An Objective Assessment Framework & Tool for Linked Data Quality

Enriching Dataset Profiles with Quality Indicators

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 a complex process as 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. Previous efforts have identified potential quality issues of Linked Data and listed quality principles for all stages of data management. We build upon these efforts 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 state of the datasets needs attention as they mostly have low completeness, provenance, licensing and comprehensibility quality scores.

Keywords: Data Quality, Linked Data, Quality Framework, Semantic Web, Dataset Profile, Profile Generation

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 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 [15].

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

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 [36,10,63]. 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 [43], 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 [51]. Ensuring data quality in Linked Open Data is a complex process as 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 [12]. The dimensionality of data quality makes it dependent on the task and users requirements. For example, DBpedia [13] and YAGO [56] are knowledge bases containing data extracted from structured and semi-structured sources. They are used in a variety of applications e.g., annotation systems [48], exploratory search [46] and recommendation engines [50]. However, their data is not 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 described in [4] and surveyed in [64]. 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 extend Roomba which is a framework to assess and build dataset profiles with an extensible quality measurement tool and evaluate it by measuring the quality of the LOD cloud group. 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 in data quality assessment methodologies. Section 3 presents our framework with its objective quality measures and indicators. Section 4 reviews the existing tools and frameworks in the Linked Open Data quality landscape. Section 5 presents our tool for evaluating those indicators. Section 6 presents concluding remarks and identifies future work.

2. Related Work

In [64], 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. At the time of writing this paper, there was only 10,555 ODI certificates issued. 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 aims at promoting graph and RDF data management systems to be an accepted industrial solution. LDBC is not focused around measuring or assessing quality. However, it focuses on creating benchmarks to measure progress in scalability, storage, indexing and query optimization techniques to become the de facto standard for publishing performance results.

In [3], 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

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 [8]:

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

[3] http://ldbc.eu/

– **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 [4], the authors identified 24 different Linked Data quality attributes. These attributes are a mix of objective and subjective meaasures that may not be derived automatically. In this paper, we refine these attributes into a condensed framework of 10 objective measures. Since these measures are rather abstract, we should rely on quality indicators that reflect data quality [24] and use them to automate calculating datasets quality.

The quality indicators are weighted. These weights give the flexibility to define multiple degrees of importance. For example, a dataset containing people can have more than one person with the same name thus it is not always true that two entities in a dataset should not have the same preferred label. As a result, the weight for that quality indicator will be set to zero and will not affect the overall quality score for the consistency measure.

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 objective quality indicators. Those indicators have been gathered by:

– Transforming the objective quality indicators presented as a set of questions in [4] into more concrete quality indicator metrics.
– Surveying the landscape of data quality tools and frameworks.
– Examining the properties of the most prominent linked data models from the survey done in [5].

Table 1: Objective Linked Data Quality Framework

Quality Attribute	Quality Category	ID	Quality Indicator
Completeness	Dataset Level	1	Existence of supporting structured metadata [33]
		2	Supports multiple serializations [64]
		3	Has different data access points
		4	Uses datasets description vocabularies
		5	Existence of descriptions about its size
		6	Existence of descriptions about its structure (MIME Type, Format)
		7	Existence of descriptions about its organization and categorization
		8	Existence of information about the kind and number of used vocabularies [64]
	Links Level	9	Existence of dereferencable links for the dataset [33,44,27]
	Model Level	10	Absence of disconnected graph clusters [44]
		11	Absence of omitted top concept [33]
		12	Has complete language coverage [44]
		13	Absence of unidirectional related concepts [33]
		14	Absence of missing labels [44]

Continued on next page

Table 1 Objective Linked Data Quality Framework

Quality Attribute	Quality Category	ID	Quality Indicator
		15	Absence of missing equivalent properties [37]
		16	Absence of missing inverse relationships [37]
		17	Absence of missing domain or range values in properties [37]
Availability	Dataset Level	18	Existence of an RDF dump that can be downloaded by users [24][33]
		19	Existence of a queryable endpoint that responds to direct queries
		20	Existence of valid dereferencable URLs (respond to HTTP request)
Licensing	Dataset Level	21	Existence of human and machine readable license information [34]
		22	Existence of de-referenceable links to the full license information [34]
		23	Specifies permissions, copyrights and attributions [64]
Freshness	Dataset Level	24	Existence of timestamps that can keep track of its modifications [25]
Correctness	Dataset Level	25	Includes the correct MIME-type for the content [33]
		26	Includes the correct size for the content
		27	Absence of syntactic errors on the instance level [33]
	Links Level	28	Absence of syntactic errors [58]
		29	Use the HTTP URI scheme (avoid using URNs or DOIs) [44]
	Model Level	30	Contains marked top concepts [44]
		31	Absence of broader concepts for top concepts [44]
		32	Absence of missing or empty labels [2,44]
		33	Absence of unprintable characters [2,44] or extra white spaces in labels [57]
		34	Absence of incorrect data type for typed literals [33,2]
		35	Absence of omitted or invalid languages tags [57,44]
		36	Absence of terms without any associative or hierarchical relationships
Comprehensibility	Dataset Level	37	Existence of at least one exemplary RDF file [64]
		38	Existence of at least one exemplary SPARQL query [64]
		39	Existence of general information (title, URL, description) for the dataset
		40	Existence of a mailing list, message board or point of contact [24]
	Model Level	41	Absence of misuse of ontology annotations [44,37]
		42	Existence of annotations for concepts [37]
		43	Existence of documentation for concepts [44,37]
Provenance	Dataset Level	44	Existence of metadata that describes its authoritative information [25]
		45	Usage of a provenance vocabulary
		46	Usage of a versioning
Coherence	Model Level	47	Absence of misplaced or deprecated classes or properties [33]
		48	Absence of relation and mappings clashes [57]
		49	Absence of blank nodes [34]
		50	Absence of invalid inverse-functional values [33]
		51	Absence of cyclic hierarchical relations [55,57,44]
		52	Absence of undefined classes and properties usage [33]
		53	Absence of solely transitive related concepts [44]
		54	Absence of redefinitions of existing vocabularies [33]
		55	Absence of valueless associative relations [44]
Consistency	Model Level	56	Consistent usage of preferred labels per language tag [35,44]
		57	Consistent usage of naming criteria for concepts [37]
		58	Absence of overlapping labels
		59	Absence of disjoint labels [44]
		60	Absence of atypical use of collections, containers and reification [33]
		61	Absence of wrong equivalent, symmetric or transitive relationships [37]
		62	Absence of membership violations for disjoint classes [33]
Security	Dataset Level	63	Uses login credentials to restrict access [64]
		64	Uses SSL or SSH to provide access to their dataset [64]

