Undefined 1 (2012) 1–5
IOS Press

Quality Assessment Methodologies for Linked Open Data

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 [a]

[a] *Universität Leipzig, Institut für Informatik, D-04103 Leipzig, Germany,*
E-mail: (zaveri, lehmann, auer)@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

Abstract. The development and standardization of semantic web technologies have resulted in an unprecedented volume of data being published on the Web as Linked Open Data (LOD). However, we observe widely varying data quality ranging from extensively curated datasets to crowd-sourced and extracted data of relatively low quality. Data quality is commonly conceived as *fitness of use*. Consequently, a key challenge is to determine the data quality wrt. a particular use case. In this article, we present the results of a systematic review of approaches for assessing the data quality of LOD. We gather existing approaches and compare and group them under a common classification scheme. In particular, we unify and formalise commonly used terminologies across papers related to data quality. Additionally, a comprehensive list of the dimensions and metrics is presented. The aim of this article is to provide researchers and data curators a comprehensive understanding of existing work, thereby encouraging further experimentation and the development of new approaches focused towards data quality.

Keywords: data quality, assessment, survey, Linked Data

1. Introduction

The development and standardization of semantic web technologies have resulted in an unprecedented volume of data being published on the Web as *Linked Open Data* (LOD). This emerging Web of Data comprises close to 50 billion facts represented as RDF triples. 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*. Specifically, biological and health care data is available as Linked Data in a great variety covering areas such as drugs, clinical trials, proteins, and diseases. 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, often contains inconsistencies as well as misrepresented and incomplete information.

Data quality is commonly conceived as *fitness for use* [31,57,34] for a certain application or use case. However, 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 ne-

glectable, 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 completely insufficient. 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 [47,35,56,3]. However, data 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. 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.

Therefore, in this paper, we present the findings of a systematic review of existing approaches that focus towards assessing the quality of Linked 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 relevant for our survey that 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,
- 26 different dimensions and their formalized definitions,
- metrics for each of the dimensions along with a distinction between them being subjective or objective and
- methodologies and supported tools used to assess data quality.

Our goal is to provide researchers 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, (b) definitions for each of the data quality dimensions and (c) metrics for each of the dimensions. In Section 4, we compare the selected approaches based on different perspectives such as, (a) dimensions, (b) metrics, (c) type of data, (d) level of automation and (e) comparing three particular tools to gauge their usability for data quality assessment. In Section 5, we conclude with ideas for future work.

2. Survey Methodology

This systematic review was conducted by two reviewers from different institutions following the systematic review procedures described in [33,44]. A systematic review can be conducted for several reasons such as: (a) the summarisation 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 summarises and compares various data quality assessment methodologies as well as identifies open problems related to Linked Open Data. Moreover, a conceptualization of the data quality assessment field is proposed. An overview of our search methodology and the number of retrieved articles at each step is shown in Figure 1.

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 [34] 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, in that, it focuses only on structured data and on approaches that aim at assessing the quality of LOD. Additionally, the prior review 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 Linked Data. Furthermore, we describe quality assessment metrics corresponding to each of the dimensions and also identified whether they can be 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 Linked Data. To achieve this goal, we aim to answer the following general research question:

How can we assess the quality of Linked 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 follows:

- Inclusion criteria:
  - Studies published in English between 2002 and 2012.
  - Studies focused on data quality assessment for Linked Data
  - Studies focused on provenance assessment of Linked Data
  - Studies that proposed and implemented an approach for data quality assessment
  - Studies that assessed the quality of Linked Data 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 methodologies
  - Studies that neither focused on Linked Data nor on other forms of structured data
  - Studies that did non 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. 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:

