Undefined 1 (2012) 1–5
IOS Press

Quality Assessment for Linked Open Data: A Survey

A Systematic Literature Review and Conceptual Framework

Amrapali Zaveri** [a,*], Anisa Rula** [b], Andrea Maurino [b], Ricardo Pietrobon [c], Jens Lehmann [a] and Sören Auer [d]

[a] *Universität Leipzig, Institut für Informatik, D-04103 Leipzig, Germany,*
E-mail: (zaveri, lehmann)@informatik.uni-leipzig.de
[b] *University of Milano-Bicocca, Department of Computer Science, Systems and Communication (DISCo), Innovative Techonologies for Interaction and Services (Lab), Viale Sarca 336, Milan, Italy*
E-mail: (anisa.rula, maurino)@disco.unimib.it
[c] *Associate Professor and Vice Chair of Surgery, Duke University, Durham, NC, USA.,*
E-mail: rpietro@duke.edu
[d] *University of Bonn, Computer Science Department, Enterprise Information Systems and Fraunhofer IAIS,*
E-mail: auer@cs.uni-bonn.de

Abstract. The development and standardization of semantic web technologies has resulted in an unprecedented volume of data being published on the Web as Linking Open Data (LOD). However, we observe widely varying data quality ranging from extensively curated datasets to crowdsourced and extracted data of relatively low quality. Data quality is commonly conceived as *fitness for use*. In this article, we present the results of a systematic review of approaches for assessing the quality of LOD. We gather existing approaches and compare and group them under a common classification scheme. In particular, we unify and formalize commonly used terminologies across papers related to data quality and provide a comprehensive list of the dimensions and metrics. Additionally, we qualitatively analyze the approaches and tools using a set of attributes. The aim of this article is to provide researchers and data curators a comprehensive understanding of existing work, thereby encouraging further experimentation and development of new approaches focused toward s data quality, specifically for LOD.

Keywords: data quality, assessment, survey, Linked Open Data

1. Introduction

The development and standardization of semantic web technologies has resulted in an unprecedented volume of data being published on the Web as *Linking Open Data* (LOD). This emerging Web of Data comprises close to 31 billion facts represented as Resource Description Framework (RDF) triples (as of 2011[1]). Although gathering and publishing such massive amounts of data is certainly a step in the right direction, data is only as useful as its quality. Datasets published on the Data Web already cover a *diverse set of domains* such as media, geography, life sciences, government etc.[2]. However, data on the Web reveals a large *variation in data quality*. For example, data extracted from semi-structured or even unstructured sources, such as DBpedia [53,40], often contains inconsistencies as well as misrepresented and incomplete information.

[*] **These authors contributed equally to this work.
[1] http://lod-cloud.net/

[2] http://lod-cloud.net/versions/2011-09-19/lod-cloud_colored.html

Data quality is commonly conceived as *fitness for use* [31,36,67] for a certain application or use case. Even datasets with quality problems might be useful for certain applications, as long as the quality is in the required range. For example, in the case of DBpedia the data quality is perfectly sufficient for enriching Web search with facts or suggestions about common sense information, such as entertainment topics. In such a scenario, DBpedia can be used to show related movies and personal information, when a user searches for an actor. In this case, it is rather neglectable, when in relatively few cases, a related movie or some personal fact is missing. For developing a medical application, on the other hand, the quality of DBpedia is probably insufficient, as shown in [69] , since data is extracted via crowdsourcing of a semi-structured source. It should be noted that even the traditional, document-oriented Web has content of varying quality and is still perceived to be extremely useful by most people. Consequently, a key challenge is to determine the quality of datasets published on the Web and make this quality information explicit. Assuring data quality is particularly a challenge in LOD as it involves a set of autonomously evolving data sources. Other than on the document Web, where information quality can be only indirectly (e.g. via page rank) or vaguely defined, there are much more concrete and measurable data quality metrics available for structured information. Such data quality metrics include correctness of facts, adequacy of semantic representation or degree of coverage.

There are already many methodologies and frameworks available for assessing data quality, all addressing different aspects of this task by proposing appropriate methodologies, measures and tools. In particular, the database community has developed a number of approaches [3,38,56,66]. However, quality on the Web of Data also includes a number of novel aspects, such as coherence via links to external datasets, data representation quality or consistency with regard to implicit information. Furthermore, inference mechanisms for knowledge representation formalisms on the web, such as OWL, usually follow an open world assumption, whereas databases usually adopt closed world semantics. Additionally, there are efforts focused towards evaluating the quality of an ontology either in the form of user reviews of an ontology, which are ranked based on inter-user trust [44] or (semi-) automatic frameworks [64]. However, in this article we focus towards quality assessment of instance data. Despite the quality in LOD being an essential concept, few efforts are currently in place to standardize how quality tracking and assurance should be implemented. Moreover, there is no consensus on how the data quality dimensions and metrics should be defined. Furthermore, the Web of Data presents new challenges that were not handled before in other research areas. Thus, adopting the existing approaches for assessing the quality in the Web of Data is not a straightforward problem. These challenges are related to the openness of the Web of Data, the diversity of the information and unbound, dynamic set of autonomous data sources and publishers. Providing semantic links is another new aspect that requires an initial exploration and understanding of the data. Therefore, in this paper, we present the findings of a systematic review of existing approaches that focus towards assessing the quality of Linked Open Data. After performing an exhaustive survey and filtering articles based on their titles, we retrieved a corpus of 118 relevant articles published between 2002 and 2012. Further analyzing these 118 retrieved articles, a total of 21 papers were found to be relevant for our survey and form the core of this paper. These 21 approaches are compared in detail and unified with respect to:

- commonly used terminologies related to data quality,
- 23 different dimensions and their formalized definitions,
- metrics for each of the dimensions along with a distinction between them being subjective or objective and
- comparison of tools used to assess data quality.

Our goal is to provide researchers, data consumers and those implementing data quality protocols with a comprehensive understanding of the existing work, thereby encouraging further experimentation and new approaches.

This paper is structured as follows: In Section 2, we describe the survey methodology used to conduct the systematic review. In Section 3, we unify and formalize (a) the terminologies related to data quality and in Section 4 we provide (b) definitions for each of the data quality dimensions and (c) metrics for each of the dimensions. In Section 5, we compare the selected approaches based on different perspectives such as, (a) dimensions, (b) metrics, (c) type of data and also distinguish the proposed tools based on a set of attributes. In Section 6, we conclude with ideas for future work.

2. Survey Methodology

This systematic review was conducted by two reviewers from different institutions (the two first authors of this article) following the systematic review procedures described in [35,52]. A systematic review can be conducted for several reasons [35] such as: (a) the summarization and comparison, in terms of advantages and disadvantages, of various approaches in a field; (b) the identification of open problems; (c) the contribution of a joint conceptualization comprising the various approaches developed in a field; or (d) the synthesis of a new idea to cover the emphasized problems. This systematic review tackles, in particular, the problems (a)−(c), in that, it summarizes and compares various data quality assessment methodologies as well as identifies open problems related to LOD. Moreover, a conceptualization of the data quality assessment field is proposed. An overview of our search methodology including the number of retrieved articles at each step is shown in Figure 1 and described in detail below.

Related surveys. In order to justify the need of conducting a systematic review, we first conducted a search for related surveys and literature reviews. We came across a study [36] conducted in 2005, which summarizes 12 widely accepted information quality frameworks applied on the World Wide Web. The study compares the frameworks and identifies 20 dimensions common between them. Additionally, there is a comprehensive review [3], which surveys 13 methodologies for assessing the data quality of datasets available on the Web, in structured or semi-structured formats. Our survey is different since it focuses only on structured data and on approaches that aim at assessing the quality of LOD. Additionally, the prior review (i.e. [36]) only focused on the data quality dimensions identified in the constituent approaches. In our survey, we not only identify existing dimensions but also introduce new dimensions relevant for assessing the quality of LOD. Furthermore, we describe quality assessment metrics corresponding to each of the dimensions and also identify whether they are objectively or subjectively measured.

Research question. The goal of this review is to analyze existing methodologies for assessing the quality of structured data, with particular interest in LOD. To achieve this goal, we aim to answer the following general research question:

How can one assess the quality of Linked Open Data employing a conceptual framework integrating prior approaches?

We can divide this general research question into further sub-questions such as:

- *What are the data quality problems that each approach assesses?*
- *Which are the data quality dimensions and metrics supported by the proposed approaches?*
- *What kind of tools are available for data quality assessment?*

Eligibility criteria. As a result of a discussion between the two reviewers a list of eligibility criteria was obtained as listed below. The articles had to satisfy the first criterion and one of the other four criteria to be included in our study.

- Inclusion criteria:
  - Studies published in English between 2002 and 2012.
  - Studies focused on data quality assessment for LOD
  - Studies focused on provenance assessment of LOD
  - Studies that proposed and/or implemented an approach for data quality assessment
  - Studies that assessed the quality of LOD or information systems based on LOD principles and reported issues
- Exclusion criteria:
  - Studies that were not peer-reviewed or published
  - Assessment methodologies that were published as a poster abstract
  - Studies that focused on data quality management
  - Studies that neither focused on LOD nor on other forms of structured data
  - Studies that did not propose any methodology or framework for the assessment of quality in LOD

Search strategy. Search strategies in a systematic review are usually iterative and are run separately by both members to avoid bias and ensure complete coverage of all related articles. Based on the research question and the eligibility criteria, each reviewer identified several terms that were most appropriate for this systematic review, such as: *data*, *quality*, *data quality*, *assessment*, *evaluation*, *methodology*, *improvement*, or *linked data*, which were used as follows:

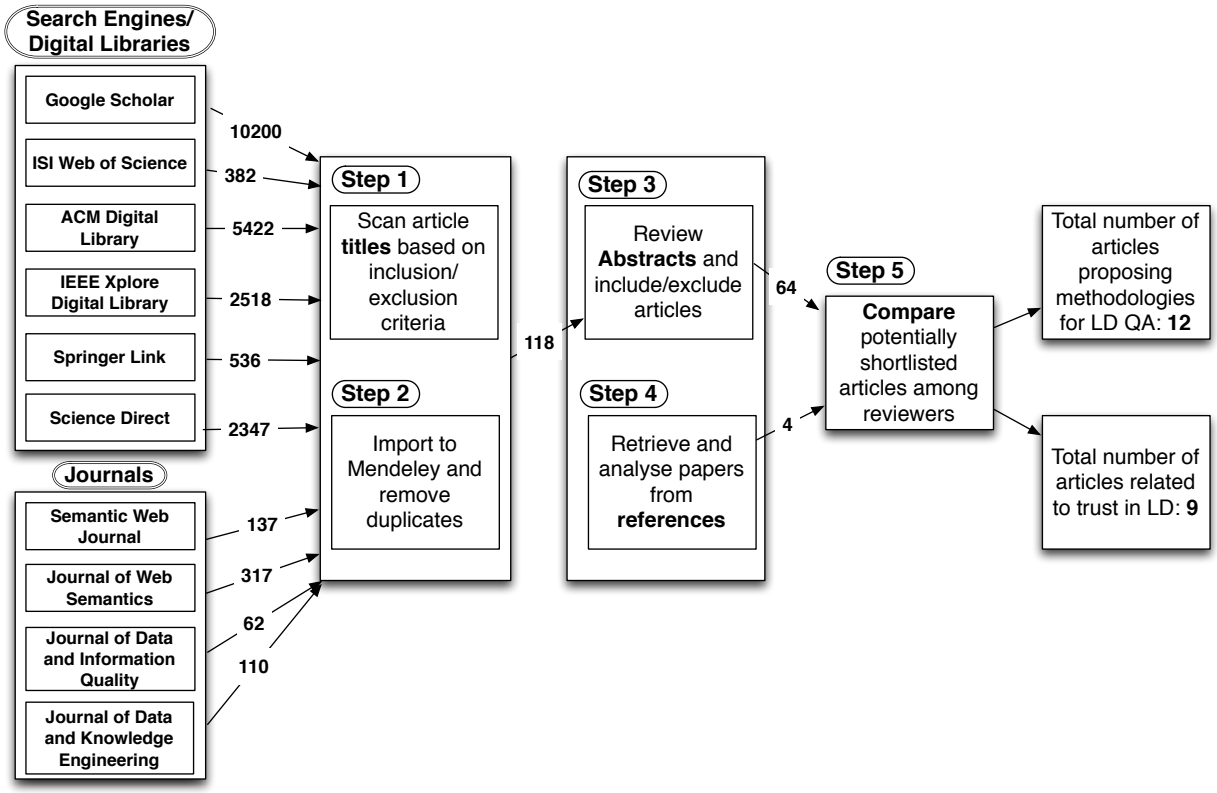

Fig. 1. Number of articles retrieved during literature search.

– *linked data* and (*quality* OR *assessment* OR *evaluation* OR *methodology* OR *improvement*)
– (*data* OR *quality* OR *data quality*) AND (*assessment* OR *evaluation* OR *methodology* OR *improvement*)

As suggested in [35,52], searching in the *title* alone does not always provide us with all relevant publications. Thus, the *abstract* or *full-text* of publications should also potentially be included. On the other hand, since the search on the full-text of studies results in many irrelevant publications, we chose to apply the search query first on the *title* and *abstract* of the studies. This means a study is selected as a candidate study if its *title* or *abstract* contains the keywords defined in the search string.

After we defined the search strategy, we applied the keyword search in the following list of search engines, digital libraries, journals, conferences and their respective workshops:
Search Engines and digital libraries:

– Google Scholar
– ISI Web of Science
– ACM Digital Library
– IEEE Xplore Digital Library
– Springer Link
– Science Direct

Journals:

– Semantic Web Journal
– Journal of Web Semantics
– Journal of Data and Information Quality
– Journal of Data and Knowledge Engineering

Conferences and their Respective Workshops:

– International Semantic Web Conference (ISWC)
– European Semantic Web Conference (ESWC)
– Asian Semantic Web Conference (ASWC)
– International World Wide Web Conference (WWW)
– Semantic Web in Provenance Management (SWPM)
– Consuming Linked Data (COLD)
– Linked Data on the Web (LDOW)

– Web Quality

Thereafter, the bibliographic metadata about the 118 potentially relevant primary studies were recorded using the bibliography management platform Mendeley[3].

Titles and abstract reviewing. Both reviewers independently screened the titles and abstracts of the retrieved 118 articles to identify the potentially eligible articles. In case of disagreement while merging the lists, the problem was resolved either by mutual consensus or by creating a list of articles to go under a more detailed review. Then, both the reviewers compared the articles and based on mutual agreement obtained a final list of 64 articles to be included.

Retrieving further potential articles. In order to ensure that all relevant articles were included, an additional strategy was applied such as:

– Looking up the references in the selected articles
– Looking up the article title in Google Scholar and retrieving the "Cited By" articles to check against the eligibility criteria
– Taking each data quality dimension individually and performing a related article search

After performing these search strategies, we retrieved 4 additional articles that matched the eligibility criteria.