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, opposite to other measures like availability where i can measure if a dataset is available or not despite of the underlying use case. For example, an entity is considered to be complete if it contains all the attributes needed for a given task, has complete language coverage [44] and has documentation properties [49,44]. Dataset completeness has some objective indicators which we include in our framework. A dataset is considered to be complete if it:

- Contains supporting structured metadata [33].
- Provides data in multiple serializations (N3, Turtle, etc.) [64].
- Contains different data access points. These can either be a queryable endpoint (i.e. SPARQL endpoint, REST API, etc.) or a data dump file.
- Uses datasets description vocabularies like DCAT[4] or VOID[5].
- Provides descriptions about its size e.g., `void:statItem`, `void:numberOfTriples` or `void:numberOfDocuments`.
- Existence of descriptions about its format.
- Contains information about its organization and categorization e.g., `dcterms:subject`.
- Contains information about the kind and number of used vocabularies [64].

Links are considered to be complete if the dataset and all its resources have defined links [33,44,27]. Models are considered to be complete if they do not contain disconnected graph clusters [44]. 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 (each concept labeled in each of the languages that are also used on the other concepts) [44], do not contain omitted top concepts or unidirectional related concepts [33] and if they are not missing labels [44], equivalent properties, inverse relationships, domain or range values in properties [37].

3.2. Availability

A dataset is considered to be available if the publisher provides data dumps e.g., RDF dump, that can be downloaded by users [24,33], its queryable endpoints e.g., SPARQL endpoint, are reachable and respond to direct queries and if all of its inbound and outbound links are dereferencable.

3.3. Correctness

A dataset is considered to be correct if it includes the correct MIME-type and size for the content [33] and doesn't contain syntactic errors [33]. Links are considered to be correct if they lack syntactic errors and use the HTTP URI scheme (avoid using URNs or DOIs) [44]. 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) [44]. Moreover, if they don't contain incorrect data type for typed literals [33][2], no omitted or invalid languages tags [57,44], does not contain "orphan terms" (orphan terms are terms without any associative or hierarchical relationships and if the labels are not empty, do not contain unprintable characters [2,44] or extra white spaces [57].

---

[4] `http://www.w3.org/TR/vocab-dcat/`
[5] `http://www.w3.org/TR/void/`

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 (two concepts having the same preferred lexical label in a given language when they belong to the same schema) [35,44], consistent preferred labels per language tag [44,57], atypical use of collections, containers and reification [33], wrong equivalent, symmetric or transitive relationships [37], consistent naming criteria in the model [44,37], overlapping labels in a given language for concepts in the same scheme [44] and membership violations for disjoint classes [33,37].

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 [25]. 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.), versioning information and verifying if the dataset uses a provenance vocabulary like PROV [41].

3.7. Licensing

Licensing is a quality attribute that is measured on the dataset level. It includes the availability of machine readable license information [34], human readable license information in the documentation of the dataset or its source [34] and the indication of permissions, copyrights and attributions specified by the author [64].

3.8. Comprehensibility

Dataset comprehensibility is identified if the publisher provides general information about the dataset (e.g., title, description, URI). In addition, if he indicates at least one exemplary RDF file and SPARQL query and provides an active communication channel (mailing list, message board or e-mail) [24]. A model is considered to be comprehensible if there is no misuse of ontology annotations and that all the concepts are documented and annotated [44,37].

3.9. Coherence

Coherence is the ability to interpret data as expected by the publisher or vocabulary maintainer [33]. The objective coherence measures are mainly associated with the modeling quality. A model is considered to be coherent when it does not contain undefined classes and properties [33], blank nodes [34], deprecated classes or properties [33], relations and mappings clashes [57], invalid inverse-functional values [33], cyclic hierarchical relations [55,57,44], solely transitive related concepts [44], redefinitions of existing vocabularies [33] and valueless associative relations [44].