- *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*

In our experience, 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

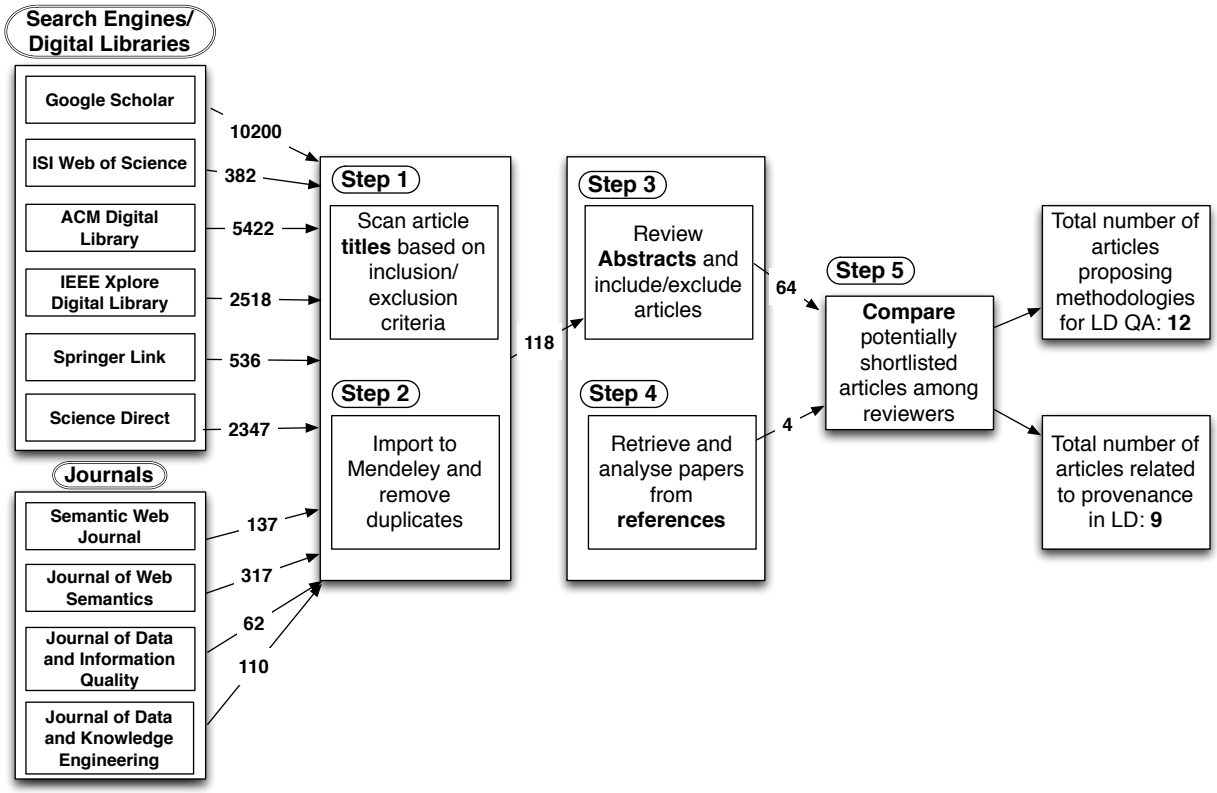

Fig. 1. Number of articles retrieved during literature search.

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[1].

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.

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

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

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 [45]	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 [38]	A framework for evaluating semantic metadata
Hartig, 2008 [23]	Trustworthiness of Data on the Web
Bizer et.al., 2009 [6]	Quality-driven information filtering using the WIQA policy framework
Böhm et.al., 2010 [7]	Profiling linked open data with ProLOD
Chen et.al., 2010 [12]	Hypothesis generation and data quality assessment through association mining
Flemming et.al., 2010 [14]	Assessing the quality of a Linked Data source
Hogan et.al.,2010 [26]	Weaving the Pedantic Web
Shekarpour et.al., 2010 [53]	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 [8]	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 [42]	Sieve: Linked Data Quality Assessment and Fusion
Rula et.al., 2012 [50]	Capturing the Age of Linked Open Data: Towards a Dataset-independent Framework

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 level of automation of the supported tools

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

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 papers. 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 Section 3.1) as well as the Linked Data quality dimensions (in Section 3.2) along with their respective metrics in detail.

3.1. General terms

RDF Dataset. In this document, we understand a data source as an access point for Linked Data in the Web. A data source provides a dataset and may support multiple methods of access. The terms RDF triple, RDF graph and RDF datasets have been adopted from the W3C Data Access Working Group [4,24,10].

Given an infinite set $\mathcal{U}$ of URIs (resource identifiers), an infinite set $\mathcal{B}$ of blank nodes, and an infinite set $\mathcal{L}$ of literals, a triple $\langle s, p, o\rangle \in (\mathcal{U} \cup \mathcal{B}) \times \mathcal{U} \times (\mathcal{U} \cup \mathcal{B} \cup \mathcal{L})$ is called an RDF triple; where s, p, o are the subject, the predicate and the object of the triple, respectively. An RDF graph G is a