Extracting data for quantitative and qualitative analysis. As a result of the search, we retrieved 21 articles from 2002 to 2012 listed in Table 1, which are the core of our survey. Of these 21, 12 propose generalized methodologies and 9 articles focus towards trust related quality assessment .

Comparison perspective of selected approaches. There exist several perspectives that can be used to analyze and compare the selected approaches, such as:

– the definitions of the core concepts
– the dimensions and metrics proposed by each approach
– the type of data that is considered for the assessment
– the comparison of the tools based on several attributes

Selected approaches differ in how they consider all of these perspectives and are thus compared and described in Section 3 and Section 4.

[3]https://www.mendeley.com

Table 1
List of the selected papers.

Citation	**Title**
Gil et al., 2002 [18]	Trusting Information Sources One Citizen at a Time
Golbeck et al., 2003 [21]	Trust Networks on the Semantic Web
Mostafavi et al., 2004 [54]	An ontology-based method for quality assessment of spatial data bases
Golbeck, 2006 [20]	Using Trust and Provenance for Content Filtering on the Semantic Web
Gil et al., 2007 [17]	Towards content trust of web resources
Lei et al., 2007 [42]	A framework for evaluating semantic metadata
Hartig, 2008 [24]	Trustworthiness of Data on the Web
Bizer et al., 2009 [5]	Quality-driven information filtering using the WIQA policy framework
Böhm et al., 2010 [6]	Profiling linked open data with ProLOD
Chen et al., 2010 [10]	Hypothesis generation and data quality assessment through association mining
Flemming, 2010 [14]	Assessing the quality of a Linked Data source
Hogan et al.,2010 [26]	Weaving the Pedantic Web
Shekarpour et al., 2010 [62]	Modeling and evaluation of trust with an extension in semantic web
Fürber et al.,2011 [15]	SWIQA − a semantic web information quality assessment framework
Gamble et al., 2011 [16]	Quality, Trust, and Utility of Scientific Data on the Web: Towards a Joint Model
Jacobi et al., 2011 [29]	Rule-Based Trust Assessment on the Semantic Web
Bonatti et al., 2011 [7]	Robust and scalable linked data reasoning incorporating provenance and trust annotations
Guéret et al., 2012 [22]	Assessing Linked Data Mappings Using Network Measures
Hogan et al., 2012 [27]	An empirical survey of Linked Data conformance
Mendes et al., 2012 [50]	Sieve: Linked Data Quality Assessment and Fusion
Rula et al., 2012 [59]	Capturing the Age of Linked Open Data: Towards a Dataset-independent Framework

Quantitative overview. Out of the 21 selected approaches, only 5 (23%) were published in a journal, particularly only in the Journal of Web Semantics. On the other hand, 14 (66%) approaches were published international conferences or workshops such as WWW, ISWC and ICDE. Only 2 (11%) of the approaches were master thesis and/or PhD workshop pa-

pers. The majority of the papers was published evenly distributed between the years 2010 and 2012 (4 papers each year − 57%), 2 papers were published in 2009 (9.5%) and the remaining 7 between 2002 and 2008 (33.5%).

3. Conceptualization

There exist a number of discrepancies in the definition of many concepts in data quality due to the contextual nature of quality [3] . Therefore, we first describe and formally define the research context terminology (in this section) as well as the LOD quality dimensions (in Section 4) along with their respective metrics in detail.

Data Quality. Data quality is commonly conceived as a multidimensional construct with a popular definition as the "fitness for use" [31]. Data quality may depend on various factors such as accuracy, timeliness, completeness, relevancy, objectivity, believability, understandability, consistency, conciseness, availability, and verifiability [67].

In terms of the Semantic Web, there are varying concepts of data quality. The semantic metadata, for example, is an important concept to be considered when assessing the quality of datasets [41]. On the other hand, the notion of link quality is another important aspect in LOD that is introduced, where it is automatically detected whether a link is useful or not [22]. It is to be noted that *data* and *information* are interchangeably used in the literature.

Data Quality Problems. Bizer et al. [5] relates data quality problems to those arising in web-based information systems which integrate information from different providers. For Mendes et al. [50], the problem of data quality is related to values being in conflict between different data sources as a consequence of the diversity of the data.

Flemming [14], on the other hand, implicitly explains the data quality problems in terms of *data diversity*. Hogan et. al. [26,27] discuss about *errors*, *noise*, *difficulties* or *modelling issues* which are prone to the non-exploitations of those data from the applications.

Thus, the term “data quality problems” refers to a set of issues that can affect the potentiality of the applications that use the data.

Data Quality Dimensions and Metrics. Data quality assessment involves the measurement of quality *dimensions* or *criteria* that are relevant to the consumer. The dimensions can be considered as the characteristics of a dataset. A data quality assessment *metric* or *measure* is a procedure for measuring a data quality dimension [5]. These metrics are heuristics that are designed to fit a specific assessment situation [43]. Since the dimensions are rather abstract concepts, the assessment metrics rely on quality *indicators* that allow for the assessment of the quality of a data source w.r.t the criteria [14]. An assessment score is computed from these indicators using a scoring function.

Bizer et al. [5], classify the data quality dimensions into three categories according to the type of information that is used as quality indicator: (1) Content Based − information content itself; (2) Context Based − information about the context in which information was claimed; (3) Rating Based − based on the ratings about the data itself or the information provider. However, we identify further dimensions (defined in Section 4) and also further categories to classify the dimensions into (1) Accessibility (2) Intrinsic (3) Trust (4) Dataset dynamicity (5) Contextual and (6) Representational dimensions.

Data Quality Assessment Methodology. A data quality assessment methodology is defined as the process of evaluating if a piece of data meets the information consumers need in a specific use case [5]. The process involves measuring the quality dimensions that are relevant to the user and comparing the assessment results with the users quality requirements.

4. Linked Open Data quality dimensions

After analyzing the 21 selected approaches in detail, we identified a core set of 23 different data quality dimensions that can be applied to assess the quality of LOD. We group the identified dimensions according to the classification introduced in [67], which is further modified and extended as:

- Accessibility dimensions
- Intrinsic dimensions
- Trust dimensions
- Dataset dynamicity dimensions
- Contextual dimensions
- Representational dimensions

In this section, we unify, formalize and adapt the definition for each dimension to the LOD context. The metrics associated with each dimension are also identified and reported. The dimensions belonging to the same group share the characteristics of the group. These groups are not strictly disjoint but can partially

overlap since there exist trade-offs between the dimensions of each group as described in Section 4.7. Additionally, we provide a use case scenario and examples for each of the dimensions. In certain cases, the examples point towards the quality of the information systems such as search engines and in other cases, about the data itself.

Use case scenario. Since data quality is described as "fitness for use", we introduce a specific use case that will allow us to illustrate the importance of each dimension with the help of an example. Our use case is about an intelligent flight search engine, which relies on acquiring (aggregating) data from several datasets. It obtains information about airports and airlines from an airline dataset (e.g., *OurAirports*[4], *OpenFlights*[5]). Information about the location of countries, cities and particular addresses is obtained from a spatial dataset (e.g., *LinkedGeoData*[6]). Additionally, aggregators pull all the information related to flights from different booking services (e.g., *Expedia*[7]) and represent this information as RDF. This allows a user to query the integrated dataset for a flight between any start and end destination for any time period. We will use this scenario throughout this section as an example of how data quality influences *fitness for use*.

4.1. Accessibility dimensions

The dimensions belonging to this category involve aspects related to the access and retrieval of data to obtain either the entire or some portion of the data for a particular use case. There are five dimensions part of this group, which are *availability*, *licensing*, *interlinking*, *security* and *performance*. Table 2 displays metrics for these dimensions and provides references to the original literature.

4.1.1. Availability.

Bizer [4] adopted the definition of availability from Pipino et al. [56] as "the extent to which information is available, or easily and quickly retrievable". On the other hand, Flemming [14] referred to availability as the proper functioning of all access methods. In the definition by Pipino et al., availability is more related to the measurement of available information rather than to the method of accessing the information as implied in the latter explanation by Flemming.

Definition 1 (Availability)**.** *Availability of a dataset is the extent to which information (or some portion of it) is present, obtainable and ready for use.*

Metrics. Availability of a dataset is measured in terms of accessibility of the server, SPARQL[8] endpoints or RDF dumps and also by the dereferencability of the Uniform Resource Identifier (URI). Furthermore, availability is also measured by the presence of structured data. The accessibility of the content is measured as the availability of dereferenced back-links or forward-links. Additionally, availability can be subjectively detected by measuring whether the content is suitable for consumption and whether the content should be accessed.

Example. Let us consider the case in which the user looks up a flight in our flight search engine. However, instead of retrieving the results, she receives an error response code such as `4xx client error`. This is an indication that a requested resource is unavailable. In particular, when the returned error code is `404 Not Found` code, she may assume that either there is no information present at that specified URI or the information is unavailable. Naturally, an apparently unreliable search engine is less likely to be used, in which case the user may not book flights from this search engine after encountering such issues.

Execution of queries over the integrated knowledge base can sometimes lead to low availability due to several reasons such as network congestion, unavailability of servers, planned maintenance interruptions, dead links or dereferencability issues. Such problems affect the usability of a dataset and thus should be avoided by methods such as replicating servers or caching information.

4.1.2. Licensing.

Licensing is a new quality dimensions not considered for relational databases but mandatory in an open data world such as LOD. Flemming [14] and Hogan et al. [27] both stated that each RDF document should contain a license under which the content can be (re-)used, in order to enable information consumers to use the data under clear legal terms. Additionally, the existence of a machine-readable indication (by including the specifications in a VoID[9] description) as well as a human-readable indication of a license is also important. Although both these studies do not provide a for-

[4] `http://thedatahub.org/dataset/ourairports`
[5] `http://thedatahub.org/dataset/open-flights`
[6] `linkedgeodata.org`
[7] `http://www.expedia.com/`
[8] `http://www.w3.org/TR/rdf-sparql-query/`
[9] `http://vocab.deri.ie/void`

Table 2

Data quality metrics related to accessibility dimensions (type S refers to a subjective metric, O to an objective one).

Dimension	Metric	Description	Type
Availability	accessibility of the SPARQL endpoint and the server	checking whether the server responds to a SPARQL query [14, 26]	O
	accessibility of the RDF dumps	checking whether a RDF dump is provided and can be downloaded [14,26]	O
	dereferencability issues	when a URI returns an error (4xx client error/ 5xx server error) response code or detection of broken links [26]	O
	no structured data available	detection of dead links or detection of a URI without any supporting RDF metadata or no redirection using the status code 303 See Other or no code 200 OK [14,26]	O
	no dereferenced back-links	detection of all local in-links or back-links: locally available triples in which the resource URI appears as an object, in the dereferenced document returned for the given resource [27]	O
	no dereferenced forward-links	detection of all forward links: locally known triples where the local URI is mentioned in the subject [27]	O
	misreported content types	detection of whether the content is suitable for consumption, and whether the content should be accessed [26]	S
Licensing	machine-readable indication of a license	detection of the indication of a license in the VoID description or in the dataset itself [14,27]	O
	human-readable indication of a license	detection of a license in the documentation of the dataset or its source [14,27]	O
	permissions to use the dataset	detection of license indicating whether reproduction, distribution, modification or redistribution is permitted [14]	O
	indication of attribution, *Copyleft* or *ShareAlike*	detection of whether the work is attributed in the same way as specified by the author or licensor [14]	O
Interlinking	interlinking degree, clustering coefficient, centrality and sameAs chains, description richness through sameAs	by using network measures [22]	O
	existence of links to external data providers	detection of the existence and usage of external URIs and owl:sameAs links [27]	S
Security	access to data is secure	use of login credentials or use of SSL or SSH [67]	O
	data is of proprietary nature	data owner allows access only to certain users [67]	O
Performance	no usage of slash-URIs	checking for usage of slash-URIs where large amounts of data is provided [14]	O
	low latency	delay between submission of a request by the user and reception of the response from the system [14,4]	O
	high throughput	no. of answered HTTP-requests per second [14]	O
	scalability of a data source	detection of whether the time to answer an amount of ten requests divided by ten is not longer than the time it takes to answer one request [14]	O

mal definition, they agree on the use and importance of licensing in terms of data quality.

Definition 2 (Licensing)**.** *Licensing is defined as the granting of permission for a consumer to re-use a dataset under defined conditions.*

Metrics. Licensing is checked by the indication of machine and human readable information associated with the dataset clearly indicating the permissions of data re-use. The indication of attribution in terms of *Copyleft* or *ShareAlike* is also used to check the licensing information of a dataset.

Example. Since our flight search engine aggregates data from several existing data sources, a clear indication of the license allows the search engine to reuse the data from the airlines websites. For example, the LinkedGeoData dataset is licensed under the Open Database License[10], which allows others to copy, distribute and use the data and produce work from the data allowing modifications and transformations. Due to the presence of this specific license, the flight search

[10] http://opendatacommons.org/licenses/odbl/

engine is able to re-use this dataset to pull geo-spatial information and feed it to the search engine.

LOD aims to provide users the capability of aggregating data from several sources, therefore the indication of an explicit license or waiver statement is necessary for each data source. A dataset can choose a license depending on what permissions it wants to issue (e.g. restrictions, liability, responsibility). Possible permissions include the reproduction of data, the distribution of data, or the modification and redistribution of data [51]. Providing licensing information increases the usability of the dataset as the consumers or third parties are thus made aware of the legal rights and permissiveness under which the pertinent data are made available. The more permissions a source grants, the more possibilities a consumer has while (re-)using the data.

4.1.3. Interlinking.

Interlinking is a relevant dimension in LOD since it supports data integration and interoperability. The interlinking is provided by RDF triples that establish a link between the entity identified by the subject with the entity identified by the object. Through the typed RDF links, data items are effectively interlinked. The importance of interlinking, also know as 'mapping coherence' can be classified in one of the four scenarios: (a) Frameworks; (b) Terminological Reasoning; (c) Data Transformation; (d) Query Processing, as identified in [48]. However, the core articles in this survey do not contain a formal definition for interlinking but instead contain metrics on how to measure this dimension.

Definition 3 (Interlinking)**.** *Interlinking refers to the degree to which entities that represent the same concept are linked to each other, be it within or between two or more linked data sources.*

Metrics. Interlinking is measured by using network measures that calculate the interlinking degree, cluster coefficient, sameAs chains, centrality and description richness through sameAs links and also by detecting the existence and usage of external URIs and owl:sameAs links.

Example. In our flight search engine, the instance of the country `"United States"` in the airline dataset should be interlinked with the instance `"America"` in the spatial dataset. This interlinking can help when a user queries for a flight, as the search engine can display the correct route from the start destination to the end destination by correctly combining information for the same country from both the datasets. Since names of various entities can have different URIs in different datasets, their interlinking can help in disambiguation.