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 [64].

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 crowdsourced 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[6]. Syntactic RDF checkers are able to detect errors in RDF documents like the W3C RDF Validator[7], RDF:about validator and Converter[8] and The Validating RDF Parser (VRP)[9]. The RDF Triple-Checker[10] is an online tool that helps find typos and common errors in RDF data. Vapour[11] [9] is a validation service to check whether semantic Web data is correctly published according to the current best practices [8].

ProLOD [14], ProLOD++ [1], Aether [45] and LODStats [7] 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 [58]. This can introduce deficiencies such as redundant concepts or conflicting relationships [28]. Getting to choose the right ontology or vocabulary is vital to ensure modeling correctness and consistency.

4.2.1. Semi-automatic Approaches

DL-Learner [42] uses supervised machine learning techniques to learn concepts from user-provided examples. CROCUS [19] 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[12] [44] scans SKOS vocabularies to provide reports on vocabulary resources and relations that are problematic. PoolParty checker[13] is an online service based on qSKOS. Skosify [57] 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 [52] evaluates OWL ontologies against a rules catalog and provides the user with a set of guidelines to solve them. ASKOSI[14] 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.

---

[6] http://www.w3.org/2001/sw/wiki/SWValidators
[7] http://www.w3.org/RDF/Validator/
[8] http://rdfabout.com/demo/validator/
[9] http://139.91.183.30:9090/RDF/VRP/index.html
[10] http://graphite.ecs.soton.ac.uk/checker/
[11] http://validator.linkeddata.org/vapour
[12] https://github.com/cmader/qSKOS
[13] http://www.poolparty.biz/
[14] http://www.w3.org/2001/sw/wiki/ASKOSI

Some errors in RDF will only appear after reasoning (incorrect inferences). In [54,59] the authors perform quality checking on OWL ontologies using integrity constraints involving the Unique Name Assumption (UNA) and the Closed World Assumption (CWA). Pellet[15] 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[16] 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[17] provides validation for many issues highlighted in [33] 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 [47] is a framework for expressing quality assessment and fusion methods. It is implemented as a component of the Linked Data Integration Framework (LDIF)[18]. 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 [26] 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)[19] 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 [13]). 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) Crowdsourcing through the Amazon Mechanical Turk[20].

TripleCheckMate [40] is a crowdsourcing 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

[15] http://clarkparsia.com/pellet
[16] http://jena.sourceforge.net/Eyeball/
[17] http://swse.deri.org/RDFAlerts/
[18] http://ldif.wbsg.de/
[19] http://spinrdf.org/
[20] https://www.mturk.com/

extraction value for object values and data types), relevancy of the extracted information, representational consistency and interlinking with other datasets.

4.3.3. Semi-automatic Approaches

Luzzu [21] 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 [3]. 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) [20]. Any additional quality metrics added to the framework should extend it.

RDFUnit[21] is a tool centered around the definition of data quality integrity constraints [39]. The input is a defined set of test cases (which can be generated manually or automatically) presented in SPARQL 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 [53] 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)[22] 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 [62] is the Open Refine[23] 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

The Project Open Data Dashboard[24] tracks and measures how US government websites implement the Open Data principles to understand the progress and current status of their public data listings. A validator analyzes machine readable files e.g., JSON files for automated metrics like the resolved URLs, HTTP status and content-type. However, deep schema information about the metadata is missing like description, license information or tags.

Similarly on the LOD cloud, the Data Hub LOD Validator[25] 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.

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

[21] http://github.com/AKSW/RDFUnit
[22] http://linkeddata.informatik.hu-berlin.de/LDSrcAss/datenquelle.php
[23] http://openrefine.org/
[24] http://labs.data.gov/dashboard/
[25] http://validator.lod-cloud.net/

way to measure the quality of Web documents search. Algorithms like PageRank [51] and HITS [38] 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 [16][18]. 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 [60]. The first adaption of PageRank for Semantic Web resources was the Ontology Rank algorithm implemented in the Swoogle search engine [23]. 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 [32] rankings are computed at query time based on two levels of granularity: resources and context graphs. DING [60] 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[26][31] 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 [27] is a fully automated approach which takes a set of RDF triples as an input and analyzes 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 [30][27] 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 [25] 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 [29] 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 [61] 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)[28] 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 [22]. 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 [17][29] the authors present their findings to measure

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

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.

5. An Extensible Objective Quality Assessment Framework

Dataset profiles are collections of data describing the internal structure of the dataset. They are presented as a set of metadata in different formats such as JSON, XML and RDF. The Linked Data publishing best practices [11] specifies that datasets should contain metadata needed to effectively understand and use them.

Data portals (or data catalogs) are the entry points to discover published datasets. They are a curated collections of datasets metadata that provide a set of complementary discovery and integrations services. Data portals can be considered as data access points providing tools to facilitate data publishing, sharing, searching and visualization. CKAN[30] is the world's leading open-source data portal platform powering websites `Datahub.io` and `publicdata.eu`.