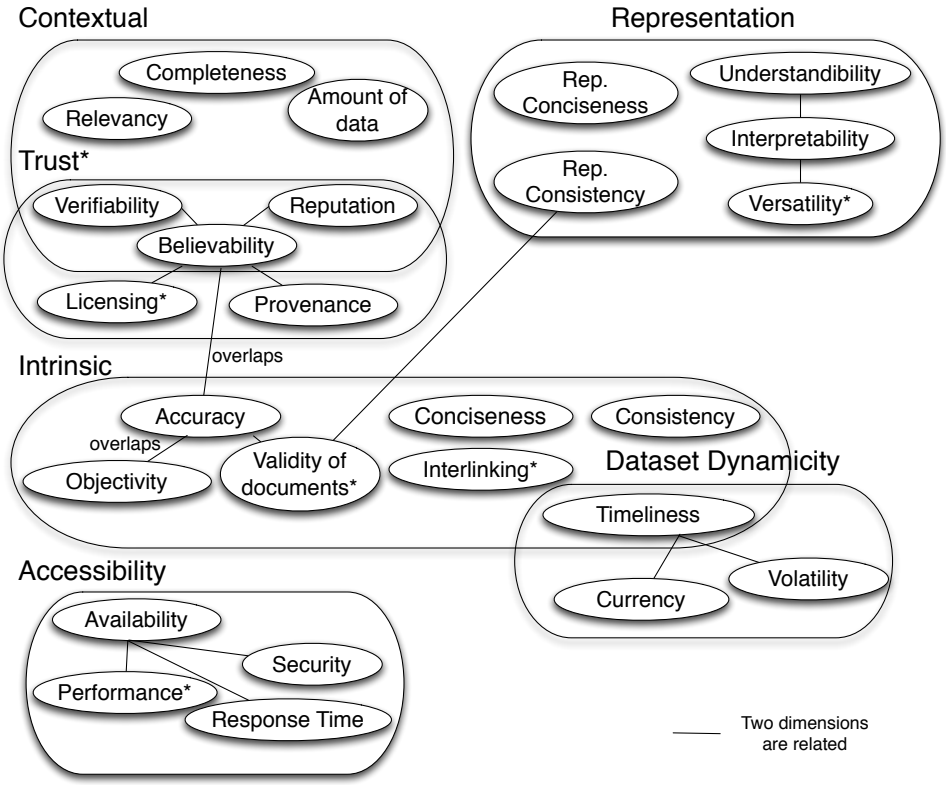

Fig. 2. Linked data quality dimensions and the relations between them. The dimensions marked with a * are newly introduced specifically for Linked Data.

set of RDF triples. A named graph is a pair $\langle G, u\rangle$, where G is called the default graph and $u \in \mathcal{U}$. An RDF dataset is a set of default and named graphs = $\{G, (u_1, G_1), (u_2, G_2), ...(u_n, G_n)\}$.

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 [57].

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 [37]. On the other hand, the notion of link quality is another important aspect in Linked Data that is introduced, where it is automatically detected whether a link is useful or not [22]. Also, it is to be noted that *data* and *information* are interchangeably used in the literature.

Data Quality Problems. Bizer et al. [6] defines data quality problems as the choice of web-based information systems design which integrate information from different providers. In [42] 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.

In [14] the author does not provide a definition but implicitly explains the problems in terms of *data diversity*. In [26] the authors discuss about *errors* or *noise* or *difficulties* and in [27] the author discuss about *modelling issues* which are prone of the non exploitations of those data from the applications.

Thus, data quality problem 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. A data quality assessment *metric* or *measure* is a procedure for measuring an information quality dimension [6]. These metrics are heuristics that are designed to fit a specific assessment situation [39]. 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 crite-

ria [14]. An assessment score is computed from these indicators using a scoring function.

In [6], the data quality dimensions are classified 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 3.2) and also further categories to classify the dimensions, namely (1) Contextual (2) Trust (3) Intrinsic (4) Accessibility (5) Representational and (6) Dataset Dynamicity dimensions, as depicted in Figure 2.

Data Quality Assessment Method. A data quality assessment methodology is defined as the process of evaluating if a piece of data meets in the information consumers need in a specific use case [6]. The process involves measuring the quality dimensions that are relevant to the user and comparing the assessment results with the users quality requirements.

3.2. Linked Data quality dimensions

After analyzing the 21 selected approaches in detail, we identified a core set of 26 different data quality dimensions that can be applied to assess the quality of Linked Data. In this section, we formalize and adapt the definition for each dimension to the Linked Data context. The metrics associated with each dimension are also identified and reported. Additionally, we group the dimensions following a classification introduced in [57] which is further modified and extended as:

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

These groups are not strictly disjoint but can partially overlap. Also, the dimensions are not independent from each other but correlations exists among dimensions in the same group or between groups. Figure 2 shows the classification of the dimensions in the above mentioned groups as well as the inter and intra relations between them.

Use case scenario. Since data quality is described as "fitness to 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[2], OpenFlights[3]). Information about the location of countries, cities and particular addresses is obtained from a spatial dataset (e.g., LinkedGeoData[4]). Additionally, aggregators in RDF pull all the information related to airlines from different data sources (e.g., Expedia[5], Tripadvisor[6], Skyscanner[7] etc.) that allows a user to query the integrated dataset for a flight from 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 to use.

3.2.1. Contextual dimensions

Contextual dimensions are those that highly depend on the context of the task at hand as well as on the subjective preferences of the data consumer. There are three dimensions *completeness*, *amount-of-data* and *relevancy* that are part of this group, which are along with a comprehensive list of their corresponding metrics listed in Table 2. The reference for each metric is provided in the table.