Interlinking is included in this group of dimensions related to Accessibility because particularly in LOD, only with the help of the interlinks between entities, users or software agents are able to navigate between data items and *access* other datasets similar to what web crawlers do for web pages. Therefore the interlinking degree can also be used to measure the likelihood that a user or software agent browsing the LOD cloud will find a given dataset. In the Web of Data, it is common to use different URIs to identify the same real-world object occurring in two different datasets. Therefore, it is the aim of LOD to link or relate these two objects in order to be unambiguous. Moreover, not only the creation of precise links but also the maintenance of these interlinks is important. An aspect to be considered while interlinking data is to use different URIs to identify the real-world object and the document that describes it. The ability to distinguish the two through the use of different URIs is critical to the interlinking coherence of the Web of Data [25]. Moreover, the correct usage of the property (e.g. `owl:sameAs, skos:related skos:broader` etc.) is important to ensure proper representation of the type of relationship between the entities [23]. An effort towards assessing the quality of a mapping (i.e. incoherent mappings), even if no reference mapping is available, is provided in [48].

4.1.4. Security.

Flemming [14] referred to security as "the possibility to restrict access to the data and to guarantee the confidentiality of the communication between a source and its consumers". However, this dimension was not included in the tool developed by Flemming because although the author considered it as an important measure, when medical or governmental datasets are concerned, the security dimension is rarely applied in LOD. This is the only article that describes this dimension (from the core set of articles included in this survey).

Definition 4 (Security)**.** *Security is the extent to which access to data can be restricted and hence protected against its illegal alteration and misuse.*

Metrics. Security is measured based on whether the data has a proprietor or requires web security techniques (e.g. SSL or SSH) for users to access, acquire

or re-use the data. The importance of security depends on whether the data needs to be protected and whether there is a cost of data becoming unintentionally available. For open data the protection aspect of security can be often neglected but the non-repudiation of the data is still an important issue. Digital signatures based on private-public key infrastructures can be employed to guarantee the authenticity of the data.

Example: In our scenario, consider a user who wants to book a flight from a city `A` to a city `B`. The search engine should ensure a secure environment to the user during the payment transaction since her personal data is highly sensitive. The data about the payment should be passed privately from the user to the data source. If there is enough identifiable public information of the user, then she can be potentially targeted by private businesses, insurance companies etc. which she is unlikely to want. Thus, the use of SSL can be used to keep the information safe.

Security covers technical aspects of the accessibility of a dataset, such as secure login and the authentication of an information source by a trusted organization. The use of secure login credentials or access via SSH or SSL is used as a mean to protect data, especially in cases of biomedical or sensitive governmental data. Additionally, adequate protection of a dataset is an important aspect to be considered against its alteration or misuse and, therefore, a reliable and secure infrastructure or methodologies should be used [63]. For example, SHI3LD, an access control framework for RDF stores is used to secure content on the Web of Data [11].

4.1.5. Performance.

Performance is a dimension that has an influence on the quality of the data source, however not on the data set itself. Low performance reduces availability and usability of the data, in particular open data, which users may not want to manage on their own machines. Performance depends on several factors such as network traffic, server workload, server capabilities and/or complexity of the user query, which affect the quality of query processing. Flemming [14] denoted performance as a quality indicator which "comprises aspects of enhancing the performance of a source as well as measuring of the actual values (of the performance)". On the other hand, Hogan et al. [27] associated performance to the issue of "using prolix RDF features" such as (i) reification, (ii) containers and (iii) collections. These features should be avoided as they are cumbersome to represent in triples and can prove to be expensive to support in performance or data intensive environments. Flemming [14] gave a general description of performance without explaining the meaning while Hogan et al. [27] described the issues related to performance. Moreover, Bizer [4], defined response-time as "the delay between submission of a request by the user and reception of the response from the system". Thus, response-time and performance point towards the same quality dimension.

Definition 5 (Performance)**.** *Performance refers to the efficiency of a system that binds to a large dataset, that is, the more performant a data source the more efficiently a system can process data.*

Metrics. Performance is measured based on the scalability of the data source as well as the delay between submission of a request by the user and reception of the response from the dataset. Additional metrics are detection of slash-URIs, low latency[11] and high throughput of the services provided for the dataset.

Example. In our use case, the performance may depend on the type and complexity of the query by a large number of users. Our flight search engine can perform well by considering response-time when deciding which sources to use to answer a query.

Achieving low latency and high performance should be the aim of a dataset service. The performance of a dataset can be improved by (i) providing the dataset additionally as an RDF dump (ii) usage of hash-URIs instead of slash-URIs, (iii) locally replicating or caching information, (iv) providing low latency, so that the user is provided with a part of the results early on. This dimension also depends on the type and complexity of the request. Low response time hinders the usability as well as accessibility of a dataset. Since LOD usually involves the aggregation of several large datasets, they should be easily and quickly retrievable. Also, the performance should be maintained even while executing complex queries over large amounts of data to provide query repeatability, explorational fluidity as well as ready accessibility.

4.1.6. Intra-relations

The dimensions in this group are related with each other as follows: performance (response-time) of a system is related to all other availability dimensions. Only if a dataset is available and has low response time, it can perform well. Security is also related to the avail-

[11] Latency is the amount of time from issuing the query until the first information reaches the user [55].

ability of a dataset because the methods used for restricting users is tied to the way a user can access a dataset.

4.2. Intrinsic dimensions

Intrinsic dimensions are those that are independent of the user's context. There are three dimensions that are part of this group, which are *accuracy*, *consistency* and *conciseness*. These dimensions focus on whether information correctly and compactly represents the real world data and whether information is logically consistent in itself. Table 3 provides metrics for these dimensions along with references to the original literature.

4.2.1. Accuracy.

Bizer [4] adopted the definition of accuracy from Wang et al. [65] as the "degree of correctness and precision with which information in an information system represents states of the real world". Based on this definition, we also considered the problems of *spurious annotation* and *inaccurate annotation* (inaccurate labeling and inaccurate classification) identified in Lei et al. [42] related to the accuracy dimension. Furthermore, Furber et al. [15] classified accuracy into syntactic and semantic accuracy.

Definition 6 (Accuracy)**.** *Accuracy is defined as the extent to which data is correct, that is, the degree to which it correctly represents the real world facts and is also free of syntax errors. Accuracy is classified into (i) syntactic accuracy, which refers to the degree to which data values are close to its corresponding definition domain*[14] *and (ii) semantic accuracy, which refers to the degree to which data values represent the correctness of the values to the actual real world values.*

Metrics. Inaccurate data is measured as the closeness of a value in LOD to the corresponding value in a defined gold standard. This comparison is made possible by using a comparison function which evaluates the distance between the correct and the inaccurate values. The detection of inaccurate values can also be identified through the violation of functional dependency rules. Yet another method is by checking accuracy against several sources where a single fact is checked individually in different datasets to determine its accuracy or even several websites [39]. Inaccuracies are also detected by identifying literals incompatible with datatype range or literals which do not abide by the lexical syntax for their respective datatype. Accuracy of the annotation, representation, labelling or classification is detected as a value between 0 and 1. The last metric of accuracy (as shown in table 3) uses a balanced distance metric (an algorithm) that calculates the distance between the extracted (or learned) concept and the target concept [46].

Example. In our use case, let us assume that the ID of the flight between Paris and New York is A123. However, in our search engine the same flight instance is represent as A231. Since this ID is included in one of the datasets, it is considered to be syntactically accurate sinced it is a valid ID. On the other hand, the instance is semantically inaccurate since the flight ID does not represent its real-world state i.e. A123.

Accuracy is one of the dimensions, which is affected by assuming a closed or open world. When assuming an open world, it is more challenging to assess accuracy, since logical constraints need to be specified for inferring logical contradictions. In general, while the problems of inaccurate and spurious annotation are considered as semantic accuracy problems, the problems of illegal values or syntax violations are considered as syntactic accuracy problems.

4.2.2. Consistency.

Bizer [4] adopted the definition of consistency from Mecella et al., [47] as when "two or more values do not conflict with each other". Similarly, Hogan et al. [26] defined consistency as "no contradictions in the data". Another definition was given by Mendes et al. [50] where "a dataset is consistent if it is free of conflicting information".

Definition 7 (Consistency)**.** *Consistency means that a knowledge base is free of (logical/formal) contradictions with respect to particular knowledge representation and inference mechanisms.*

Metrics. In LOD, semantic knowledge representation techniques are employed, which come with certain inference and reasoning strategies for revealing implicit knowledge, which then might render a contradiction. Consistency is relative to a particular logic (set of inference rules) for identifying contradictions. A consequence of our definition of consistency is that a dataset can be consistent wrt. the RDF inference rules, but inconsistent when taking, e.g., the OWL2-

[12] predicates are often misused when no applicable predicate exists

[13] detection of an instance mapped back to more than one real world object leading to more than one interpretation

[14] A domain is a region characterized by a specific feature.

Table 3
Data quality metrics related to intrinsic dimensions (type S refers to a subjective metric, O to an objective one).

Dimension	Metric	Description	Type
Accuracy	detection of outliers	by using distance-based, deviations-based and distribution-based method [5]	O
	inaccurate values	by using functional dependencies rules between the values of two or more different predicates [15,68]	O
	inaccurate facts	a single fact is checked individually in different datasets [39]	O
	malformed datatype literals	detection of ill-typed literals which do not abide by the lexical syntax for their respective datatype [26]	O
	literals incompatible with datatype range	detection of a datatype clash that can then occur if the property is given a value (i) that is malformed, or (ii) that is a member of an incompatible datatype [26]	O
	erroneous annotation/representation erroneous	$1 - \frac{\text{erroneous instances}}{\text{total no. of instances}}$ [42]	O
	inaccurate annotation, labelling, classification	$1 - \frac{\text{inaccurate instances}}{\text{total no. of instances}} * \frac{\text{balanced distance metric}}{\text{total no. of instances}}$ [42]	O
Consistency	entities as members of disjoint classes	$\frac{\text{no. of entities described as members of disjoint classes}}{\text{total no. of entities described in the dataset}}$ [14,26]	O
	usage of homogeneous datatypes	$\frac{\text{no. of properties used with homogeneous units in the dataset}}{\text{total no. of properties used in the dataset}}$ [14]	O
	invalid usage of undefined classes and properties	detection of classes and properties used without any formal definition [26]	O
	misplaced classes or properties	using entailment rules that indicate the position of a term in a triple [26]	O
	misuse of owl:datatypeProperty or owl:objectProperty	by using weighting scheme that identifies that most usage is contrary to the vocabulary constraint [26]	O
	use of members of owl:DeprecatedClass or owl:-DeprecatedProperty	based on a manual mapping between deprecated terms and compatible term [26]	O
	provide a blacklist for void values	list all bogus owl:Inverse-FunctionalProperty values [26]	O
	ontology hijacking	detection of the redefinition by third parties of external classes/ properties such that reasoning over data using those external terms is affected [26]	O
	misuse of predicates[12]	profiling statistics support the detection of such discordant values or misused predicates and facilitate to find valid formats for specific predicates [6]	O
	ambiguous annotation	$1 - \frac{\text{no. of ambiguous instances}^{13}}{\text{no. of the instances contained in the semantic metadata set}}$ [42]	O
Conciseness	intensional conciseness	$\frac{\text{no. of unique attributes of a dataset}}{\text{total no. of attributes in a target schema}}$ [50]	O
	extensional conciseness	$\frac{\text{no. of unique objects of a dataset}}{\text{total number of objects representations in the dataset}}$ [50]	O
	duplicate instance	$1 - \frac{\text{total no. of instances that violate the uniqueness rule}}{\text{total no. of relevant instances}}$ [15,42]	O

EL profile[15] into account. For assessing consistency, an inference engine or a reasoner supports the respective expressivity of the underlying knowledge representation. The reasoner is employed to detect violations of entities defined as members of disjoint classes, usage of homogeneous datatypes, invalid usage of undefined classes and properties, those classes and properties used without any formal definition, usage of members of deprecated classes or properties, ambiguous annotations or ontology hijacking. Other techniques such as data profiling can detect misuse of predicates (predicates are often misused when no applicable predicate exists), misuse of owl:datatypeProperty or owl:objectProperty by using weighting scheme that identifies that most usage is contrary to the vocabulary constraint or provide a blacklist for void values (bogus owl:Inverse-FunctionalProperty values).

Example. Let us assume a user looking for flights between Paris and New York on the 21st of December, 2012. Her query returns the following results:

```
Flight From    To      Arrival Departure
A123 Paris NewYork     14:50   22:35
A123 Paris Singapore   14:50   22:35
```

The results show that the flight number `A123` has two different destinations at the same date and same time of arrival and departure, which is inconsistent with the ontology definition that one flight can only have one destination at a specific time and date. This contradiction arises due to inconsistency in data

[15] `http://www.w3.org/TR/owl2-profiles/`

representation, which is detected by using inference and reasoning.

In practice, RDF-Schema inference and reasoning with regard to the different OWL profiles is used to measure consistency in a dataset. For domain specific applications, consistency rules are defined, for example, according to the SWRL [28] or RIF standards [34] and processed using a rule engine.

4.2.3. Conciseness.

Mendes et al. [50] classified conciseness into schema and instance level conciseness. On the schema level (intensional), "a dataset is concise if it does not contain redundant attributes (two equivalent attributes with different names)". Thus, intensional conciseness measures the number of unique properties of a dataset in relation to the overall number of properties in a target schema. On the data (instance) level (extensional), "a dataset is concise if it does not contain redundant objects (two equivalent objects with different identifiers)". Thus, extensional conciseness measures the number of unique objects in relation to the overall number of objects in the dataset. This definition of conciseness (reported in [50]) is very similar to the definition of 'uniqueness' defined by Furber et al. [15] as the "degree to which data is free of redundancies, in breadth, depth and scope". This comparison shows that uniqueness and conciseness point to the same dimension.

Definition 8 (Conciseness)**.** *Conciseness refers to the redundancy of entities, be it at the schema or the data level. Conciseness is classified into (i) intensional conciseness (schema level) which refers to the case when the data does not contain redundant attributes and (ii) extensional conciseness (data level) which refers to the case when the data does not contain redundant objects.*

Metrics. As conciseness is classified in two categories, it is measured as the ratio between the number of unique attributes (properties) or unique objects (instances) compared to the overall number of attributes or objects respectively present in a dataset.

Example. In our flight search engine, an example of intensional conciseness would be a particular flight, say `A123`, being represented by two different properties in the same dataset, such as `http://flights.org/airlineID#A123` and `http://flights.org/name#A123`. This redundancy can ideally be solved by fusing the two and keeping only one unique identifier. On the other hand, an example of extensional conciseness is when both these different identifiers of the same flight have the same information associated with them in both the datasets, thus duplicating the information.