Looking at the list of objective quality indicators, we found out that a large amount of those indicators can be examined automatically from attached datasets metadata found in data portals. As a result, we have chosen to extend Roomba, a scalable automatic approach for extracting, validating, correcting and generating descriptive linked dataset profiles [6]. Roomba is built as a Command Line Interface (CLI) application using Node.js. Instructions on installing and running the framework are available on its public Github repository. Figure 1 shows the main steps which are the following: (i) Data portal identification; (ii) metadata extraction; (iii) instance and resource extraction; (iv) profile validation (v) profile and report generation. Roomba's advantages lay in being easy to extend as it uses a modular pluggable approach and because it already performs several pre-processing steps needed to fetch, sample, cache and validate datasets metadata.

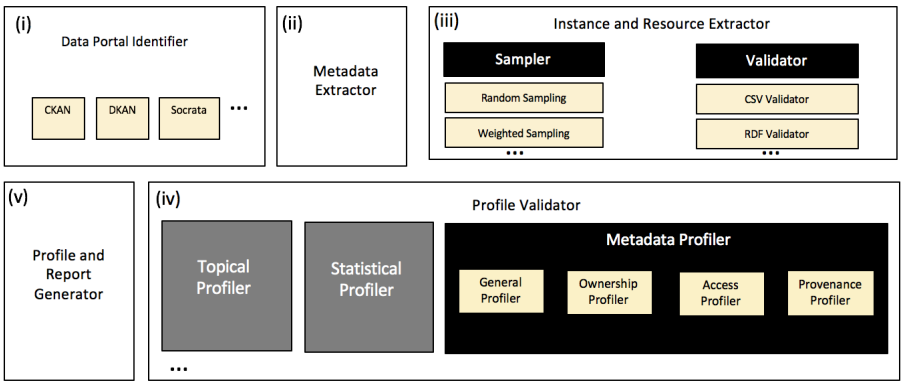

Fig. 1. Processing pipeline for objective dataset quality assessment

In our framework, we have presented 30 objective quality indicators related to dataset and links quality. The remainder 34 indicators are related to the entities and models quality and cannot be checked through the attached metadata. Excluding security related quality indicators as LOD cloud group members should not restrict access to their datasets, the Roomba quality extension is able to assess and score 23 of them (82%).

[30] `http://ckan.org`

We have extended Roomba with 7 submodules that will check various dataset quality indicators shown in table 2. Some indicators have to be examined against a finite set. For example, to measure the quality indicator no.3 (having different data access points), we need to have a defined set of access points in order to calculate a quality score. Since Roomba runs on CKAN-based data portals, we built our quality extension to calculate the scores against the CKAN standard model[31].

Quality Indicator	Assessment Method
1	Check if there is a valid metadata file by issuing a `package_show` request to the CKAN API
2	Check if the `format` field for the dataset resources is defined and valid
3	Check the `resource_type` field with the following possible values `file, file.upload, api, visualization, code, documentation`
4	Check the resources `format` field for `meta/void` value
5	Check the resources `size` or the `triples` extras fields
6	Check the `format` and `mimetype` fields for resources
7	Check if the dataset has a `topic` tag and if it is part of a valid group in CKAN
9	Check if the dataset and all its resources have has a valid URI
18	Check if there is a dereferencable resource with a description containing string *dump*
19	Check if there is a dereferencable resource with `resource_type` of type `api`
20	Check if all the links assigned to the dataset and its resources are dereferencable
21	Check if the dataset contains valid `license_id` and `license_title`
22	Check if the `license_url` is dereferencable
24	Check if the dataset and its resources contain the following metadata fields `metadata_created, metadata_modified, revision_timestamp, cache_last_updated`
25	Check if the `content-type` extracted from the a valid HTTP request is equal to the corresponding `mimetype` field.
26	Check if the `content-length` extracted from the a valid HTTP request is equal to the corresponding `size` field.
28,29	Check that all the links are valid HTTP scheme URIs
37	Check if there is at least one resource with a `format` value corresponding to one of `example/rdf+xml, example/turtle, example/ntriples, example/x-quads, example/rdfa, example/x-trig`
39	Check if the dataset and its tags and resources contain general metadata `id, name, type, title, description, URL, display_name, format`
40	Check if the dataset contain valid `author_email` or `maintainer_email` fields
44	Check if the dataset and its resources contain provenance metadata `maintainer, owner_org, organization, author, maintainer_email, author_email`
46	Check if the dataset contain and its resources contain versioning information `version, revision_id`

Table 2
Objective Quality Assessment Methods for CKANbased Data Portals

[31] `http://demo.ckan.org/api/3/action/`

5.1. Quality Score Calculation

A CKAN dataset model describes four main sections in addition to the core dataset's properties. These sections are:

- **Resources**: The distributable parts containing the actual raw data. They can come in various formats (JSON, XML, RDF, etc.) and can be downloaded or accessed directly (REST API, SPARQL endpoint).
- **Tags**: Provide descriptive knowledge on the dataset content and structure. They are used mainly to facilitate search and reuse.
- **Groups**: A dataset can belong to one or more group that share common semantics. A group can be seen as a cluster or a curation of datasets based on shared categories or themes.
- **Organizations**: A dataset can belong to one or more organization controlled by a set of users. Organizations are different from groups as they are not constructed by shared semantics or properties, but solely on their association to a specific administration party.