3.2.1.1. Completeness. Completeness is defined as "the degree to which information is not missing" in [5]. This dimension is further classified in [15] into the following categories: (a) Schema completeness, which is the degree to which entities and attributes are not missing in a schema; (b) Column completeness, which is a function of the missing values for a specific property or column; and (c) Population completeness, which refers to the ratio of entities represented in an information system to the complete population. In [42], completeness is distinguished on the schema level 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, the authors in [5] give a general definition whereas the authors in [15] provide a set of sub-categories for completeness. On the other

[2] http://thedatahub.org/dataset/ourairports
[3] http://thedatahub.org/dataset/open-flights
[4] linkedgeodata.org
[5] http://www.expedia.com/
[6] http://www.tripadvisor.com/
[7] http://www.skyscanner.de/

Dimension	Metric	Description	Type
Completeness	degree to which classes and properties are not missing	detection of the degree to which the classes and properties of an ontology are represented [5,15,42]	S
	degree to which values for a property are not missing	detection of no. of missing values for a specific property [5, 15]	O
	degree to which real-world objects are not missing	detection of the degree to which all the real-world objects are represented [5,15,26,42]	O
	degree to which interlinks are not missing	detection of the degree to which instances in the dataset are interlinked [22]	O
Amount-of-data	appropriate volume of data for a particular task	no. of triples, instances per class, internal and external links in a dataset [14,12]	O
	coverage	scope and level of detail [14]	S
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 [5]	S
	retrieval of relevant documents	sorting documents according to their relevancy for a given query [5]	S

Table 2

Comprehensive list of data quality metrics of the contextual dimensions, how it can be measured and it's type - "S"ubjective or "O"bjective

hand, in [42], the two types defined i.e. the schema and data level completeness can be mapped to the two categories (a) Schema completeness and (c) Population completeness provided in [15].

Definition 1 (Completeness)**.** *Completeness refers to the degree to which all required information is present in a particular dataset. In general, completeness is the extent to which data is of sufficient depth, breadth and scope for the task at hand. In terms of Linked Data, we classify completeness as follows: (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 Linked Data and refers to the degree to which instances in the dataset are interlinked.*

Metrics. Completeness can be measured by detecting the number of classes, properties, values and interlinks that are present in the dataset compared to the original dataset (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.

Example. In our use case, the flight search engine should contain complete information so that it includes all offers for flights (population completeness). For example, a user residing in Germany wants to visit her friend in America. Since the user is a student, low price is of high importance. But, looking for flights individually on the airlines websites shows her flights with very expensive fares. However, using our flight search engine she finds all offers, even the less expensive ones and is also able to compare fares from different airlines and choose the most affordable one. The more complete the information for flights is, including cheap flights, the more visitors the site attracts. Moreover, sufficient interlinks between the datasets will allow her to query the integrated dataset so as to find an optimal route from the start to the end destination (interlinking completeness) in cases when there is no direct flight.

Particularly in Linked Data, completeness is of prime importance when integrating datasets from several sources where one of the goals is to increase completeness. That is, the more sources are queried, the more complete the results will be. 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 requires the presence of a gold standard or the original data source to compare with.

3.2.1.2. Amount-of-data. The amount-of-data dimension can be defined as "the extent to which the volume of data is appropriate for the task at hand" [5]. "The amount of data provided by a data source influences its usability [14] and should be enough to approximate

the "true" scenario precisely" [12]. While the authors in [5] provide a formal definition, the author in [14] explains the dimension by mentioning its advantages and metrics. In case of [12], we analyzed the mentioned problem and the respective metrics and mapped them to this dimension since it best fits the definition.

Definition 2 (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 can be measured in terms of bytes (most coarse-grained), triples, instances, and/or links present in the dataset. This amount should represent an appropriate volume of data for a particular task, that is, appropriate scope and level of detail.

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. This helps to provide the user with better travel plans, which includes smaller cities (airports). For example, when a user wants to travel from Connecticut to Santa Barbara, she might not find direct or indirect flights by searching individual flights websites. But, using our example search engine, she is suggested convenient flight connections between the two destinations, because it contains a large amount of data so as to cover all the airports. She is also offered convenient combinations of planes, trains and buses. The provision of such information also necessitates the presence of a large amount of internal as well as externals links between the datasets so as to provide a fine grained search for flights between specific places.

In essence, this dimension conveys the importance of not having too much unnecessary information, which might overwhelm the data consumer or reduce query performance. An appropriate volume of data, in terms of quantity and coverage, should be a main aim of a dataset provider. In the Linked Data community there is often a focus on large amounts of data being available. 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.

3.2.1.3. Relevancy. In [5], relevancy is explained as "the extent to which information is applicable and helpful for the task at hand" .

Definition 3 (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 can be measured by using meta-information attributes for assessing whether the content is relevant for a particular task. Additionally, retrieval of relevant documents 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 times, duration and cost per person should be provided. If a lot of irrelevant data is included in the spatial data, e.g. post offices, trees etc. (present in LinkedGeoData), query performance can decrease. The user may thus get lost in the silos of information and may not be able to browse through it efficiently to get only what she requires.