While integrating data from two different datasets, if both use the same schema or vocabulary to represent the data, then the intensional conciseness is high. On the other hand, if the integration leads to the duplication of values, that is the same information is stored in different ways, this leads to high extensional conciseness. This may lead to contradictory values and can be solved by fusing duplicate entries and merging common properties.

4.2.4. Intra-relations

The dimensions in this group are related to each other as follows: Data can be accurate by representing the real world state but still can be inconsistent. However, if we merge accurate datasets, we will most likely get less inconsistencies than merging inaccurate datasets.

4.3. Trust dimensions

The dimensions belonging to this group are those that focus on the perceived trustworthiness of the dataset. There are four dimensions that are part of this group, namely *reputation*, *believability*, *verifiability* and *objectivity*. Table 4 displays metrics for these dimensions along with references to the original literature.

4.3.1. Reputation.

Gil et al. [17] associated reputation of an entity or a dataset either as a result from direct experience or recommendations from others. They proposed the tracking of reputation either through a centralized authority or via decentralized voting. This is the only article that describes this dimension (from the core set of articles included in this survey).

Definition 9 (Reputation)**.** *Reputation is a judgment made by a user to determine the integrity of a data source.*

Metrics. Reputation is usually a score, for example, a real value between 0 (low) and 1 (high). There are different possibilities to determine reputation, which can be classified into manual or (semi-)automated approaches. The manual approach is via a survey in a community or by questioning other members who can help to determine the reputation of a source through explicit ratings about the data, data sources or data providers. The (semi-)automated approach uses exter-

Table 4
Data quality metrics related to the trust dimensions (type S refers to a subjective metric, O to an objective one).

Dimension	Metric	Description	Type
Reputation	reputation of the dataset	by assigning explicit ratings to the dataset (manual) and analyzing external links or page rank (semi-automated) [50]	S
Believability	meta-information about the identity of information provider	checking whether the provider/contributor is contained in a list of trusted providers [4]	O
	indication of metadata about a dataset (provenance information)	presence of the title, content and URI of the dataset [14]	O
	computing the trustworthiness of RDF statements	computing a trust value based on the provenance information which can be either unknown or a value in the interval [-1,1] where 1: absolute belief, -1: absolute disbelief and 0:lack of belief/disbelief [24]	O
	computing the trust of an entity	construction of decision networks informed by provenance graphs [16]	O
	accuracy of computing the trust between two entities	by using a combination of (1) a propagation algorithm which utilizes statistical techniques for computing trust values between 2 entities through a path and (2) an aggregation algorithm based on a weighting mechanism for calculating the aggregate value of trust over all paths [62]	O
	acquiring content trust from users	based on associations that transfer trust from entities to resources [17]	O
	assigning trust values to data/ sources/rules	use of trust ontologies that assign content-based or metadata-based trust values that can be transferred from known to unknown data [29]	O
	determining trust value for data	using annotations for data such as (i) blacklisting, (ii) authoritativeness and (iii) ranking and using reasoning to incorporate trust values to the data [7]	O
	computing personalized trust recommendations	using provenance of existing trust annotations in social networks [20]	S
	detection of reliability and credibility of a data source	use of trust annotations made by several individuals to derive an assessment of the sources' reliability and credibility [18]	S
	computing the trustworthiness of RDF statements	computing a trust value based on user-based ratings or opinion-based method [24]	S
	detect the reliability and credibility of the dataset publisher	indication of the level of trust for the publisher on a scale of $1-9$ [17,21]	S
Verifiability	authenticity of the dataset	verifying authenticity of the dataset based on a provenance vocabulary such as the author and his contributors, the publisher of the data and its sources if any [14]	O
	usage of digital signatures	by signing a document containing an RDF serialization or signing an RDF graph [9,14]	O
	correctness of the dataset	verifying correctness of the dataset with the help of unbiased trusted third party [4]	S
Objectivity	objectivity of the information	checking for bias or opinion expressed when a data provider interprets or analyzes facts [4]	S
	objectivity of the source	checking whether independent sources confirm a fact [4]	S
	no biased data provided by the publisher	checking whether the dataset is neutral or the publisher has a personal influence on the data provided [4]	S

nal links or page ranks to determine the reputation of a dataset.

Example. The provision of information on the reputation of data sources allows conflict resolution. For instance, several data sources report conflicting prices (or times) for a particular flight number. In that case, the search engine can decide to trust only the source with higher reputation. Older, long-established data sources typically have a higher reputation.

Reputation is a social notion of trust [19] where the data publisher should be identifiable for a certain (part of a) dataset. This dimension is mainly associated with a data publisher, a person, organization, group of people or community of practice rather than being a characteristic of a dataset. Trust is often represented in a web of trust, where nodes are entities and edges are the trust value based on a metric that reflects the reputation one entity assigns to another [17]. Based on the information presented to a user, she forms an opinion or makes a judgment about the reputation of the dataset or the publisher and the reliability of the statements. It should be noted that *credibility* can be used as a synonym for reputation.

4.3.2. Believability.

Bizer [4] adopted the definition of believability from Pipino et al. [56] as "the extent to which information is regarded as true and credible". Jacobi et al. [29] termed believability as "trustworthiness" and similar to Pipino et al., they referred to believability as a subjective measure of a user's belief that the data is "true". Thus, trustworthiness can be used as a synonym for believability.

Definition 10 (Believability)**.** *Believability is defined as the degree to which the information is accepted to be correct, true, real and credible.*

Metrics. Believability is measured by checking whether the contributor is contained in a list of trusted providers and by analyzing the metadata of the dataset. Computing trust values (i) based on provenance information, (ii) by constructing decision networks using the provenance information, (iii) using statistical techniques, (iv) based on associations that transfer trust from entities to resources, (v) using trust ontologies, (vi) using annotation for data and using reasoning to incorporate the trust values to the data, are other metrics that are used to measure the believability of a dataset. Also, by using provenance of existing trust annotations (i) in social networks, (ii) those made by several individuals (iii) on user-based ratings or opinions or (iv) on the level of trust for the dataset publisher, one can measure the believability of the dataset subjectively.

Example. In our flight search engine use case, if the flight information is provided by trusted and well-known flights companies such as Lufthansa, British Airways, etc. then the user believes the information provided by their websites. She does not need to assess their credibility since these are well-known international flight companies. On the other hand, if the user retrieves information about an airline previously unknown, she can decide whether to believe this information by checking whether the airline is well-known or if it is contained in a list of trusted providers. Moreover, she will need to check the source website from which this information was obtained.

This dimension involves the decision of which information to believe. Users can make these decisions based on factors such as the source, their prior knowledge about the subject, the reputation of the source and their prior experience [29]. Another method proposed by Tim Berners-Lee was that Web browsers should be enhanced with an "Oh, yeah?" button to support the user in assessing the reliability of data encountered on the web[16]. Pressing of such a button for any piece of data or an entire dataset would help contribute towards assessing the believability of the dataset.

4.3.3. Verifiability.

Bizer [4] adopted the definition of verifiability from Naumann et al. [55] as the "degree and ease with which the information can be checked for correctness". Similarly, Flemming [14] referred to the verifiability dimension as the means a consumer is provided with to examine the data for correctness. Without such means, the assurance of the correctness of the data would come from the consumer's trust in that source. It can be observed here that on the one hand Naumann et al. provided a formal definition whereas Flemming described the dimension by providing its advantages and metrics.

Definition 11 (Verifiability)**.** *Verifiability refers to the degree by which a data consumer can assess the correctness of a dataset.*

Metrics. Verifiability is measured (i) based on provenance information, (ii) by the presence of a digital signature or (iii) by an unbiased third party, if the dataset itself points to the source.

Example. In our use case, if we assume that the flight search engine crawls information from arbitrary

[16] `http://www.w3.org/DesignIssues/UI.html`

airline websites, which publish flight information according to a standard vocabulary, there is a risk for receiving incorrect information from malicious websites. For instance, if such a website publishes cheap flights just to attract a large number of visitors. In that case, the use of digital signatures for published RDF data allows to restrict crawling only to verified datasets.

This dimension allows data consumers to decide whether to accept the provided information. One means of verification in LOD is to provide basic provenance information along with the dataset, such as using existing vocabularies like SIOC, Dublin Core, Provenance Vocabulary, the OPMV[17] or the recently introduced PROV vocabulary[18]. Yet another mechanism is by the usage of digital signatures [9], whereby a source can sign either a document containing an RDF serialization or an RDF graph. Using a digital signature, the data source can vouch for all possible serializations that can result from the graph thus ensuring the user that the data she receives is in fact the data that the source has vouched for. Moreover, an unbiased third party can help verify the correctness of the dataset.

4.3.4. Objectivity.

Bizer [4] adopted the definition of objectivity from Pipino et al. [56] as "the extent to which information is unbiased, unprejudiced and impartial." This is the only article that describes this dimension (from the core set of articles included in this survey).

Definition 12 (Objectivity)**.** *Objectivity is defined as the degree to which the interpretation and usage of data is unbiased, unprejudiced and impartial.*

Metrics. Objectivity cannot be quantified (since it highly depends on the type of information) but is measured indirectly by checking for (i) bias or opinions expressed when a data provider interprets or analyzes facts, (ii) whether independent sources can confirm a single fact or (iii) by checking whether the dataset is neutral or the publisher has a personal influence on the data provided.

Example. In our use case, consider the reviews regarding the safety, comfort and prices available for each airline. It may happen that an airline belonging to a particular alliance is ranked higher than others when in reality it is not so. This could be an indication of a bias where the review is falsified due to the providers preference or intentions. This kind of bias or partiality affects the user as she might be provided with incorrect information from expensive flights or from malicious websites.

One of the possible ways to detect biased information is to compare the information with other datasets providing the same information. However, objectivity is measured only with factual data. This is done by checking a single fact individually in different datasets for confirmation [39]. However, the bias in information can lead to errors in judgment and decision making and should be avoided.

4.3.5. Intra-relations

The dimensions in this group are related as follows: Reputation affects believability but the vice-a-versa does not hold true. We consider a dataset with high reputation to be believable. For example a source with high reputation may provide some information that may be judged to be not believable. Although a resource is believable based on a high reputation score, it is not believable when the resource describes wrong information. For example the Frankfurt international airport outputs a wrong timetable about a flight, the user considers it non-trustworthy. Although some specific statements are considered to be incorrect, the source's reputation still holds. However, in the case where the airport outputs a wrong timetable more than a certain number of times, the reputation of the source will suffer.

While verifiability concerns the verifications of the correctness of a dataset, believability is about trust to a dataset without checking. Thus, verifiability is related to the believability dimension, but differs from it because even though verification can find whether information is correct or incorrect, belief is the degree to which a user thinks an information is correct. Verifiability is an important dimension when a dataset has low believability or reputation e.g., because it is new and/or does not provide enough information with which to make that judgement. Low reputation may have been established because the source is new and there is not enough information with which to make that judgment. Objectivity is also related to the verifiability dimension, that is, the more verifiable a source is, the more objective it is likely to be.

4.4. Dataset dynamicity dimensions

Important aspects of data dynamicity are its freshness over time, the frequency of change over time and

[17] http://open-biomed.sourceforge.net/opmv/ns.html

[18] http://www.w3.org/TR/prov-o/

its freshness over time for a specific task. These three aspects are captured by the dataset dynamicity dimensions: *currency*, *volatility*, and *timeliness*. As shown in Batini et al. [2], the definitions provided for currency and timeliness, in different articles, are interchangeably used. Although some articles such as Fürber et al. [15] do not distinguish between these two dimensions, we keep them separated and describe the differences in this section. Table 5 surveys metrics for these three dimensions and provides references to the original literature.

4.4.1. Currency.

Rula et al. [59] defined currency as "the age of a value, where the age of a value is computed as the difference between the current time (the observation time) and the time when the value was last modified", with value referring either to documents or to statements. *TimeCloseness*, referring also to currency, is defined by Mendes et al. [50] as "the distance between the input date from the provenance graph to the current date". Both definitions show different ways of measuring currency but none of them provides a definition for the currency dimension. Furthermore, Bizer [4] and Flemming [14] named currency as timeliness. Although they called this dimension timeliness, it refers to currency as can be seen from the definitions. Bizer [4] adopted the definition from Kahn et al. [33] as "the degree to which information is up-to-date" and Flemming [14] defined the timeliness dimension as "the currentness of the data provided by a source".

Definition 13 (Currency)**.** *Currency measures how promptly the data is updated.*

Metrics. The currency of a value is basically computed as the difference between the observation time and the time when the value was last modified. Thus, it can relies on two components: (i) the time when the data was last modified (represented in the data model) and (ii) the observation time. Besides the above components, other approaches consider the publishing time[19] as an additional component in the formula. Alternatively, currency is measured as is measured as the ratio between the outdated values (if the system is able to identify the outdated values) and all the values, known as the *age* of a value. To determine whether the information is out-dated, we need temporal metadata to be available and represented by one of the data models proposed in the literature [58].

Example. Consider a user checking the flight timetable for her flight from a city A to a city B. Suppose that the result is a list of triples comprising of the description of the resource A such as the connecting airports, the time of arrival, the terminal, the gate, etc. The currency of the above set of triples is calculated based on the last update time of the document containing the set of triples.

Let us consider the case when the user knows when a change occurred in real-world data. In this case, it is possible to use an alternative metric for currency that measures the speed with which the values in the information system are updated after the real-world values change.

4.4.2. Volatility.

There is no definition for the volatility dimension in LOD but a recent study provides a comprehensive analysis regarding the change frequency of data sources and resources [32]. Based on this work, we provide a definition which applies to LOD.

Definition 14 (Volatility)**.** *Volatility refers to the frequency with which data varies in time.*

Metrics. Volatility can be measured as the length of time during which data remains valid. Volatility needs two components: (i) the expiry time (the time when the data becomes invalid) and (ii) the input time (the time when the data was first issued). Both these components are combined to measure the distance between the expiry time and the input time of the published data to obtain the time interval in which the information remains valid. Additionally, volatility can be measured by the "changefrequency" in a Semantic Sitemap.

Example. Let us consider the aforementioned example we used for currency where a user is interested in checking her flight timetable from a city A to a city B. The timetable is considered to be volatile information as it changes frequently. For instance, it is estimated that the flight timetable is updated every 10 minutes i.e. data remains valid for 10 minutes. As we showed before in the example for currency, a user is not able to interpret the result as a good or bad indicator of freshness except the case when the either the user knows a prior the change frequency of a document or statement or the system explicitly states those information. Therefore, volatility is used to support currency by interpreting if the result returned is current or not. Considering a time validity interval of 10 minutes, the

[19] identifies the time when data is first published in LOD

Table 5
Data quality metrics related to dataset dynamicity dimensions (type S refers to a subjective metric, O to an objective one).