A CKAN portal contains a set of datasets $\mathbf{D} = \{D_1, ...D_n\}$. We denote the set of resources $R_i = \{r_1, ..., r_k\}$, groups $G_i = \{g_1, ..., g_k\}$ and tags $T_i = \{t_1, ..., t_k\}$ for $D_i \in \mathbf{D}(i = 1, ..., n)$ by $\mathbf{R} = \{R_1, ..., R_n\}, \mathbf{G} = \{G_1, ..., G_n\}$ and $\mathbf{T} = \{T_1, ..., t_n\}$ respectively.

Our quality framework contains a set of measures $\mathbf{M} = \{M_1, ..., M_n\}$. We denote the set of quality indicators $Q_i = \{q_1, ..., q_k\}$ for $M_i \in \mathbf{M}(i = 1, ..., n)$ by $\mathbf{Q} = \{Q_1, ..., Q_n\}$. Each quality indicator has a weight, context and a score $Q_i < weight, context, score >$. In Roomba, all the weights are equal and set to 1. However, they can be adjusted manually to rank the quality indicators. Each Q_i of M_i (for $i =$ 1,...n) is applied to one or more of the resources, tags or groups. The indicator context is defined where $\exists Q_i \in \mathbf{R} \cup \mathbf{G} \cup \mathbf{T}$.

The quality indicator score is based on a ratio between the number of violations **V** and the total number of instances where the rule applies **T** multiplied by the specified weight for that indicator. In some cases, the quality indicator score is a boolean value (0 or 1). For example, checking if there is a valid metadata file (QI.1) or checking if the `license_url` is dereferencable (QI.22).

$$Q \text{ weightedscore} = (V/T) * Q < weight > \tag{1}$$

Q weightedscore is an error ratio. A quality measure score should reflect the alignment of the dataset with respect to the quality indicators. The quality measure score **M** is calculated by dividing the weighted quality indicator scores sum by the total number of instances in its context, as the following formula shows:

$$M = 1 - ((\sum_{i=1}^{n} Q\text{i weightedscore})/ \mid Qi \text{ context} \mid) \tag{2}$$

5.2. Evaluation

In our evaluation, similarily to Roomba we focused on two aspects: i) *quality profiling correctness* which manually assesses the validity of the errors generated in the report, and ii) *quality profiling completeness* which assesses if Roomba covers all the quality indicators in table 2.

Profiling Correctness

To measure profile correctness, we need to make sure that the issues reported by Roomba are valid. On the dataset level, we chose five datasets from the LOD Cloud detailed in table 3.

After running Roomba and examining the results on the selected datasets and groups, we found out that our framework provides 100% correct results on the individual dataset level. Roomba's aggregation have been evaluated in [6], thus we can infer that the quality profiler at the group and portal level also produces correct profiles.

Dataset ID	dbpedia	event-media	geolinkeddata	nytimes-linked-open-data	yovisto
Resources	10	9	4	5	6
Tags	21	15	13	14	20

Table 3
Datasets chosen for the correctness evaluation

Profiling Completeness

We analyzed the completeness of our framework by manually constructing a synthetic set of profiles[32]. These profiles cover the indicators in table 2. After running our framework at each of these profiles, we measured the completeness and correctness of the results. We found out that our framework covers indeed all the quality problems discussed.

5.3. Experiments and Analysis

In this section, we provide the experiments done using the proposed framework. Listing 1 shows an excerpt of the generated quality report. All the experiments are reproducible by Roomba and their results are available on its Github repository. We have run the framework on the LOD cloud containing 259 datasets at the time of writing this paper. We ran the instance and resource extractor in order to cache the metadata files for these datasets locally and ran the quality assessment process which took around two hours on a 2.6 Ghz Intel Core i7 processor with 16GB of DDR3 memory machine.

We found out that licensing, availability and comprehensibility had the worst quality measures scores: 19.59%, 26.22% and 31.62% respectively. On the other hand, the LOD cloud datasets have good quality scores for freshness, correctness and provenance as most of the datasets have an average of 75% for each one of those measures.

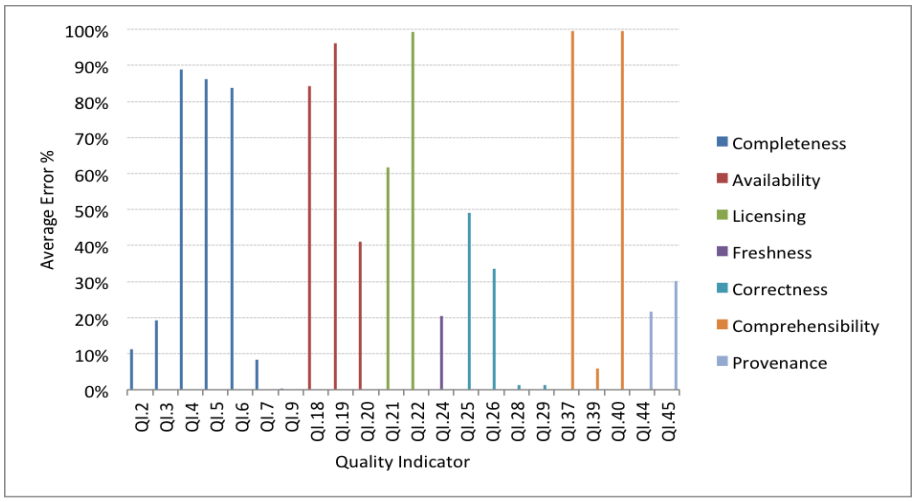

Fig. 2. Average Error % per quality indicator for LOD group

Figure 2 shows the average errors percentage in quality indicators grouped by the corresponding measures. The error percentage is the inverse quality. For example, 86.3% of the datasets resources do not have information about its size, which means that only 13.7% of the datasets are considered in good quality for this indicator. After examining the results, we notice that the worst quality indicators scores are for the