When a user is provided with relevant information as a response to her query, higher recall with respect to query answering can be obtained. Data polluted with irrelevant information affects the usability as well as typical query performance of the dataset. Using external links or `owl:sameAs` links can help to pull in additional relevant information about a resource and is thus encouraged in LOD.

Relations between dimensions. For instance, if a dataset is incomplete for a particular purpose, the amount of data is often insufficient. However, if the amount of data is too large, it could be that irrelevant data is provided, which affects the relevance dimension.

3.2.2. Trust dimensions

Trust dimensions are those that focus on the trustworthiness of the dataset. There are five dimensions that are part of this group, namely, *provenance*, *verifiability*, *believability*, *reputation* and *licensing* which are displayed along with their respective metrics in Table 3. The reference for each metric is provided in the table.

3.2.2.1. Provenance. There are many definitions in the literature that emphasize different views of provenance.

Definition 4 (Provenance)**.** *Provenance refers to the contextual metadata that focuses on how to represent, manage and use information about the origin of the source. Provenance helps to describe entities to enable trust, assess authenticity and allow reproducibility.*

Dimension	Metric	Description	Type
Provenance	indication of metadata about a dataset	presence of the title, content and URI of the dataset [14]	O
	computing personalised trust recommendations	using provenance of existing trust annotations in social networks [20]	S
	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 [23]	O
	computing the trustworthiness of RDF statements	computing a trust value based on user-based ratings or opinion-based method [23]	S
	detect the reliability and credibility of a person (publisher)	indication of the level of trust for people a person knows on a scale of 1 - 9 [21]	S
	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) propagation algorithm which utilises 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 [53]	O
	acquiring content trust from users	based on associations that transfer trust from entities to resources [17]	O
	detection of trustworthiness, reliability and credibility of a data source	use of trust annotations made by several individuals to derive an assessment of the sources' trustworthiness, reliability and credibility [18]	S
	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 [8]	O
Verifiability	verifying publisher information	stating the author and his contributors, the publisher of the data and its sources [14]	S
	verifying authenticity of the dataset	whether the dataset uses a provenance vocabulary, eg. the use of the Provenance Vocabulary [14]	O
	verifying correctness of the dataset	with the help of unbiased trusted third party [5]	S
	verifying usage of digital signatures	signing a document containing an RDF serialisation or signing an RDF graph [14]	O
Reputation	reputation of the publisher	survey in a community questioned about other members [17]	S
	reputation of the dataset	analyzing references or page rank or by assigning a reputation score to the dataset [42]	S
Believability	meta-information about the identity of information provider	checking whether the provider/contributor is contained in a list of trusted providers [5]	O
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	detection of whether the work is attributed in the same way as specified by the author or licensor [14]	O
	indication of *Copyleft* or *ShareAlike*	checking whether the derivated contents are published under the same license as the original [14]	O

Table 3

Comprehensive list of data quality metrics of the trust dimensions, how it can be measured and it's type - "S"ubjective or "O"bjective

Metrics. Provenance can be measured by analyzing the metadata associated with the source. This provenance information can in turn be used to assess the trustworthiness, reliability and credibility of a data source, an entity, a publishers or individual RDF statements. There exists an inter-dependancy between the data provider and the data itself. On the one hand, data is likely to be accepted as true if it is provided by a trustworthy provider. On the other hand, the data provider is trustworthy if it provides true data. Thus, both can be checked to measure the trustworthiness.

Example. Our flight search engine constitutes information from several airline providers. In order to verify the reliability of these different airline data providers, provenance information from the aggregators can be analyzed and re-used so as enable users of the flight search engine to trust the authenticity of the data provided to them.

In general, if the source information or publisher is highly trusted, the resulting data will also be highly trusted. Provenance data not only helps determine the trustworthiness but additionally the reliability and credibility of a source [18]. Provenance assertions are a form of contextual metadata and can themselves become important records with their own provenance.

3.2.2.2. Verifiability. Verifiability is described as the "degree and ease with which the information can be checked for correctness" [5]. Similarly, in [14] the verifiability criterion is used as the means a consumer is provided with, which can be used 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 the authors in [5] provide a formal definition whereas the author in [14] describes the dimension by providing its advantages and metrics.

Definition 5 (Verifiability)**.** *Verifiability refers to the degree by which a data consumer can assess the correctness of a dataset and as a consequence its trustworthiness.*

Metrics. Verifiability can be measured either by an unbiased third party, if the dataset itself points to the source or by the presence of a digital signature.

Example. In our use case, if we assume that the flight search engine crawls information from arbitrary airline websites, which publish flight information according to a standard vocabulary, there is a risk for receiving incorrect information from malicious websites. For instance, 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.