Dimension	Metric	Description	Type
Currency	currency of documents/statements	$1 - \frac{\text{observation time} - \text{last modified time}}{\text{observation time} - \text{publishing time}}$ [59]	O
	time since modification	observation time - last modified time [50]	O
	exclusion of outdated data	$1 - \frac{\text{outdated data}}{\text{total amount of data}}$ [14]	O
Volatility	frequency of change	refer to the `changefrequency` attribute in a Semantic Sitemap for value of the frequency or updates of a data source [14]	O
	time validity interval	expiry time − input time of the semantic web source [15]	O
Timeliness	timeliness between the semantic source web and original source	a positive difference between last modified time of the original source and last modified time of the semantic web source implies data source to be outdated [15]	O
	timeliness of the resource	a positive difference between current and expiry time of the resource implies data source to be outdated [15]	O
	timeliness between the ideal freshness and the data source freshness	$1 - \frac{\text{observation time} - \text{last modified time}}{\text{ideal freshness}}$ [49]	O

resource representing the timetable is considered current if it is last modified within the 10 minutes interval. Otherwise it can be concluded that the timetable information is outdated, and thus users can seek other sources which provide current flight information.

In general, studying the change frequency of data is relevant for understanding whether the last update is opportunely provided in time. Thus, volatility of data provides further assessment not only related to the freshness of the information but also to the validity.

4.4.3. Timeliness.

Gamble et al. [16] defined timeliness as "a measure of utility is a comparison of the date the annotation was updated with the consumer's requirement".

Definition 15. *Timeliness measures how up-to-data data is, relative to a specific task.*

Metrics. Timeliness is usually measured by combining the two dimensions: currency and volatility. The first metric of timeliness refers to the delay between a change of the real-world value and the resulting modification of the value in the data source. The second metric measures the difference between the observation time and the invalid time. An alternative metric measures how far the ideal freshness for a given task is from the data source freshness.

Example. Let us suppose the user is consulting the flight timetable and she is aware of the volatility of this type of information, which is determined as a time distance of 10 minutes. In terms of dataset dynamicity dimensions, the information related to the flight is recorded and reported every 7 minutes meaning that the information remains current within the time validity interval determined by the volatility dimension. Although the flight information is updated on time, the information received by the user and the information recorded by the system can be inconsistent. This shows that there is a need for another dimension in order to support such problems, that is timeliness.

Data should be recorded and reported as frequently as the source values change and thus never become outdated. However, this may not be necessary nor ideal for the user's purposes, let alone practical, feasible or cost-effective. Thus, timeliness is an important quality dimension, with its value determined by the user's judgement of whether information is recent enough, given the rate of change of the source value and the user's domain and purpose of interest.

4.4.4. Intra-relations

The dimensions in this group are related as follows: Although timeliness is part of the dataset dynamicity group, it can be also considered as part of intrinsic quality dimensions because it is independent of the users context. However, timeliness depends on both – the currency and volatility dimensions and furthermore requires to ensure that data is available before the planned usage time.

Currency depends not only on the last update of the information in the information system but also on the time data remains valid. This is captured by volatility aiming to measure how corresponding real-world values change. Thus, in order to have current data, the last update should be within the interval determined by volatility. While currency is crucial for data with high volatility, it is less important for data with low volatility.

4.5. Contextual dimensions

Contextual dimensions are those that highly depend on the context of the task at hand. There are three dimensions that are part of this group, namely *completeness*, *amount-of-data* and *relevancy*. These dimensions along with their corresponding metrics and references to the original literature are presented in Table 6.

4.5.1. Completeness.

Bizer [4] adopted the definition of completeness from Pipino et al. [56] as "the degree to which information is not missing". Furber et al. [15] further classified completeness into: (a) Schema completeness, which is the degree to which classes and properties are not missing in a schema; (b) Column completeness, which is a function of the missing property values for a specific property/column; and (c) Population completeness, which refers to the ratio between classes represented in an information system and the complete population. Mendes et al. [50] distinguish completeness on the schema and the data level. On the schema level, a dataset is complete if it contains all of the attributes needed for a given task. On the data (i.e. instance) level, a dataset is complete if it contains all of the necessary objects for a given task. As can be observed, Pipino et al. provided a general definition whereas Furber et al. provided a set of sub-categories for completeness. On the other hand, the two types of completeness defined in Mendes et al. can be mapped to the two categories (a) Schema completeness and (c) Population completeness provided by Furber et al.

Definition 16 (Completeness)**.** *Completeness refers to the degree to which all required information is present in a particular dataset. In terms of LD, completeness comprises the following aspects: (a) Schema completeness, the degree to which the classes and properties of an ontology are represented, thus can be called "ontology completeness", (b) Property completeness, measure of the missing values for a specific property, (c) Population completeness is the percentage of all real-world objects of a particular type that are represented in the datasets and (d) Interlinking completeness has to be considered especially in LOD and refers to the degree to which instances in the dataset are interlinked.*

Metrics. Completeness is measured by detecting the number of classes, properties, values and interlinks that are present in the dataset compared with an ideal (or gold standard) dataset. It should be noted, that in this case, users should assume a closed-world-assumption where a gold standard dataset is available and can be used to compare against the converted dataset.

Example. In our use case, the flight search engine contains complete information to include all the airport and airport codes such that it allows a user to find an optimal route from the start to the end destination (even in cases when there is no direct flight). For example, the user wants to travel from Santa Barbara to San Francisco. Since our flight search engine contains interlinks between these close airports, the user is able to locate a direct flight easily.

Particularly in LOD, completeness is of prime importance when integrating datasets from several sources where one of the goals is to increase completeness. The completeness of interlinks between datasets is also important so that one can retrieve all the relevant facts about a resource when querying the integrated data sources. However, measuring completeness of a dataset usually mandates the presence of a gold standard or the original data source to compare with.

4.5.2. Amount-of-data.

Bizer [4] adopted the definition for the amount-of-data dimension from Pipino et al. [56] as "the extent to which the volume of data is appropriate for the task at hand". Flemming [14] defined amount-of-data as the "criterion influencing the usability of a data source". Additionally, Chen et al. [10] stated that "the amount of data should be enough to approximate the 'true' scenario precisely". While Pipino et al. provided a formal definition, Flemming and Chen et al. explained the dimension by mentioning its advantages.

Definition 17 (Amount-of-data)**.** *Amount-of-data refers to the quantity and volume of data that is appropriate for a particular task.*

Metrics. The amount-of-data is measured in terms of triples, instances, and/or links present in the dataset. This amount should represent an appropriate volume of data for a particular task along with appropriate scope (no. of entities) and level of detail (no. of properties).

Example. In our use case, the flight search engine acquires enough amount of data so as to cover all, even small, airports. In addition, the data also covers alternative means of transportation (e.g. bus, taxi etc.). This helps to provide the user with better travel plans, which includes smaller cities (airports). For example, when a user wants to travel from Hartford to Santa Barbara, she does not find direct or indirect flights by searching

Table 6
Data quality metrics related to contextual dimensions (type S refers to a subjective metric, O to an objective one).

Dimension	Metric	Description	Type
Completeness	schema completeness	no. of classes and properties represented / total no. of classes and properties [4,15,50]	O
	property completeness	no. of values represented for a specific property / total no. of values for a specific property [4,15]	O
	population completeness	no. of real-world objects are represented / total no. of real-world objects [4,15,26,50]	O
	interlinking completeness	no. of instances in the dataset that are interlinked / total no. of instances in a dataset [22]	O
Amount-of-data	appropriate volume of data for a particular task	ratio of no. of semantically valid association rules to the no. of non-trivial rules[20] [10]	O
	appropriate amount of data	use of the apriori algorithm to detect poor predicates based on the occurrence dependencies among predicates [10]	O
	amount of triples	no. of triples present in a dataset [14]	O
	coverage	scope (no. of entities) and level of detail (no. of properties) [14]	O
Relevancy	usage of meta-information attributes	counting the occurrence of relevant terms within these attributes or using vector space model and assigning higher weight to terms that appear within the meta-information attributes [4]	S
	retrieval of relevant resources	sorting documents according to their relevancy for a given query [4]	S

individual flights websites. Flights are only suggested to the nearby bigger airports such as New York and San Francisco. But, using our example search engine, she is suggested convenient flight connections between the two destinations, because it contains an appropriate amount of data so as to cover all the airports. She is also offered convenient combinations of flights, trains and buses. The provision of such information also necessitates the presence of a sufficient amount of internal as well as externals links between the datasets so as to provide a fine grained search for connections between specific places.

An appropriate volume of data, in terms of quantity and coverage, should be a main aim of a dataset provider. However, a small amount of data, appropriate for a particular task, does not violate this definition. The aim should be to have sufficient breadth and depth as well as sufficient scope (number of entities) and detail (number of properties applied) in a given dataset.

4.5.3. Relevancy.

Bizer [4] adopted the definition of relevancy from Pipino et al. [56] as "the extent to which information is applicable and helpful for the task at hand". This is the only article that describes this dimension (from the core set of articles included in this survey).

Definition 18 (Relevancy)**.** *Relevancy refers to the provision of information which is in accordance with the task at hand and important to the users' query.*

Metrics. Relevancy is highly context dependent and is measured by using meta-information attributes for assessing whether the content is relevant for a particular task (use case). Additionally, retrieval of relevant resources can be performed using a combination of hyperlink analysis and information retrieval methods.

Example. When a user is looking for flights between any two cities, only relevant information i.e. start and end airports and times, duration and cost per person should be provided. Spatial datasets might, in addition to relevant information (e.g. airports), also contain much irrelevant data (e.g. post offices, lakes, trees etc. as present in LinkedGeoData) and as a consequence the user may get lost in the information. Instead, restricting a dataset to a domain of only flight related information which is smaller than the general Web domain, an application is more likely to return only relevant results.

In LOD, the external links or `owl:sameAs` links can help pull in additional relevant information about a resource. However, data polluted with irrelevant information affects the usability as well as typical query performance of the dataset.

4.5.4. Intra-relations

The three dimensions in this group are dependent on each other as follows: By establishing completeness (i.e. if the dataset is complete), amount-of-data becomes a function of (or depends on) relevancy. "Appropriate" amount-of-data means having relevant in-

formation for the task at hand. However, the term "appropriate" is difficult to be quantified, but we say that the amount of data is "appropriate" if we verify that the dataset meets the requirements of a user's query. On the other hand, if the amount-of-data is too large, known also as *information overload* [45], it decreases the effectiveness and efficiency of relevant data.

By fixing amount-of-data, completeness becomes a function of relevancy. In most cases, low completeness points towards low relevancy. However, data is 'complete' if it contains all the relevant information. By fixing relevancy, amount-of-data becomes a function of the completeness. The amount of data is often insufficient if a dataset is incomplete for a particular purpose.

4.6. Representational dimensions

Representational dimensions capture aspects related to the design of the data such as the *representational-conciseness*, *representational-consistency*, *understandability*, *interpretability* as well as the *versatility* of the data. Table 7 displays metrics for these dimensions along with references to the original literature.

4.6.1. Representational-conciseness.

Bizer [4], adopted the definition of representational-conciseness from Pipino et al. [56] as "the extent to which information is compactly represented". This is the only article that describes this dimension (from the core set of articles included in this survey).

Definition 19 (Representational-conciseness)**.** *Representational-conciseness refers to the representation of the data which is compact and well formatted on the one hand and clear and complete on the other hand.*

Metrics. Representational-conciseness is measured by detecting the use of short URIs and the absence of prolix RDF features in a dataset.

Example. Our flight search engine represents the URIs for the destination compactly with the use of the airport codes. For example, LEJ is the airport code for Leipzig, therefore the URI is `http://airlines.org/LEJ`. Similarly, CGN is the airport code for Cologne/Bonn and therefore the URI is compactly represented as `http://airlines.org/CGN`. This short representation of URIs helps users share and memorize them easily and also provides efficient processing of frequently used RDF data.

The binary RDF representation for publication and exchange[21] is a format for publishing and exchanging RDF data which consists of Header-Dictionary-Triples (HDT). This format is used to represent data *compactly* and on a large scale as it supports the splitting of huge RDF graphs into several parts. The concise representation of data not only contributes to the human readability of that data, but also influences the performance of data when queried. For example, Hogan et al. [27] associate the use of very long URIs (or those that contain query parameters) as an issue related to the representational conciseness of the data. Keeping URIs concise and human readable is highly recommended for large scale and/or frequent processing of RDF data as well as for efficient indexing and serialization.

4.6.2. Representational-consistency.

Bizer [4] adopted the definition of representational-consistency from Pipino et al. [56] as "the extent to which information is represented in the same format". The definition of the representational-consistency dimension is very similar to the definition of *uniformity* which refers to the re-use of established format to represent data as described by Flemming [14]. Additionally, as stated in Hogan et al. [27], the re-use of well-known terms to describe resources in a uniform manner increases the interoperability of data published in this manner and contributes towards representational consistency of the dataset.

Definition 20 (Representational-consistency)**.** *Representational-consistency is the degree to which the format and structure of the information conforms to previously returned information as well as data from other sources.*

Metrics. Representational consistency is assessed by detecting whether the dataset re-uses existing vocabularies and/or terms from existing established vocabularies to represent its entities.

Example. Let us consider different airline datasets using different notations for representing temporal data, e.g. one dataset uses the time ontology while another dataset uses XSD notations. This makes querying the integrated dataset difficult as it requires users to understand the various heterogeneous schema. Additionally, with the difference in the vocabularies used to represent the same concept (in this case time), the consumers are faced with problem of how the data can

[21] `http://www.w3.org/Submission/2011/SUBM-HDT-20110330/`

Table 7
Data quality metrics related to representational dimensions (type S refers to a subjective metric, O to an objective one).

Dimension	Metric	Description	Type
Representational-conciseness	keeping URIs short	detection of long URIs or those that contain query parameters [27]	O
	no use of prolix RDF features	detect use of RDF primitives i.e. RDF reification, RDF containers and RDF collections [27]	O
Representational-consistency	re-use existing terms	detect whether existing terms from other vocabularies have been reused [27]	O
	re-use existing vocabularies	usage of established vocabularies [14]	S
Understandability	human-readable labelling of classes, properties and entities	percentage of entities having an `rdfs:label` or `rdfs:comment` [14]	O
	dereferenced representations: providing human-readable metadata	detecting the use of `rdfs:label` to attach labels or names to resources [14]	O
	indication of one or more exemplary URIs	detecting whether the pattern of the URIs is provided [14]	O
	indication of a regular expression that matches the URIs of a dataset	detecting whether a regular expression that matches the URIs is present [14]	O
	indication of an exemplary SPARQL query	detecting whether examples of SPARQL queries are provided [14]	O
	indication of the vocabularies used in the dataset	checking whether a list of vocabularies used in the dataset is provided [14]	O
	provision of message boards and mailing lists	checking the effectiveness and the efficiency of the usage of the mailing list and/or the message boards [14]	O
Interpretability	use of self-descriptive formats	identifying objects and terms used to define these objects with globally unique identifiers [4]	O
	interpretability of terms	use of various schema languages to provide definitions for terms [4]	S
	interpretability of data	detect the use of appropriate language, symbols, units and clear definitions [4]	S
	misinterpretation of missing values	detecting use of blank nodes [27]	O
	atypical use of collections, containers and reification	detect non-standard usage of collections, containers and reification [26,27] (since these features are discouraged from use by Linked Data guidelines)	O
Versatility	provision of the data in different serialization formats	checking whether data is available in different serialization formats [14]	O
	provision of the data in various languages	checking whether data is available in different languages [1,14,37]	O
	accessing of data in different ways	checking whether the data is available as SPARQL endpoint and for download as RDF dump [14]	O
	application of content negotiation	checking whether data can be retrieved in accepted formats and languages by adding a corresponding accept-header to an HTTP request [14]	O

be interpreted and displayed. In order to avoid these interoperability issues, we provide data based on the Linked Data principles, which are designed to support heterogeneous description models necessary to handle different formats of data.