[32] https://github.com/ahmadassaf/opendata-checker/tree/master/test

comprehensibility measure where 99.61% of the datasets did not have valid exemplary RDF file (QI.37) and did not define valid point of contact (QI.40). Moreover, we noticed that 96.41% of the datasets queryable endpoints (SPARQL endpoints) failed to respond to direct queries (QI.19). After careful examination, we found that the cause was incorrect assignment for metadata fields. Data publishers specified the resource `format` field as an `api` instead of the specifying the `resource_type` field.

```
Dataset Quality Report

completeness quality Score      :   50.22%
availability quality Score      :   26.22%
licensing quality Score         :   19.59%
freshness quality Score         :   79.49%
correctness quality Score       :   72.06%
comprehensibility quality Score :   31.62%
provenance quality Score        :   74.07%
Average total quality Score     :   50.47%

Quality Indicators Average Error %

Quality Indicator : Supports multiple serializations: 11.35%
Quality Indicator : Has different data access points: 19.31%
Quality Indicator : Uses datasets description vocabularies: 88.80%
Quality Indicator : Existence of descriptions about its size: 86.30%
Quality Indicator : Existence of descriptions about its structure: 83.67%
```

Listing 1: Excerpt of the LOD cloud group quality report

To drill down more on the availability issues, we generated a metadata profile assessment report using Roomba's metadata profiler. We found out that 25% of the datasets access information (being the dataset URL and any URL defined in its groups) has issues related to them (missing or unreachable URLs). Three datasets (1.15%) did not have a URL defined while 45 datasets (17.3%) defined URLs were not accessible at the time writing this paper. Out of the 1068 defined resources 31.27% were not reachable. All these issues resulted in a 26.22% average availability score. This can highly affect the usability of those datasets especially in an enterprise context.

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 the models against the objective quality indicators mentioned. OOPS! covers all of them with additional support for the other common modeling pitfalls in [37]. 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. Table 3 summarizes the automatic dataset quality approaches that have implemented tools (full circle denotes full quality indicator assessment, while half circle denoted partial assessment). As can be seen in this table Roomba covers most of the quality indicators with its focus on completeness, correctness provenance and licensing. Roomba is not able to check the existence of information about the kind and number of used vocabularies (QI.8), license permissions, copyrights and attributes (QI.23), exemplary SPARQL query (QI.38), usage of provenance vocabulary (QI.45) and is not able to check the dataset for syntactic errors (QI.27).

These shortcomings are mainly due to the limitations in the CKAN dataset model. However, syntactic checkers and additional modules to examine vocabularies usage could be easily integrated in Roomba to fix QI.27, QI.8 and QI.45. Roomba's metadata quality profiler can fix QI.23 as we have manually created a

mapping file standardizing the set of possible license names and their information[33]. We have also used the open source and knowledge license information[34] to normalize license information and add extra metadata like the domain, maintainer and open data conformance.

Tool\Indicator	1	2	3	4	5	6	7	8	9	18	19	20	21	22	23	24	25	26	27	28	29	37	38	39	40	44	45	46	63	64
LOV	●		●	●	◗		◗		●	●		●	◗									●		◗		●		●		
Data.gov	●				◗	●			●			●				◗	●							●		◗				
Roomba	●	●	●	●	●	●	●		●	●	●	●	●	●		●	●	●	●	●		●		●	●	●		●		

Table 4
Functional Comparison of Automatic Linked Data quality Tools

6. Conclusions and Future Work

In this paper, we have presented a comprehensive objective quality framework applied to the Linked Open Data. We have built upon previous efforts with focus on objective data quality measures. We have identified a total of 64 quality indicators that were mapped when suitable to four main categories (entity, dataset, links, models). We have also surveyed more than 30 different tools that measure different quality aspects of Linked Open Data. We identified several gaps in the current tools and identified the need for a comprehensive evaluation and assessment framework and specifically for measuring quality on the dataset level. As a result, we presented an extension of Roomba (An extensible tool to assess and generate dataset profiles) that covers 82% of the suggested datasets objective quality indicators. Based on our experiments running Roomba on the LOD cloud, we discovered that the general state of the datasets needs attention as most of them have low completeness, provenance, licensing and comprehensibility quality scores.

In future work, we plan to integrate tools assessing models quality in addition to syntactic checkers with Roomba. This will provide a complete coverage of the proposed quality indicators. We also intend to suggest ranked quality indicators to improve the quality report. We also plan to run this tool on various CKAN based data portals and schedule periodic reports to monitor their quality evolution. Finally, at some stage, we plan to extend this tool for other data portal types like DKAN and Socrata.