Verifiability is an important dimension when a dataset includes sources with low believability or reputation. This dimension allows data consumers to decide whether to accept provided information. One means of verification in Linked Data is to provide basic provenance information along with the dataset, such as using existing vocabularies like SIOC, Dublin Core, Provenance Vocabulary, the OPMV[8] or the recently introduced PROV vocabulary[9]. Yet another mechanism is by the usage of digital signatures [11], whereby a source can sign either a document containing an RDF serialisation or an RDF graph. Using a digital signature, the data source can vouch for all possible serialisations 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.

3.2.2.3. Reputation. The authors in [17] associate reputation of an entity either as a result from direct experience or recommendations from others. They propose the tracking of reputation either through a centralized authority or via decentralized voting.

Definition 6 (Reputation)**.** *Reputation is a judgement made by a user to determine the integrity of a source. It is mainly associated with a data publisher, a person, organisation, group of people or community of practice rather than being a characteristic of a dataset. The data publisher should be identifiable for a certain (part of a) dataset.*

Metrics. Reputation is usually a score, for example, a real value between 0 (low) and 1 (high). There are different possibilities to determine reputation and 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 or by the person who published a dataset. The (semi-)automated approach can be performed by the use of external links or page ranks.

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,

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

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

the search engine can decide to trust only the source with higher reputation.

Reputation is a social notion of trust [19]. 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 judgement about the reputation of the dataset or the publisher and the reliability of the statements.

3.2.2.4. Believability. In [5], believability is explained as "the extent to which information is regarded as true and credible". Believability can also be termed as "trustworthiness" as it is the subjective measure of a users belief that the data is "true" [29].

Definition 7 (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. In Linked Data, believability can be subjectively measured by analyzing the provenance information of the dataset.

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 verify 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]. Either all of the information that is provided can be trusted if it is well-known or only by looking at the data source can the decision of its credibility and believability be determined. 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[10]. Pressing of such a button for any piece of data or an entire dataset would contribute towards assessing the believability of the dataset.

3.2.2.5. Licensing. "In order to enable information consumers to use the data under clear legal terms, each RDF document should contain a license under which the content can be (re-)used" [27,14]. Additionally, the existence of a machine-readable indication (by including the specifications in a VoID description) as well as a human-readable indication of a license is also important. Although in both [27] and [14], the authors do not provide a formal definition, they agree on the use and importance of licensing in terms of data quality.

Definition 8 (Licensing)**.** *Licensing is defined as a granting of the permission for a consumer to re-use a dataset under defined conditions.*

Metrics. Licensing can be checked by the indication of machine and human readable information associated with the dataset clearly indicating the permissions of data re-use.

Example. Since our flight search engine aggregates data from several data sources, a clear indication of the license allows the search engine to re-use the data from the airlines websites. For example, the LinkedGeoData dataset is licensed under the Open Database License[11], 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 engine is able to reuse this dataset to pull geo-spatial information and feed it to the search engine.

Linked Data aims to provide users the capability to aggregate 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 permits it wants to issue. Possible permissions include the reproduction of data, the distribution of data, and the modification and redistribution of data [43]. 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. Additional triples should be added to a dataset clearly indicating the type of license or waiver or license details should be mentioned in the VoID[12] file.

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

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

[12] http://vocab.deri.ie/void

Relations between dimensions. 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. The provenance information associated with a dataset assists a user in verifying the believability of a dataset. Therefore, believability is affiliated to the provenance of a dataset. Moreover, if a dataset has a high reputation, it implies high believability and vice-a-versa. Licensing is also part of the provenance of a dataset and contributes towards its believability. Believability, on the other hand, can be seen as expected accuracy. Moreover, the verifiability, believability and reputation dimensions are also included in the contextual dimensions group (as shown in Figure 2) because they highly depend on the context of the task at hand.

3.2.3. Intrinsic dimensions

Intrinsic dimensions are those that are independent of the user's context. These dimensions focus on whether information correctly represents the real world and whether information is logically consistent in itself. Table 4 lists the five dimensions with their respective metrics which are part of this group, namely, *accuracy*, *objectivity*, *validity-of-documents*, *interlinking*, *consistency* and *conciseness*. The reference for each metric is provided in the table.

3.2.3.1. Accuracy. In [5], accuracy is defined as the "degree of correctness and precision with which information in an information system represents states of the real world". Also, we mapped the problems of inaccurate annotation such as *inaccurate labelling* and *inaccurate classification* mentioned in [38] to the accuracy dimension. In [15] there are two types of accuracy, which were identified: syntactic and semantic. We associate the accuracy dimension mainly to semantic accuracy and the syntactic accuracy to the validity-of-documents dimension.

Definition 9 (Accuracy)**.** *Accuracy can be 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 error. In particular, we associate accuracy mainly to semantic accuracy which relates to the correctness of a value to the actual real world value, that is, accuracy of the meaning.*

Metrics. Accuracy can be measured by checking the correctness of the data in a data source. That is, the detection of outliers or identification of semantically incorrect values through the violation of functional dependency rules. 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 more logical constraints need to be specified for inferring logical contradictions.