Re-use of well known vocabularies, rather than inventing new ones, not only ensures that the data is consistently represented in different datasets but also supports data integration and management tasks. In practice, for instance, when a data provider needs to describe information about people, FOAF[22] should be the vocabulary of choice. Moreover, re-using vocabularies maximizes the probability that data can be consumed by applications that may be tuned to well-known vocabularies, without requiring further pre-processing of the data or modification of the application. Suitable terms can be found in *Linked Open Vocabularies*[23], *SchemaWeb*[24], *SchemaCache*[25] and *Swoogle*[26]. Additionally, a comprehensive survey done in [61] lists a set of naming conventions that should be used to avoid

[22] `http://xmlns.com/foaf/spec/`

[23] `http://lov.okfn.org/dataset/lov/`
[24] `http://www.schemaweb.info/`
[25] `http://schemacache.com/`
[26] `http://swoogle.umbc.edu/`

inconsistencies[27]. Another possibility is to use *LODStats* [12], which allows to perform a search for frequently used properties and classes in the LOD cloud.

4.6.3. Understandability.

Bizer [4] adopted the definition of understandability from Pipino et al. [56] stating that "understandability is the extent to which data is easily comprehended by the information consumer". Flemming [14] related understandability also to the comprehensibility of data i.e. the ease with which human consumers can understand and utilize the data. Thus, comprehensibility can be interchangeably used with understandability.

Definition 21 (Understandability)**.** *Understandability refers to the ease with which data can be comprehended, without ambiguity, and used by a human information consumer.*

Metrics. Understandability is measured by detecting whether human-readable labels for classes, properties and entities are available in a dataset. Provision of human-readable metadata of a dataset also contribute towards improving its understandability. Additionally, the dataset should present exemplary URIs and SPARQL queries along with the vocabularies used so that can users can understand how it can be used. Moreover, presence of message boards and mailing lists helps users to develop an understanding of the dataset.

Example. Let us assume that our flight search engine allows a user to enter a start and end destination address. In that case, strings entered by the user need to be matched to entities in the spatial dataset, probably via string similarity. Understandable labels for cities, places etc. improve the search performance. For instance, when a user looks for a flight to "NYC" (label), then the search engine should return flights to New York City.

Understandability, in general, measures how well a source presents its data so that a user is able to understand its semantic value. In LOD, data publishers are encouraged to re-use well-known formats, vocabularies, identifiers, human-readable labels and descriptions of defined classes, properties and entities to ensure clarity and understandability of their data by the consumers.

[27] However, they only restrict themselves to only considering the needs of the OBO foundry community but still can be applied to other domains

4.6.4. Interpretability.

Bizer [4] adopted the definition of interpretability from Pipino et al. [56] as the "extent to which information is in appropriate languages, symbols and units, and the definitions are clear". This is the only article that describes this dimension (from the core set of articles included in this survey).

Definition 22 (Interpretability)**.** *Interpretability refers to technical aspects of the data, that is, whether information is represented using an appropriate notation and whether it conforms to the technical ability of the consumer.*

Metrics. Interpretability is measured a) by detecting the use of self-descriptive formats, b) by identifying objects and the terms used to describe these objects with globally unique identifiers and c) by employing various schema languages to provide definitions for terms. Additionally, avoiding blank nodes helps prevent misinterpretation because due to the ongoing debate on the semantics of blank nodes the blank node could either refer to a single unnamed entity or represent a existential quantification. Thus, the absence of blank nodes in a dataset can be used as an indicator for interpretability. Similarly, the atypical use of collections, containers and reification can help to measure the interpretability of the dataset.

Example. Consider our flight search engine and a user that is looking for a flight from `Milan` to `Boston`. Data related to Boston in the integrated data, for the required flight, contains the following entities:

- `http://rdf.freebase.com/ns/m.049jnng`
- `http://rdf.freebase.com/ns/m.043j22x`
- Boston Logan Airport

For the first two items no human-readable label is available, therefore the URI is displayed, which does not represent anything meaningful to the user besides that the information is from Freebase. The third entity, however, contains a human-readable label, which the user can easily interpret.

The more interpretable an LOD source is, the easier it is to integrate with other data sources. Also, interpretability contributes towards the usability of a data source. Use of existing, well-known terms, self-descriptive formats and globally unique identifiers increase the interpretability of a dataset.

4.6.5. Versatility.

Flemming [14] defined versatility as the "alternative representations of the data and its handling." This is the only article that describes this dimension (from the core set of articles included in this survey).

Definition 23 (Versatility)**.** *Versatility refers to the availability of the data in an internationalized way, the availability of alternative representations of data and the provision of alternative access methods for a dataset.*

Metrics. Versatility is measured by checking the availability of a dataset in different languages, the representation of the data in non-region specific ways (e.g. telephone numbers), the availability of different serialization formats as well as different access methods. Additionally, the application of content negotiation can help check whether the data can be retrieved in accepted formats and languages by adding a corresponding accept-header to an HTTP request.

Example. Consider a user who does not understand English but only Spanish and wants to use our flight search engine. In order to cater to the needs of such a user, either the original data sources should provide labels and other language-dependent information in Spanish (and possibly other languages) so that any user has the capability to understand it.

Provision of LOD in different languages contributes towards the versatility of the dataset with the use of language tags for literal values. Also, providing a SPARQL endpoint as well as an RDF dump as access points is an indication of the versatility of the dataset. Provision of resources in HTML format in addition to RDF, as suggested by the Linked Data principles, is also recommended to increase human readability. Similar to the uniformity dimension, versatility also enhances the probability of consumption and ease of processing of the data. In order to handle the versatile representations, content negotiation should be enabled whereby a consumer can specify accepted formats and languages by adding a corresponding accept header to an HTTP request.

4.6.6. Intra-relations

The dimensions in this group are related as follows: Understandability is related to the interpretability dimension as it refers to the capability of the information consumer to comprehend information. Interpretability mainly refers to the technical aspects of the data, that is if the data is represented using an appropriate notation. Also, interpretability is related to the representational-consistency of data since the consistent representation (e.g. re-use of established vocabularies) ensures that a system will be able to interpret the data correctly [13]. Versatility is also related to the interpretability of a dataset as the more versatile forms a dataset is represented in (e.g. in different languages), the more interpretable a dataset is.

4.7. Inter-relationships between dimensions

The 23 data quality dimensions explained in the previous section are not independent of each other but correlations exist among them. If one dimension is considered more important than the others for a specific application (or use case), then the choice of favoring it may imply negative consequences on the others. Investigating the relationships among dimensions is an interesting problem, as shown by the following examples of the possible interrelations between them. In this section, we describe the intra-relations between the 23 dimensions, as shown in Figure 2.

First, relationships exist between the dimensions belonging to the trust group and the accuracy and currency dimensions. When assessing the trust of a LOD dataset, it is not sufficient just to determine the reputation, believability and objectivity of the dataset. Also, accuracy and the currency of the dataset should be assessed. The accuracy and trust dimensions are often inter-related. Frequently the assumption is made, that a publisher with a high reputation will produce data that is also accurate and current. Furthermore, we can also state that data with a high believability is considered to be accurate and current. Moreover as shown in Figure 2, the dimensions in the trust group i.e. verifiability, believability and reputation are also included in the contextual dimensions group because they highly depend on the context of the task at hand.

Second, relationships occur between timeliness and the accuracy, completeness and consistency dimensions. Indeed, having accurate, complete or consistent data may require time and thus timeliness can be negatively affected. Conversely, having timely data may cause low accuracy, incompleteness and/or inconsistency. Most web applications prefer timeliness as opposed to accurate, complete or consistent data. As the time constraints are often very stringent for web available data, it may happen that such data are deficient with respect to other quality dimensions. For instance, a list of courses published on a university website must be timely, although there could be accuracy or consistency errors and some fields specifying courses

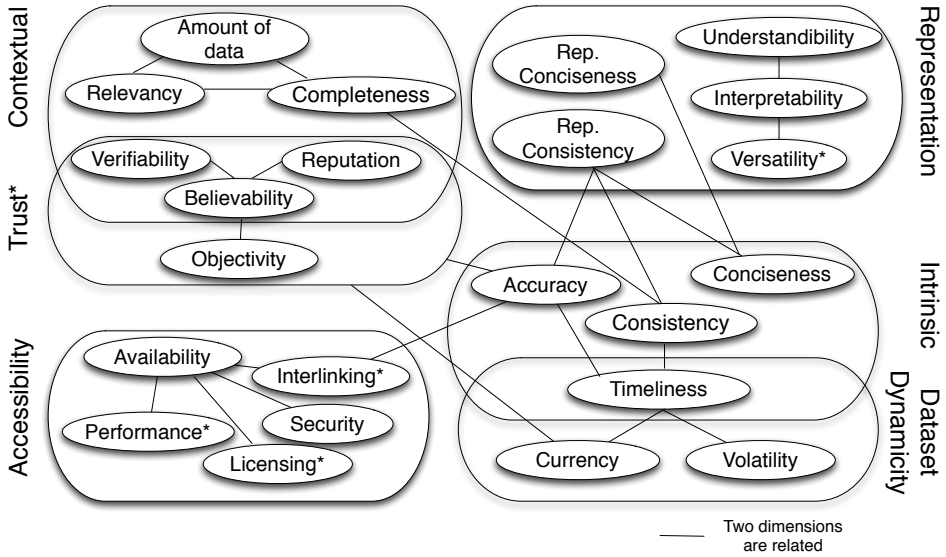

Fig. 2. Linked Open Data quality dimensions and the relations between them. The dimensions marked with '*' are specific for Linked Open Data.

could be missing. Conversely, when considering an e-banking application, accuracy, consistency and completeness requirements are more stringent than timeliness, and therefore delays are allowed in favor of correctness of data provided. Additionally, the timeliness dimension is also included in the intrinsic group because it does not depend on the user's context.

A further significant case of relationship exists between the consistency and completeness dimensions. Even though data is complete, it can be inconsistent. For example, statistical data analysis typically requires to have significant data in order to perform analysis and the approach is to favor completeness, tolerating inconsistencies, or adopting techniques to solve them. Conversely, if considering an application that calculates the salaries of a company's employees, it is more important to have a list of consistently checked salaries rather than a complete list.

The representational-conciseness dimension (belonging to the representational group) and the conciseness dimension (belonging to the intrinsic group) are also closely related with each other. On the one hand, representational-conciseness refers to the conciseness of representing the data (e.g. short URIs) while conciseness refers to the compactness of the data itself (redundant attributes and objects). Both dimensions thus point towards the compactness of the data. Additionally, the representational-consistency dimension (belonging to the representational group) is inter-related with the syntactic accuracy dimension (belonging to the intrinsic group), because the invalid usage of vocabularies may lead to inconsistency in the data. Another dimension in the representational group, versatility is related to the accessibility dimension since provision of data via different means (e.g. SPARQL endpoint, RDF dump) inadvertently points towards the ways in which data can be accessed. Furthermore, there exists an inter-relation between the conciseness and the relevancy dimensions. Conciseness frequently positively affects relevancy since removing redundancies increases the amount of relevant data.

The interlinking dimension is associated with the syntactic accuracy dimension. It is important to choose the correct similarity relationship such as *same, matches, similar* or *related* between two entities to capture the most appropriate relationship [23] thus contributing towards the syntactic accuracy of the data. Also, the amount-of-data dimension is related to the completeness dimension. In certain cases the amount of data, for example, can be an indicator for the completeness. Moreover, large amounts of data also affect the performance of the system or querying of the dataset.

These examples of inter-relations between the dimensions, belonging to different groups, indicate the interplay between them and show that these dimensions are to be considered differently in different data quality assessment scenarios.

5. Comparison of selected approaches

In this section, we compare the 21 selected approaches based on the different perspectives discussed in Section 2 (Comparison perspective of selected approaches). In particular, we analyze each approach based on the dimensions (Section 5.1), their respective metrics (Section 5.2), types of data (Section 5.3), and compare the proposed tools based on several attributes (Section 5.4).

5.1. Dimensions

Linked Open Data touches three different research and technology areas, namely the *Semantic Web* to generate semantic connections among datasets, the *World Wide Web* to make the data available, preferably under an open access license, and *Data Management* for handling large quantities of heterogeneous and distributed data. Previously published literature provides a thorough classification of the data quality dimensions [8,30,55,57,65,67]. By analyzing these classifications, it is possible to distill a core set of dimensions, namely accuracy, completeness, consistency and timeliness. These four dimensions constitute the focus of most authors [60]. However, no consensus exists which set of dimensions defines data quality as a whole or the exact meaning of each dimension, which is also a problem occurring in LOD.

As mentioned in Section 3, data quality assessment involves the measurement of data quality dimensions that are relevant for the user. We therefore gathered all data quality dimensions that have been reported as being relevant for LOD by analyzing the 21 selected approaches. An initial list of data quality dimensions was obtained from [4]. Thereafter, the problem addressed by each approach was extracted and mapped to one or more of the quality dimensions. For example, the problems of dereferencability, the non-availability of structured data, and content misreporting as mentioned in [27] were mapped to the dimensions of completeness as well as availability. However, not all problems related to LOD could be mapped to the initial set of dimensions, including the problem of the alternative data representation and its handling, i.e. the dataset versatility. Therefore, we obtained a further set of quality dimensions from [14], which was one of the first studies focusing specifically on data quality dimensions and metrics applicable to LOD. Still, there were some problems that did not fit in this extended list of dimensions such as incoherency of interlinking between datasets or the different aspects of the timeliness of datasets. Thus, we introduced new dimensions such as *interlinking, volatility and currency* in order to cover all the identified problems in all of the included approaches, while also mapping them to at least one dimension.