References

[1] Z. Abedjan, T. Gruetze, A. Jentzsch, and F. Naumann. Profiling and mining RDF data with ProLOD++. In *30th IEEE International Conference on Data Engineering (ICDE)*, pages 1198–1201, 2014.

[2] M. Acosta, A. Zaveri, E. Simperl, and D. Kontokostas. Crowdsourcing Linked Data quality assessment. In *12th International Semantic Web Conference (ISWC)*, 2013.

[3] R. Anisa and A. Zaveri. Methodology for Assessment of Linked Data Quality. In *1st Workshop on Linked Data Quality (LDQ)*, 2014.

[4] A. Assaf and A. Senart. Data Quality Principles in the Semantic Web. In *6th International Conference on Semantic Computing ICSC '12*, 2012.

[5] A. Assaf, R. Troncy, and A. Sénart. HDL-Towards a Harmonized Dataset Model for Open Data Portals. In *2nd International Workshop on Dataset PROFIling & fEderated Search for Linked Data*, Portoroz, Slovenia, 2015.

[6] A. Assaf, R. Troncy, and A. Sénart. Roomba: An Extensible Framework to Validate and Build Dataset Profiles. In *12th European Semantic Web Conference (ESWC)*, Portoroz, Slovenia, 2015.

[7] S. Auer, J. Demter, M. Martin, and J. Lehmann. LODStats - an Extensible Framework for High-performance Dataset Analytics. In *18th International Conference on Knowledge Engineering and Knowledge Management (EKAW)*, pages 353–362, Galway, Ireland, 2012.

[8] T. Berners-Lee. Linked Data - Design Issues. W3C Personal Notes, 2006. `http://www.w3.org/DesignIssues/LinkedData`.

[33] `https://github.com/ahmadassaf/opendata-checker/blob/master/util/licenseMappings.json`
[34] `https://github.com/okfn/licenses`

[9] D. Berrueta, S. Fern"andez, and I. Frade. Cooking HTTP content negotiation with Vapour. In *4th Workshop on Scripting for the Semantic Web (SFSW'08)*, 2008.

[10] G. L. S. Besiki, , M. B. Twidale, and L. C. Smith. A framework for information quality assessment. *Journal of the American Society for Information Science and Technology*, 2007.

[11] C. Bizer. Evolving the Web into a Global Data Space. In *28th British National Conference on Advances in Databases*, 2011.

[12] C. Bizer and R. Cyganiak. Quality-driven information filtering using the wiqa policy framework. *Jorunal of Web Semantics*, 7(1), 2009.

[13] C. Bizer, J. Lehmann, G. Kobilarov, S. Auer, C. Becker, R. Cyganiak, and S. Hellmann. DBpedia - A Crystallization Point for the Web of Data. *Journal of Web Semantics*, 7(3), 2009.

[14] C. Bohm, F. Naumann, Z. Abedjan, D. Fenz, T. Grutze, D. Hefenbrock, M. Pohl, and D. Sonnabend. Profiling linked open data with ProLOD. In *26th International Conference on Data Engineering Workshops (ICDEW)*, 2010.

[15] 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.

[16] S. Brin and L. Page. The anatomy of a large-scale hypertextual Web search engine. In *7th International Conference on World Wide Web (WWW'98)*, 1998.

[17] C. Buil-Aranda and A. Hogan. SPARQL Web-Querying Infrastructure: Ready for Action? In *12th International Semantic Web Conference (ISWC)*, 2013.

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

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

[20] 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)*, 2014.

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

[22] R. Delbru, N. Toupikov, and M. Catasta. Hierarchical link analysis for ranking web data. In *7th European Semantic Web Conference (ESWC)*, 2010.

[23] L. Ding, T. Finin, A. Joshi, R. Pan, and R. Cost. Swoogle: A semantic web search and metadata engine. In *13st ACM International Conference on Information and Knowledge Management (CIKM)*, 2004.

[24] A. Flemming. Quality Characteristics of Linked Data Publishing Datasources. Master's thesis, Humboldt-Universitt zu Berlin, 2010.

[25] G. Flouris, Y. Roussakis, and M. Poveda-Villalón. Using provenance for quality assessment and repair in linked open data. In *2nd Joint Workshop on Knowledge Evolution and Ontology Dynamics (EvoDyn'12)*, 2012.

[26] C. Fürber and M. Hepp. SWIQA - A Semantic Web information quality assessment framework. 2011.

[27] C. Guéret, P. Groth, C. Stadler, and J. Lehmann. Assessing Linked Data Mappings Using Network Measures. In *9th European Semantic Web Conference (ESWC)*, 2012.

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

[29] A. Harth, S. Kinsella, and S. Decker. Using naming authority to rank data and ontologies for web search. In *8th International Semantic Web Conference (ISWC)*, 2009.

[30] O. Hartig and J. Zhao. Using web data provenance for quality assessment. In *8th International Semantic Web Conference (ISWC)*, 2009.

[31] B. Haslhofer and N. Popitsch. DSNotify: Detecting and Fixing Broken Links in Linked Data Sets. In *8th International Workshop on Web Semantics*, 2009.

