is a tool centered around the definition of data quality integrity constraints [37]. The input is a defined set of test cases (which can be generated manually or automatically) presented in SPRAQL query templates. One of the main advantages for this approach is the ability to discover quality problems beyond conventional quality heuristics by encoding domain specific semantics in the test cases.

LiQuate [50] is based on probabilistic models to analyze the quality of data and links. It consists of two main components: A Bayesian Network builder and an ambiguity detector. They rely on data experts to represent probabilistic rules. LiQuate identifies redundancies (redundant label names for a given resource), incompleteness (incomplete links among a given set of resources) and inconsistencies (inconsistent links).

Quality Assessment of Data Sources (Flemming's Data Quality Assessment Tool)[25] 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.

LODGRefine [60] is the Open Refine[26] 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.

4.3.4. Automatic Ranking Approaches

Link-based Approaches

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 [39] and HITS [36] 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 [10][12].

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 [59].

The first adaption of PageRank for Semantic Web resources was the Ontology Rank algorithm implemented in the Swoogle search engine [19]. 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 [29] rankings are computed at query time based on two levels of granularity: resources and context graphs. DING [59] 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 up-

[24]http://github.com/AKSW/RDFUnit

[25]http://linkeddata.informatik.hu-berlin.de/LDSrcAss/datenquelle.php

[26]http://openrefine.org/

dated. DSNotify[27][28] 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 event log which pushes notifications to registered applications.
LinkQA [24] is a fully automated approach which takes a set of RDF triples as an input and analyses it to extract topological measures (links quality). However, the authors depend only on five metrics to determine the quality of data (degree, clustering coefficient, centrality, sameAs chains and descriptive richness through sameAs).

Provenance-based Approaches

Provenance-based assessment methods are an important step towards transparency of data quality in the Semantic Web. In [27][28] 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 [21] 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 [26] the authors introduce the notion of naming authority which connects an identifier with the source to establish a connection to its provenance. They construct a naming authority graph that acts as input to derive PageRank scores for the data sources.

Entity-based Approaches

Sindice [16] 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)[29] 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 [17]. 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 end-points. The standard query language for Semantic Web resources is SPARQL. As a result, we focus on tools measuring the quality of SPARQL endpoints. In [11][30] 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.

Summary

We notice that there is a plethora of tools (syntactic checkers or statistical profilers) that automatically check the quality of information at the entities level. Moreover, various tools can automatically check

[27]http://www.cibiv.at/ niko/dsnotify/
[28]http://trdf.sourceforge.net
[29]http://siren.sindice.com/
[30]http://labs.mondeca.com/sparqlEndpointsStatus/

the models against the objective quality indicators mentioned. OOPS! covers all of them with additional support for the other common modeling pitfalls in [35]. PoolParty covers also a wide set of those indicators but it targets SKOS vocabularies only. However, we notice a lack in automatic tools to check the dataset quality especially in its completeness, licensing and provenance measures.
As a result, we have identified the need for a complete quality framework that can automatically assess all the objective quality indicators. In section 5, we propose an extensible framework were existing tools that can automatically measure the quality of various sections (entities and models) can be plugged in as well the introduction of our dataset quality checker that covers all the mentioned dataset and links quality indicators.

5. An Extensible Objective Quality Assessment Framework

6. Experiments and Evaluation

7. 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 identified several gaps in the current tools and identified the need for a comprehensive evaluation and assessment framework.

In future work, 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] Z. Abedjan, T. Gruetze, A. Jentzsch, and F. Naumann. Profiling and mining rdf data with prolod++. In *Data Engineering (ICDE), 2014 IEEE 30th International Conference on*, pages 1198–1201, March 2014.

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

[3] A. Assaf and A. Senart. Data quality principles in the semantic web. In *Proceedings of the 2012 IEEE Sixth International Conference on Semantic Computing*, ICSC '12.

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

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

[6] C. Bizer and R. Cyganiak. Quality-driven information filtering using the wiqa policy framework. *Web Semant.*

[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*, 2009.

[8] C. Bohm, F. Naumann, Z. Abedjan, F. Dandy, T. Grutze, D. Hefenbrock, M. Pohl, and D. Sonnabend. ProÔ¨Åling Linked Open Data with ProLOD. *ICDE 2010*, 2010.

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

[10] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. In *The seventh international conference on World Wide Web 7*, 1998.

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

[12] 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.

[13] D. Cherix, R. Usbeck, A. Both, and J. Lehmann. CROCUS: Cluster-based ontology data cleansing. In *Proceedings of the 2nd International Workshop on Semantic Web Enterprise Adoption and Best Practice*, 2014.

[14] J. Debattista, C. Lange, and S. Auer. daq, an ontology for dataset quality information. In *Workshop on Linked Data on the Web co-located with the 23rd International World Wide Web Conference (WWW 2014), Seoul, Korea, April 8, 2014.*, 2014.

[15] J. Debattista, S. Londoño, C. Lange, and S. Auer. LUZZU - A framework for linked data quality assessment. *CoRR*, abs/1412.3750, 2014.