Example. In our use case, suppose a user is looking for flights between Paris and New York. Instead of returning flights starting from Paris, France, the search returns flights between Paris in Texas and New York. This kind of semantic inaccuracy in terms of labelling as well as classification can lead to erroneous results.

A possible method for checking accuracy is an alignment with high quality datasets in the domain (reference dataset), if available. Yet another method is by manually checking accuracy against several sources where a single fact is checked individually in different datasets to determine its accuracy [36].

3.2.3.2. Objectivity. In [5], objectivity is expressed as "the extent to which information is unbiased, unprejudiced and impartial."

Definition 10 (Objectivity)**.** *Objectivity is defined as the degree to which the interpretation and usage of data is unbiased, unprejudiced and impartial. This dimension highly depends on the type of information and therefore is classified as a subjective dimension.*

Metrics. Objectivity can not be measured qualitatively but indirectly by checking the authenticity of the source responsible for the information, whether the dataset is neutral or the publisher has a personal influence on the data provided. Additionally, it can be measured by checking whether independent sources can confirm a single fact.

Example. In our use case, consider the reviews available for each airline regarding the safety, comfort and prices for each. 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

[13] Not being what it purports to be; false or fake

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

Dimension	Metric	Description	Type
Accuracy	detection of poor attributes i.e. those that do not contain useful values for data entries	using association rules (using the Apriori Algorithm [1]) or inverse relations or foreign key relationships [38]	O
	detection of outliers	statistical methods such as distance-based, deviations-based and distribution-based method [5]	O
	detection of semantically incorrect values	checking the data source by integrating queries for the identification of functional dependencies violations [15,7]	O
Objectivity	objectivity of the information	checking for bias or opinion expressed when a data provider interprets or analyzes facts [5]	S
	objectivity of the source	checking whether independent sources confirm a fact	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	S
Validity-of-documents	no syntax errors	detecting syntax errors using validators [14]	O
	invalid usage of undefined classes and properties	detection of classes and properties which are used without any formal definition [14]	O
	use of members of deprecated classes or properties	detection of use of OWL classes owl:DeprecatedClass and owl:-DeprecatedProperty [14]	O
	invalid usage of vocabularies	detection of the improper usage of vocabularies [14]	O
	malformed datatype literals	detection of ill-typed literals which do not abide by the lexical syntax for their respective datatype [14]	O
	erroneous[13] annotation / representation	1 - (erroneous instances / total no. of instances) [38]	O
	inaccurate annotation, labelling, classification	(1 - inaccurate instances / total no. of instances) * (balanced distance metric [40] / total no. of instances) [38] (Balanced distance metric is an algorithm that calculates the distance between the extracted (or learned) concept and the target concept)	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
Consistency	entities as members of disjoint classes	no. of entities described as members of disjoint classes / total no. of entities described in the dataset [14]	O
	valid usage of inverse-functional properties	detection of inverse-functional properties that do not describe entities stating empty values [14]	S
	no redefinition of existing properties	detection of existing vocabulary being redefined [14]	S
	usage of homogeneous datatypes	no. of properties used with homogeneous units in the dataset / total no. of properties used in the dataset [14]	O
	no stating of inconsistent property values for entities	no. of entities described inconsistently in the dataset / total no. of entities described in the dataset [14]	O
	ambiguous annotation	detection of an instance mapped back to more than one real world object leading to more than one interpretation [38]	S
	invalid usage of undefined classes and properties	detection of classes and properties used without any formal definition [27]	O
	misplaced classes or properties	detection of a URI defined as a class is used as a property or a URI defined as a property is used as a class [27]	O
	misuse of owl:datatypeProperty or owl:objectProperty	detection of attribute properties used between two resources, and relation properties used with literal values [27]	O
	use of members of deprecated classes or properties	detection of use of OWL classes owl:DeprecatedClass and owl:-DeprecatedProperty [27]	O
	bogus owl:Inverse-FunctionalProperty values	detecting uniqueness & validity of inverse-functional values [27]	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
	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[14]	profiling statistics support the detection of such discordant values or misused predicates and facilitate to find valid formats for specific predicates [7]	O
	negative dependencies/correlation among predicates	using association rules [7]	O
Conciseness	does not contain redundant attributes/properties (intensional conciseness)	number of unique attributes of a dataset in relation to the overall number of attributes in a target schema [42]	O
	does not contain redundant objects/instances (extensional conciseness)	number of unique objects in relation to the overall number of object representations in the dataset [42]	O
	no unique values for functional properties	assessed by integration of uniqueness rules [15]	O
	no unique annotations	check if several annotations refer to the same object [38]	O

Table 4

Comprehensive list of data quality metrics of the intrinsic dimensions, how it can be measured and it's type - "S"ubjective or "O"bjective

can be measured only with quantitative (factual) data. This can be done by checking a single fact individually in different datasets for confirmation [36]. Measuring the bias in qualitative data is far more challenging. However, the bias in information can lead to errors in judgment and decision making and should be avoided.