[32] A. Hogan, A. Harth, and S. Decker. ReConRank: A Scalable Ranking Method for Semantic Web Data with Context. In *2nd Workshop on Scalable Semantic Web Knowledge Base Systems*, 2006.

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

[34] A. Hogan, J. Umbrich, A. Harth, R. Cyganiak, A. Polleres, and S. Decker. An empirical survey of Linked Data conformance. *Journal of Web Semantics*, 2012.

[35] A. Isaac and E. Summers. SKOS Simple Knowledge Organization System Primer. W3C Working Group Note, 2009.

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

[37] C. Keet, M. del Carmen Suárez-Figueroa, and M. Poveda-Villalón. The Current Landscape of Pitfalls in Ontologies. In *International Conference on Knowledge Engineering and Ontology Development (KEOD)*, 2013.

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

[39] D. Kontokostas, P. Westphal, S. Auer, S. Hellmann, J. Lehmann, R. Cornelissen, and A. Zaveri. Test-driven Evaluation of Linked Data Quality. In *23rdInternational Conference on World Wide Web (WWW'14)*, 2014.

[40] 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.

[41] T. Lebo, S. Sahoo, and D. McGuinness. PROV-O: The PROV Ontology. W3C Recommendation, 2013. `http://www.`

w3.org/TR/prov-o.
[42] J. Lehmann and S. Sonnenburg. DL-Learner: Learning Concepts in Description Logics. *Journal of Machine Learning Research*, 2009.
[43] J. J. M. and A. B. Godfrey. *Juran's quality handbook*. McGraw Hill, 1999.
[44] C. Mader, B. Haslhofer, and A. Isaac. Finding quality issues in SKOS vocabularies. *Theory and Practice of Digital Libraries*, 2012.
[45] E. Mäkelä. Aether - Generating and Viewing Extended VoID Statistical Descriptions of RDF Datasets. In *11th European Semantic Web Conference (ESWC), Demo Track*, Heraklion, Greece, 2014.
[46] 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.
[47] P. Mendes, H. Mühleisen, and C. Bizer. Sieve: linked data quality assessment and fusion. 2012.
[48] P. N. Mendes, M. Jakob, A. García-Silva, and C. Bizer. DBpedia Spotlight: Shedding Light on the Web of Documents. In *7th International Conference on Semantic Systems*, 2011.
[49] A. Miles and S. Bechhofer. SKOS Simple Knowledge Organization System Reference. W3C Recommendation, 2009. http://www.w3.org/TR/skos-reference/.
[50] T. D. Noia, R. Mirizzi, V. O. Claudio, D. Romito, and M. Zanker. Linked Open Data to Support Content-based Recommender Systems. In *8th International Conference on Semantic Systems - I-SEMANTICS '12*, 2012.
[51] L. Page, S. Brin, M. Rajeev, and W. Terry. The PageRank Citation Ranking: Bringing Order to the Web. Technical report, 1998.
[52] M. Poveda-Villalón, M. Suárez-Figueroa, and A. Gmez-Pérez. Validating Ontologies with OOPS! In *18th International Conference on Knowledge Engineering and Knowledge Management (EKAW)*, 2012.
[53] E. Ruckhaus, O. Baldizan, and M.-E. Vidal. Analyzing Linked Data Quality with LiQuate. In *11th European Semantic Web Conference (ESWC)*, 2014.
[54] E. Sirin, M. Smith, and E. Wallace. Opening, Closing Worlds - On Integrity Constraints. In *5th OWLED Workshop on OWL: Experiences and Directions*, 2008.
[55] D. Soergel. Thesauri and ontologies in digital libraries. In *2nd ACM/IEEE-CS Joint Conference on Digital Libraries*, 2002.
[56] F. Suchanek, G. Kasneci, and G. Weikum. Yago: A Core of Semantic Knowledge. In *16th International World Wide Web Conference (WWW)*, 2007.
[57] 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.
[58] O. Suominen and C. Mader. Assessing and Improving the Quality of SKOS Vocabularies. *Journal on Data Semantics*, 2013.
[59] J. Tao, L. Ding, and D. McGuinness. Instance Data Evaluation for Semantic Web-Based Knowledge Management Systems. In *42nd Hawaii International Conference on System Sciences, HICSS'09*, pages 1–10, 2009.
[60] N. Toupikov, J. Umbrich, and R. Delbru. DING! Dataset ranking using formal descriptions. In *2nd International Workshop on Linked Data on the Web (LDOW)*, 2009.
[61] G. Tummarello, R. Delbru, and E. Oren. Sindice.com: Weaving the open linked data. In *6th International Semantic Web Conference (ISWC)*, 2007.
[62] M. Verlic. LODGrefine - LOD-enabled Google Refine in Action. In *8th International Conference on Semantic Systems - I-SEMANTICS '12*, 2012.
[63] R. Y. Wang and D. M. Strong. Beyond accuracy: What data quality means to data consumers. *Journal of Management Information Systems*, 1996.
[64] A. Zaveri, A. Rula, A. Maurino, R. Pietrobon, J. Lehmann, and S. Auer. Quality Assessment Methodologies for Linked Open Data. *Semantic Web Journal*, 2012.