Table 8 shows the complete list of 23 LOD quality dimensions along with their respective frequency of occurrence in the included approaches. This table can be intuitively divided into the following three groups: (a) a set of approaches focusing only on trust [7,16,17, 18,20,21,24,29,62]; (b) a set of approaches covering more than four dimensions [5,14,27,50,15,26] and (c) a set of approaches focusing on very few and specific dimensions [6,10,22,26,42,54,59]. Overall, it is observed that the dimensions believability, consistency, currency, completeness, accuracy and availability are the most frequently used. Additionally, the categories intrinsic, trust, contextual, dataset dynamicity, accessibility and representational rank in descending order of importance based on the frequency of occurrence of dimensions. Finally, we can conclude that none of the approaches covers all data quality dimensions that are relevant for LOD and most of the dimensions are discussed in two articles.

5.2. Metrics

As defined in Section 3, a data quality metric is *a procedure for measuring an information quality dimension*. We notice that most of the metrics take the form of a ratio, which measures the occurrence of observed instances out of the occurrence of the desired instances, where by instances we mean properties or classes [42] For example, for the representational-consistency dimension, the metric for determining the re-use of existing vocabularies takes the form of the ratio:

$$\frac{\textit{no. of erroneous instances}}{\textit{total no. of instances}}$$

Other metrics, which cannot be measured as a ratio, can be assessed using algorithms. Table 2, Table 3, Table 4, Table 5, Table 6 and Table 7 provide the metrics for each of the dimensions.

For some of the included approaches, the problem, its corresponding metric, and a dimension were clearly mentioned [4,14]. However, for the other approaches, we first extracted the problem addressed along with the way in which it was assessed i.e. the metric. There-

after, we mapped each problem and the corresponding metric to a relevant data quality dimension. For example, the problem related to keeping URIs short (identified in [27]) measured by the presence of long URIs or those containing query parameters, was mapped to the representational-conciseness dimension. On the other hand, the problem related to the re-use of existing terms (also identified in [27]) was mapped to the representational-consistency dimension.

It is worth noting that for a particular dimension there are several metrics associated with it but each metric is only associated with one dimension. Additionally, there are several ways of measuring one dimension either individually or by combining different metrics. As an example, the availability dimension is measured by a combination of three other metrics, namely accessibility of the (i) server, (ii) SPARQL end-point, and (iii) RDF dumps. Additionally, availability is individually measured by the availability of structured data, misreported content types, or by the absence of dereferencability issues (cf. Table 2). We also classify each metric as being *Objectively* (quantitatively) or *Subjectively* (qualitatively) assessed. Objective metrics are those that are quantified or for which a concrete value can be calculated. For example, for the completeness dimension, the metrics such as schema completeness or property completeness are quantified. The ratio form of the metrics is generally applied to those metrics which can be measured quantitatively (objectively). On the other hand, subjective dimensions are those which cannot be quantified but depend on the users perception of the respective dimension (e.g. via surveys). For example, metrics belonging to dimensions such as objectivity, relevancy highly depend on the user and can only be measured subjectively. There are cases when the metrics of particular dimensions are either entirely subjective (for example relevancy, objectivity) or entirely objective (for example accuracy, conciseness). But, there are also cases when a particular dimension is measured both objectively as well as subjectively. For example, although completeness is perceived as a dimension which is measured objectively, it also includes metrics which are measured subjectively. That is, the schema or ontology completeness is measured subjectively whereas the property, instance and interlinking completeness is measured objectively. Similarly, for the amount-of-data dimension, on the one hand the number of triples, instances per class, internal and external links in a dataset is measured objectively but on the other hand, the scope and level of detail is measured subjectively.

5.3. Type of data

The goal of a data quality assessment activity is the analysis of data in order to measure the quality of datasets along relevant quality dimensions. Therefore, the assessment involves the comparison between the obtained measurements and the references values, in order to enable a diagnosis of quality. The assessment considers different types of data that describe real world objects in a format that can be stored, retrieved, and processed by a software procedure and communicated through a network. Thus, in this section, we distinguish between the types of data considered in the various approaches in order to obtain an overview of how the assessment of LOD operates on such different levels. The assessment is associated with small-scale units of data such as assessment of RDF triples to the assessment of entire datasets which potentially affect the whole assessment process. In LOD, we distinguish the assessment process operating on three types of data:

- RDF triples, which focus on individual triple assessment.
- RDF graphs, which focus on entities assessment where entities are described by a collection of RDF triples [25].
- Datasets, which focus on datasets assessment where a dataset is considered as a set of default and named graphs.

In Table 9, we can observe that most of the methods are applicable at the triple or graph level and to a lesser extend on the dataset level. Additionally, it can be seen that 9 approaches assess data both at triple and graph level [6,7,10,14,15,18,42,54,59], 2 approaches assess data both at graph and dataset level [16,22] and 4 approaches assess data at triple, graph and dataset levels [5,20,26,27]. There are 2 approaches that apply the assessment only at triple level [24,50] and 4 approaches that only apply the assessment at the graph level [17,21,29,62].

In most cases, if the assessment is provided at the triple level, this assessment can usually be propagated at a higher level such as graph or dataset level. For example, in order to assess the rating of a single source, the overall rating of the statements associated to the source can be used [18]. On the other hand, if the assessment is performed at the graph level, it is further

Table 8: Consideration of data quality dimensions in each of the included approaches.

Approaches / Dimensions	Bizer et al.,2009	Flemming,2010	Böhm et al.,2010	Chen et al.,2010	Guéret et al,2011	Hogan et al.,2010	Hogan et al.,2012	Lei et al.,2007	Mendes et al., 2012	Mostafavi et al., 2004	Fürber et al.,2011	Rula et al., 2012	Hartig, 2008	Gamble et al., 2011	Shekarpour et al., 2010	Golbeck, 2006	Gil et al., 2002	Golbeck et al., 2003	Gil et al., 2007	Jacobi et al., 2011	Bonatti et al., 2011
Believability	✔	✔											✔	✔	✔	✔	✔	✔	✔	✔	✔
Consistency	✔	✔	✔			✔		✔	✔	✔											
Currency	✔	✔							✔			✔									
Volatility											✔										
Timeliness											✔			✔							
Accuracy	✔					✔		✔			✔										
Completeness	✔				✔	✔			✔		✔										
Amount-of-data	✔	✔		✔																	
Availability	✔	✔				✔	✔														
Understandability	✔	✔																			
Relevancy	✔																				
Reputation									✔										✔		
Verifiability	✔	✔																			
Interpretability	✔						✔														
Rep.-conciseness	✔						✔														
Rep.-consistency	✔	✔					✔														
Licensing		✔					✔														
Performance	✔	✔					✔														
Objectivity	✔																				
Security		✔																			
Versatility		✔																			
Conciseness								✔	✔		✔										
Interlinking					✔		✔														

Table 9
Qualitative evaluation of the 21 core frameworks included in this survey.

Paper	**Application**	**Goal**	**Type of data**			**Tool support**	
	General/ **S**pecific		**RDF Triple**	**RDF Graph**	**Dataset**	**Tool implemented**	**URL**
Gil et al., 2002	G	Approach to derive an assessment of a data source based on the annotations of many individuals	✔	✔	–	✔	http://www.isi.edu/ikcap/trellis/demo.html
Golbeck et al., 2003	G	Trust networks on the semantic web	–	✔	–	✔	http://trust.mindswap.org/trustMail.shtml
Mostafavi et al.,2004	S	Spatial data integration	✔	✔	–	–	–
Golbeck, 2006	G	Algorithm for computing personalized trust recommendations using the provenance of existing trust annotations in social networks	✔	✔	✔	–	
Gil et al., 2007	S	Trust assessment of web resources	–	✔	–	–	–
Lei et al., 2007	S	Assessment of semantic metadata	✔	✔	–	–	–
Hartig, 2008	G	Trustworthiness of Data on the Web	✔	–	–	✔	http://trdf.sourceforge.net/tsparql.shtml
Bizer et al., 2009	G	Information filtering	✔	✔	✔	✔	http://wifo5-03.informatik.uni-mannheim.de/bizer/wiqa/
Böhm et al., 2010	G	Data integration	✔	✔	–	✔	http://tinyurl.com/prolod-01
Chen et al., 2010	G	Generating semantically valid hypothesis	✔	✔	–	–	–
Flemming, 2010	G	Assessment of published data	✔	✔	–	✔	http://linkeddata.informatik.hu-berlin.de/LDSrcAss/
Hogan et al., 2010	G	Assessment of published data by identifying RDF publishing errors and providing approaches for improvement	✔	✔	✔	–	–
Shekarpour et al., 2010	G	Method for evaluating trust	-	✔	–	–	–
Fürber et al., 2011	G	Assessment of published data	✔	✔	–	–	–
Gamble et al., 2011	G	Application of decision networks to quality, trust and utility assessment	-	✔	✔	–	–
Jacobi et al., 2011	G	Trust assessment of web resources	–	✔	–	–	–
Bonatti et.al., 2011	G	Provenance assessment for reasoning	✔	✔	–	–	–
Guéret et al., 2012	G	Assessment of quality of links	–	✔	✔	✔	https://github.com/LATC/24-7-platform/tree/master/latc-platform/linkqa
Hogan et al., 2012	G	Assessment of published data	✔	✔	✔	–	–
Mendes et al., 2012	G	Data integration	✔	–	–	✔	http://sieve.wbsg.de/
Rula et al., 2012	G	Assessment of time related quality dimensions	✔	✔	–	–	–

Table 10
Comparison of quality assessment tools according to several attributes.

	Trellis, Gil et al., 2002	**TrustBot, Golbeck et al., 2003**	**tSPARQL, Hartig, 2008**	**WIQA, Bizer et al., 2009**	**ProLOD, Böhm et al., 2010**	**Flemming, 2010**	**LinkQA, Gueret et al., 2012**	**Sieve, Mendes et al., 2012**
Accessibility/ Availability	–	–	✔	–	–	✔	✔	✔
Licensing	Open-source	–	GPL v3	Apache v2	–	–	Open-source	Apache
Automation	Semi-automated	Semi-automated	Semi-automated	Semi-automated	Semi-automated	Semi-automated	Automated	Semi-automated
Collaboration	Allows users to add observations and conclusions	No	No	No	No	No	No	No
Customizability	✔	✔	✔	✔	✔	✔	No	✔
Scalability	–	No	Yes	–	–	No	Yes	Yes
Usability/ Documentation	2	4	4	2	2	3	2	4
Maintenance (Last updated)	2005	2003	2012	2006	2010	2010	2011	2012

propagated either to a more fine grained level that is the RDF triple level or to a more generic one, that is the dataset level. For example, the evaluation of trust of a data source (graph level) is propagated to the statements (triple level) that are part of the Web source associated with that trust rating [62]. However, there are no approaches that perform an assessment only at the dataset level (cf. Table 9). A reason is that the assessment of a dataset always involves the assessment of a fine grained level (such as triple or entity level) and this assessment is then propagated to the dataset level.

5.4. Comparison of tools

In Table 10, we compare the tools proposed by eight of the 21 core articles based on eight different attributes. These tools implement the methodologies and metrics defined in the respective approaches.

Accessibility/Availability. The URL for accessing each tool is available in Table 9. In Table 10, only the tools marked with a ✔are available to be used for quality assessment. The other tools are either available only as a demo or screencast (Trellis, ProLOD) or not available at all (TrustBot, WIQA).

Licensing. Each of the tools is available using a particular software license, which specifies the restrictions with which it can be redistributed. The Trellis and LinkQA tools are open-source and as such by default they are protected by copyright which is *All Rights Reserved*. Also, WIQA and Sieve are both available with open-source licence: the Apache Version 2.0[28] and Apache licenses respectively. tSPARQL is distributed under the GPL v3 license[29]. However, no licensing information is available for TrustBot, ProLOD and Flemming's tool.

Automation. The automation of a system is the ability to automatically perform its intended tasks thereby reducing the need for human intervention. In this context, we classify the eight tools into semi-automated and automated approaches. As seen in Table 10, all the tools are semi-automated except for LinkQA, which is completely automated as there is no user involvement. LinkQA automatically selects a set of resources, information from the Web of Data (i.e. SPARQL endpoints and/or dereferencable resources) and a set of new triples as input and generates the respective quality assessment reports. On the other hand, the WIQA and Sieve tools require a high degree of user involvement. Specifically in Sieve, the definition of metrics has to be done by creating an XML file which contains specific configurations for a quality assessment task. Although it gives the users the flexibility of tweaking the tool to match their needs, it requires much time for understanding the required XML file structure and specification. The other semi-automated tools, Trellis, TrurstBot, tSPARQL, ProLOD and Flemming's tool require a minimum amount of user involvement. For example, Flemming's Data Quality Assessment Tool requires the user to answer a few questions regarding the dataset (e.g. existence of a human-readable license) or they have to assign weights to each of the pre-defined data quality metrics.

[28] http://www.apache.org/licenses/LICENSE-2.0

[29] http://www.gnu.org/licenses/gpl-3.0.html

Collaboration. Collaboration is the ability of a system to support co-operation between different users of the system. None of the tools, except Trellis, supports collaboration between different users of the tool. The Trellis user interface allows several users to express their trust value for a data source. The tool allows several users to add and store their observations and conclusions. Decisions made by users on a particular source are stored as annotations, which can be used to analyze conflicting information or handle incomplete information.

Customizability. Customizability is the ability of a system to be configured according to the users' needs and preferences. In this case, we measure the customizability of a tool based on whether the tool can be used with any dataset that the user is interested in. Only LinkQA can not be customized since the user cannot add any dataset of her choice. The other seven tools can be customized according to the use case. For example, in TrustBot, an IRC bot that makes trust recommendations to users (based on the trust network it builds), the users have the flexibility to submit their own URIs to the bot at any time while incorporating the data into a graph. Similarly, Trellis, tSPARQL, WIQA, ProLOD, Flemming's tool and Sieve can be used with any dataset.

Scalability. Scalability is the ability of a system, network, or process to handle a growing amount of work or its ability to be enlarged to accommodate that growth. Out of the eight tools only three, the tSPARQL, LinkQA and Sieve, tools are scalable, that is, they can be used with large datasets. TrustBot and Flemming's tool are reportedly not scalable for large datasets. Trellis, WIQA and ProLOD do not provide any information on the scalability.

Usability/Documentation. Usability is the ease of use and learnability of a human-made object, in this case the quality assessment tool. We assess the usability of the tools based on the ease of use as well as the complete and precise documentation available for each of them. We score them based on a scale from 1 (low usability) – 5 (high usability). TrustBot, tSPARQL and Sieve score high in terms of usability and documentation followed by Flemming's data quality assessment tool. Trellis, WIQA, ProLOD and LinkQA rank lower in terms of ease of use since they do not contain useful documentation of how to use the tool.

Maintenance/Last updated. While TrustBot, Trellis and WIQA have not been updated since they were first introduced in 2003, 2005 and 2006 respectively, ProLOD and Flemming's tool have been updated in 2010. The recently updated tools are LinkQA (2011), tSRARQL (2012) and Sieve (2012) and are currently being maintained.