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

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

[18] J. Demter, S. Auer, M. Martin, and J. Lehmann. Lodstats—an extensible framework for high-performance dataset analytics. In *EKAW 2012*, Lecture Notes in Computer Science (LNCS) 7603. Springer, 2012.

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

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

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

[22] 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*, 2012.

[23] C. Fürber and M. Hepp. SWIQA - A Semantic Web information quality assessment framework. *ECIS 2011*, 2011.

[24] C. Guéret, P. T. Groth, C. Stadler, and J. Lehmann. Assessing linked data mappings using network measures. In *The 9th Extended Semantic Web Conference*, 2012.

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

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

[27] 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.

[28] B. Haslhofer and N. Popitsch. DSnotify: Detecting and fixing broken links in linked data sets. In *8th International Workshop on Web Semantics (WebS ’09), co-located with DEXA 2009*, 2009.

[29] 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.

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

[31] A. Hogan, J. Umbrich, A. Harth, R. Cyganiak, A. Polleres, and S. Decker. An empirical survey of linked data conformance. *Web Semant.*, 2012.

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

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

[34] B. K. Kahn, D. M. Strong, and R. Y. Wang. Information quality benchmarks: product and service performance. *Communications of the ACM*, 2002.

[35] C. M. Keet, M. del Carmen Suárez-Figueroa, and M. Poveda-Villalón. The current landscape of pitfalls in ontologies. In *KEOD 2013 - Proceedings of the International Conference on Knowledge Engineering and Ontology Development, Vilamoura, Algarve, Portugal, 19-22 September, 2013*, 2013.

[36] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. *J. ACM*, 1999.

[37] D. Kontokostas, P. Westphal, S. Auer, S. Hellmann, J. Lehmann, R. Cornelissen, and A. Zaveri. Test-driven evaluation of linked data quality. In *Proceedings of the 23rd International Conference on World Wide Web*, 2014.

[38] 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*, 2013.

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

[40] J. Lehmann and S. Sonnenburg. Dl-learner: Learning concepts in description logics. *Journal of Machine Learning Research*, 2009.

[41] H. Living. Review of: Hedden, heather. the accidental taxonomist medford, nj: Information today, inc., 2010. *Inf. Res.*, 2010.

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

[43] N. Marie, F. Gandon, M. Ribière, and F. Rodio. Discovery hub: On-the-fly linked data exploratory search. In *The 9th International Conference on Semantic Systems*, 2013.

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

[45] P. N. Mendes, M. Jakob, A. García-Silva, and C. Bizer. Dbpedia spotlight: Shedding light on the web of documents. In *The 7th International Conference on Semantic Systems*, 2011.

[46] A. Miles and S. Bechhofer. Skos simple knowledge organization system reference. w3C recommendation 18 August 2009., 2009.

[47] R. Mirizzi, T. D. Noia, A. Ragone, V. C. Ostuni, and E. D. Sciascio. Movie recommendation with dbpedia. CEUR Workshop Proceedings, 2012.
[48] E. Mkel. Aether – generating and viewing extended void statistical descriptions of rdf datasets. In *ESWC 2014 demo track, Springer-Verlag*, 2014.
[49] M. Poveda-Villal$\surd\geq$n, M. Su$\surd$degrez-Figueroa, and A. G$\surd\geq$mez-P$\surd$©rez. Validating ontologies with OOPs! In *Knowledge Engineering and Knowledge Management*. Springer Berlin Heidelberg, 2012.
[50] E. Ruckhaus, O. Baldizan, and M.-E. Vidal. Analyzing linked data quality with liquate. In *OTM Workshops*, Lecture Notes in Computer Science, 2013.
[51] A. Rula and A. Zaveri. Methodology for assessment of linked data quality. In *The 1st Workshop on Linked Data Quality co-located with 10th International Conference on Semantic Systems, LDQ@SEMANTiCS 2014, Leipzig, Germany, September 2nd, 2014.*, 2014.
[52] E. Sirin, M. Smith, and E. Wallace. Opening, closing worlds - on integrity constraints. 2008.
[53] D. Soergel. Thesauri and ontologies in digital libraries. In *JCDL*. ACM, 2005.
[54] 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*, 2007.
[55] O. Suominen and E. Hyvönen. Improving the quality of skos vocabularies with skosify. In *The 18th international conference on Knowledge Engineering and Knowledge Management*, 2012.
[56] O. Suominen and C. Mader. Assessing and Improving the Quality of SKOS Vocabularies. *Journal on Data Semantics*, 2013.
[57] 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.
[58] B.-L. Tim. Linked data. Technical report, W3C, 2006.
[59] N. Toupikov, J. Umbrich, and R. Delbru. DING! Dataset ranking using formal descriptions. *WWW09*, 2009.
[60] M. Verlic. Lodgrefine - lod-enabled google refine in action. In *I-SEMANTICS (Posters & Demos)*. CEUR-WS.org, 2012.
[61] R. Wang and D. Strong. Beyond accuracy: What data quality means to data consumers. *Journal of management information systems*, 1996.