6. Conclusions and future work

In this paper, we have presented, to the best of our knowledge, the most comprehensive systematic review of data quality assessment methodologies applied to LOD. The goal of this survey is to obtain a clear understanding of the differences between such approaches, in particular in terms of quality dimensions, metrics, type of data and tools available.

We surveyed 21 approaches and extracted 23 data quality dimensions along with their definitions and corresponding metrics. We analyzed the approaches in terms of the dimensions, metrics and type of data they focus on. Additionally, we identified tools proposed by eight approaches (out of the 21) and compared them using eight different attributes.

We observed that most of the publications focusing on data quality assessment in Linked Open Data are presented at either conferences or workshops. As our literature review reveals, the number of publications published in the span of 10 years (i.e. 21) is rather low. This can be attributed to the infancy of this research area. Additionally, in most of the surveyed literature, the metrics were often not explicitly defined or did not consist of precise statistical measures. Moreover, only few approaches were actually accompanied by an implemented tool. Also, there was no formal validation of the methodologies that were implemented as tools. We also observed, that none of the existing implemented tools covered all the data quality dimensions. In fact, the best coverage in terms of dimensions was achieved by Flemming's data quality assessment tool with 11 covered dimensions. Our survey shows that the flexibility of the tools, with regard to the level of automation and user involvement, needs to be improved. Some tools required a considerable amount of configuration while some others were easy-to-use but provided only results with limited usefulness or required a high-level of interpretation.

Meanwhile, there is much research on data quality being done and guidelines as well as recommendations on how to publish "good" data are currently available. However, there is less focus on how to use this "good" data. We deem our data quality dimensions to be very useful for data consumers in order to assess the quality of datasets. As a next step, we aim to integrate the

various data quality dimensions into a comprehensive methodological framework for data quality assessment comprising the following steps:

1. Requirements analysis,
2. Data quality checklist,
3. Statistics and low-level analysis,
4. Aggregated and higher level metrics,
5. Comparison,
6. Interpretation.

We aim to develop this framework for data quality assessment allowing a data consumer to select and assess the quality of suitable datasets according to this methodology. In the process, we also expect new metrics to be defined and implemented.

References

[1] Sören Auer, Matthias Weidl, Jens Lehmann, Amrapali Zaveri, and Key-Sun Choi. I18n of semantic web applications. In *ISWC 2010*, volume 6497 of *LNCS*, pages 1 – 16, 2010.

[2] Carlo Batini, Cinzia Cappiello, Chiara Francalanci, and Andrea Maurino. Methodologies for data quality assessment and improvement. *ACM Computing Surveys*, 41(3), 2009.

[3] Carlo Batini and Monica Scannapieco. *Data Quality: Concepts, Methodologies and Techniques*. Springer-Verlag New York, Inc., 2006.

[4] Christian Bizer. *Quality-Driven Information Filtering in the Context of Web-Based Information Systems*. PhD thesis, Freie Universität Berlin, March 2007.

[5] Christian Bizer and Richard Cyganiak. Quality-driven information filtering using the wiqa policy framework. *Journal of Web Semantics*, 7(1):1–10, 2009.

[6] Christoph Böhm, Felix Naumann, Ziawasch Abedjan, Dandy Fenz, Toni Grütze, Daniel Hefenbrock, Matthias Pohl, and David Sonnabend. Profiling linked open data with ProLOD. In *ICDE Workshops*, pages 175–178. IEEE, 2010.

[7] Piero A. Bonatti, Aidan Hogan, Axel Polleres, and Luigi Sauro. Robust and scalable linked data reasoning incorporating provenance and trust annotations. *Journal of Web Semantics*, 9(2):165 – 201, 2011.

[8] Matthew Bovee, Rajendra P. Srivastava, and Brenda Mak. A conceptual framework and belief-function approach to assessing overall information quality. *International Journal of Intelligent Systems*, 18(1):51–74, 2003.

[9] Jeremy J Carroll. Signing RDF graphs. In *ISWC*, pages 369–384. Springer, 2003.

[10] Ping Chen and Walter Garcia. Hypothesis generation and data quality assessment through association mining. In *IEEE ICCI*, pages 659–666. IEEE, 2010.

[11] Luca Costabello, Serena Villata, and Fabien Gandon. Context-aware access control for rdf graph stores. In *20th ECAI*, 2012.

[12] Jan Demter, Sören Auer, Michael Martin, and Jens Lehmann. LODStats – an extensible framework for high-performance dataset analytics. In *EKAW*, LNCS. Springer, 2012.

[13] Li Ding and Tim Finin. Characterizing the semantic web on the web. In *5th ISWC*, 2006.

[14] Annika Flemming. Quality characteristics of linked data publishing datasources. `http://sourceforge.net/apps/mediawiki/trdf/index.php?title=Quality_Criteria_for_Linked_Data_sources`, 2010.

[15] Christian Fürber and Martin Hepp. SWIQA - a semantic web information quality assessment framework. In *ECIS*, 2011.

[16] Matthew Gamble and Carole Goble. Quality, trust, and utility of scientific data on the web: Towards a joint model. In *ACM WebSci*, pages 1–8, June 2011.

[17] Yolanda Gil and Donovan Artz. Towards content trust of web resources. *Web Semantics*, 5(4):227 – 239, December 2007.

[18] Yolanda Gil and Varun Ratnakar. Trusting information sources one citizen at a time. In *ISWC*, pages 162 – 176, 2002.

[19] Jennifer Golbeck. Inferring reputation on the semantic web. In *WWW*, 2004.

[20] Jennifer Golbeck. Using trust and provenance for content filtering on the semantic web. In *Workshop on Models of Trust on the Web at the 15th World Wide Web Conference*, 2006.

[21] Jennifer Golbeck, Bijan Parsia, and James Hendler. Trust networks on the semantic web. In *CIA*, 2003.

[22] Christophe Guéret, Paul Groth, Claus Stadler, and Jens Lehmann. Assessing linked data mappings using network measures. In *ESWC*, 2012.

[23] Harry Halpin, Pat Hayes, James P. McCusker, Deborah McGuinness, and Henry S. Thompson. When owl:sameas isn't the same: An analysis of identity in linked data. In *9th ISWC*, volume 1, pages 53–59, 2010.

[24] Olaf Hartig. Trustworthiness of data on the web. In *STI Berlin and CSW PhD Workshop, Berlin, Germany*, 2008.

[25] Tom Heath and Christian Bizer. *Linked Data: Evolving the Web into a Global Data Space*, chapter 2, pages 1 – 136. Number 1:1 in Synthesis Lectures on the Semantic Web: Theory and Technology. Morgan and Claypool, 1st edition, 2011.

[26] Aidan Hogan, Andreas Harth, Alexandre Passant, Stefan Decker, and Axel Polleres. Weaving the pedantic web. In *3rd Linked Data on the Web Workshop at WWW*, 2010.

[27] Aidan Hogan, Jürgen Umbrich, Andreas Harth, Richard Cyganiak, Axel Polleres, and Stefan Decker. An empirical survey of Linked Data conformance. *Journal of Web Semantics*, 2012.

[28] I Horrocks, P.F. Patel-Schneider, H Boley, S Tabet, B Grosof, and M Dean. SWRL: A semantic web rule language combining OWL and RuleML. Technical report, W3C, May 2004.

[29] Ian Jacobi, Lalana Kagal, and Ankesh Khandelwal. Rule-based trust assessment on the semantic web. In *RuleML*, pages 227 – 241, 2011.

[30] Matthias Jarke, Maurizio Lenzerini, Yannis Vassiliou, and Panos Vassiliadis. *Fundamentals of Data Warehouses*. Springer Publishing Company, 2nd edition, 2010.

[31] Joseph Juran. *The Quality Control Handbook*. McGraw-Hill, New York, 1974.

[32] Tobias Käfer, Ahmed Abdelrahman, Jürgen Umbrich, Patrick O'Byrne, and Aidan Hogan. Observing linked data dynamics. In *ESWC*, pages 213–227, 2013.

[33] Beverly K. Kahn, Diane M. Strong, and Richard Y. Wang. Information quality benchmarks: product and service performance. *Commun. ACM*, 45(4):184–192, April 2002.

[34] Michael Kifer and Harold Boley. RIF overview. Technical report, W3C, June 2010. `http://www.w3.org/TR/2012/NOTE-rif-overview-20121211/`.

[35] B. Kitchenham. Procedures for performing systematic reviews. Technical report, Joint Technical Report Keele Univer-

sity Technical Report TR/SE-0401 and NICTA Technical Report 0400011T.1, 2004.

[36] S.A. Knight and J. Burn. Developing a framework for assessing information quality on the World Wide Web. *Information Science*, 8:159 – 172, 2005.

[37] Jose Emilio Labra Gayo, Dimitris Kontokostas, and Sören Auer. Multilingual linked open data patterns. *Semantic Web Journal*, 2012.

[38] Yang W. Lee, Diane M. Strong, Berverly K. Kahn, and Richard Y. Wang. AIMQ: a methodology for information quality assessment. *Information Management*, 40(2):133 – 146, 2002.

[39] Jens Lehmann, Daniel Gerber, Mohamed Morsey, and Axel-Cyrille Ngonga Ngomo. DeFacto - Deep Fact Validation. In *ISWC*. Springer Berlin / Heidelberg, 2012.

[40] Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N. Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick van Kleef, Sören Auer, and Christian Bizer. DBpedia - a large-scale, multilingual knowledge base extracted from wikipedia. *Semantic Web Journal*, 2013. Under review.

[41] Yuangui Lei, Andriy Nikolov, Victoria Uren, and Enrico Motta. Detecting quality problems in semantic metadata without the presence of a gold standard. In *Workshop on "Evaluation of Ontologies for the Web" (EON) at the WWW*, pages 51–60, 2007.

[42] Yuangui Lei, Victoria Uren, and Enrico Motta. A framework for evaluating semantic metadata. In *4th International Conference on Knowledge Capture*, number 8 in K-CAP '07, pages 135 – 142. ACM, 2007.

[43] David Kopcso Leo Pipino, Ricahrd Wang and William Rybold. *Developing Measurement Scales for Data-Quality Dimensions*, volume 1. M.E. Sharpe, New York, 2005.

[44] Holger Michael Lewn. *Facilitating Ontology Reuse Using User-Based Ontology Evaluation*. PhD thesis, Karlsruher Instituts für Technologie (KIT), 2010.

[45] ZJ Lipowski. Sensory and information inputs overload: Behavioral effects. *Comprehensive Psychiatry*, 1975.

[46] Diana Maynard, Wim Peters, and Yaoyong Li. Metrics for evaluation of ontology-based information extraction. In *Workshop on "Evaluation of Ontologies for the Web" (EON) at WWW*, May 2006.

[47] Massimo Mecella, Monica Scannapieco, Antonino Virgillito, Roberto Baldoni, Tiziana Catarci, and Carlo Batini. Managing data quality in cooperative information systems. In *Confederated International Conferences DOA, CoopIS and ODBASE*, pages 486–502, 2002.

[48] Christian Meilicke and Heiner Stuckenschmidt. Incoherence as a basis for measuring the quality of ontology mappings. In *3rd International Workshop on Ontology Matching (OM) at the ISWC*, 2008.

[49] P Mendes, Z Bizer, J.P. Miklos, A. Moraru Clabimonte, and Flouris G. D2.1: Conceptual model and best practices for high-quality metadata publishing. Technical report, PlanetData Deliverable, 2012.

[50] P Mendes, H Mühleisen, and C Bizer. Sieve: Linked data quality assessment and fusion. In *LWDM*, March 2012.

[51] Paul Miller, Rob Styles, and Tom Heath. Open data commons, a license for open data. In *LDOW*, 2008.

[52] David Moher, Alessandro Liberati, Jennifer Tetzlaff, Douglas G Altman, and PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *PLoS medicine*, 6(7), 2009.

[53] Mohamed Morsey, Jens Lehmann, Sören Auer, Claus Stadler, and Sebastian Hellmann. DBpedia and the Live Extraction of Structured Data from Wikipedia. *Program: electronic library and information systems*, 46:27, 2012.

[54] MA. Mostafavi, Edwards G., and R. Jeansoulin. Ontology-based method for quality assessment of spatial data bases. In *International Symposium on Spatial Data Quality*, volume 4, pages 49–66, 2004.

[55] Felix Naumann. *Quality-Driven Query Answering for Integrated Information Systems*, volume 2261 of *Lecture Notes in Computer Science*. Springer-Verlag, 2002.

[56] L. Pipino, Yang W. Lee, and Richard Y. Wang. Data quality assessment. *Communications of the ACM*, 45(4), 2002.

[57] Thomas C. Redman. *Data Quality for the Information Age*. Artech House, 1st edition, 1997.

[58] Anisa Rula, Matteo Palmonari, Andreas Harth, Steffen Stadtmüller, and Andrea Maurino. On the Diversity and Availability of Temporal Information in Linked Open Data. In *ISWC*, 2012.

[59] Anisa Rula, Matteo Palmonari, and Andrea Maurino. Capturing the Age of Linked Open Data: Towards a Dataset-independent Framework. In *IEEE International Conference on Semantic Computing*, 2012.

[60] Monica Scannapieco and Tiziana Catarci. Data quality under a computer science perspective. *Archivi & Computer*, 2:1–15, 2002.

[61] Daniel Schober, Smith. Barry, E. Suzanna Lewis, Waclaw Kusnierczyk, Jane Lomax, Chris Mungall, F. Chris Taylor, Philippe Rocca-Serra, and Susanna-Assunta Sansone. Survey-based naming conventions for use in OBO foundry ontology development. *BMC Bioinformatics*, 10(125), 2009.

[62] Saeedeh Shekarpour and S.D. Katebi. Modeling and evaluation of trust with an extension in semantic web. *Web Semantics: Science, Services and Agents on the World Wide Web*, 8(1):26 – 36, March 2010.

[63] Fabian M. Suchanek, David Gross-Amblard, and Serge Abiteboul. Watermarking for ontologies. In *ISWC*, 2011.

[64] Denny Vrandecic. *Ontology Evaluation*. PhD thesis, Karlsruher Instituts für Technologie (KIT), 2010.

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

[66] Richard Y. Wang. A product perspective on total data quality management. *Communications of the ACM*, 41(2):58 – 65, Feb 1998.

[67] Richard Y. Wang and Diane M. Strong. Beyond accuracy: what data quality means to data consumers. *Journal of Management Information Systems*, 12(4):5–33, 1996.

[68] Yang Yu and Jeff Heflin. Extending functional dependency to detect abnormal data in rdf graphs. In *ISWC*, volume 7031, pages 794–809. Springer Berlin Heidelberg, 2011.

[69] Amrapali Zaveri, Dimitris Kontokostas, Mohamed A. Sherif, Lorenz Bühmann, Mohamed Morsey, Sören Auer, and Jens Lehmann. User-driven quality evaluation of DBpedia. In *9th International Conference on Semantic Systems, I-SEMANTICS '13 (To Appear)*. ACM, 2013.