3.2.3.3. Validity-of-documents. "Validity of documents consists of two aspects influencing the usability of the documents: the valid usage of the underlying vocabularies and the valid syntax of the documents" [14].

Definition 11 (Validity-of-documents)**.** *Validity-of-documents refers to the valid usage of the underlying vocabularies and the valid syntax of the documents (syntactic accuracy).*

Metrics. A syntax validator can be employed to assess the validity of a document, i.e. its syntactic correctness. The syntactic accuracy of entities can be measured by detecting the erroneous or inaccurate annotations, classifications or representations. An RDF validator can be used to parse the RDF document and ensure that it is syntactically valid, that is, to check whether the document is in accordance with the RDF specification.

Example. In our use case, let us assume that the user is looking for flights between two specific locations, for instance Paris (France) and New York (United States). However, the user is returned with no results. A possible reason for this is that one of the data sources incorrectly uses the property `geo:lon` to specify the longitude of "Paris" instead of using `geo:long`. This causes a query to retrieve no data when querying for flights starting near to a particular location.

Invalid usage of vocabularies means referring to not existing or deprecated resources. Moreover, invalid usage of certain vocabularies can result in consumers not able to process data as intended. Syntax errors, typos, use of deprecated classes and properties all add to the problem of the invalidity of the document as the data can neither be processed nor a consumer can perform reasoning on such data.

3.2.3.4. interlinking. When an RDF triple contains URIs from different namespaces in subject and object position, this triple basically establishes a link between the entity identified by the subject (described in the source dataset using namespace A) with the entity identified by the object (described in the target dataset using namespace B). Through the typed RDF links data items are effectively interlinked. The importance of 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 [41].

Definition 12 (interlinking)**.** *interlinking refers to the degree to which entities that represent the same concept are linked to each other.*

Metrics. interlinking can be measured by using network measures that calculate the interlinking degree, cluster coefficient, sameAs chains, centrality and description richness through 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.

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 Linked Data to link or relate these two objects in order to be unambiguous. The interlinking refers to not only the interlinking between different datasets but also internal links within the dataset itself. 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]. An effort towards assessing the quality of a mapping (i.e. incoherent mappings), even if no reference mapping is available, is provided in [41].

3.2.3.5. Consistency. Consistency implies that "two or more values do not conflict with each other" [5]. Similarly, in [14] and [26] consistency is defined as "no contradictions in the data" . A more generic definition is that "a dataset is consistent if it is free of conflicting information" [42].

Definition 13 (Consistency)**.** *Consistency means that a knowledge base is free of (logical/formal) contradictions with respect to particular knowledge representation and inference mechanisms.*

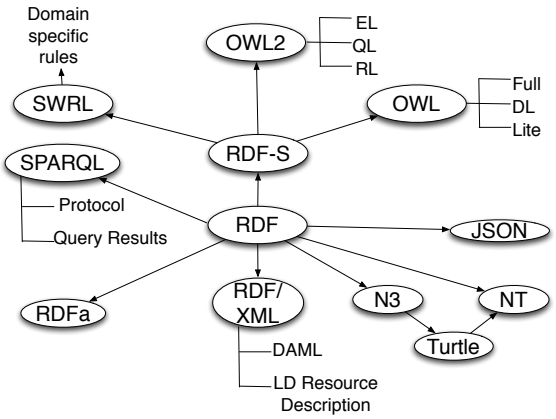

Fig. 3. Different components related to the RDF representation of data.

Metrics. On the Linked Data Web, 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 the OWL2-QL reasoning profile into account. For assessing consistency, we can employ an inference engine or a reasoner, which supports the respective expressivity of the underlying knowledge representation formalism. Additionally, we can detect functional dependency violations such as domain/range violations.

In practice, RDF-Schema inference and reasoning with regard to the different OWL profiles can be used to measure consistency in a dataset. For domain specific applications, consistency rules can be defined, for example, according to the SWRL [28] or RIF standards [32] and processed using a rule engine. Figure 3 shows the different components related to the RDF representation of data where consistency mainly applies to the schema (and the related components) of the data rather than RDF and its representation formats.

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
B123 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 representation, which can be detected by using inference and reasoning.

3.2.3.6. Conciseness. The authors in [42] distinguish the evaluation of conciseness at the schema and the instance level. 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 attributes of a dataset in relation to the overall number of attributes 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 object representations in the dataset. The definition of conciseness is very similar to the definition of ’uniqueness’ defined in [15] as the “degree to which data is free of redundancies, in breadth, depth and scope". This comparison shows that conciseness and uniqueness can be used interchangeably.

Definition 14 (Conciseness)**.** *Conciseness refers to the redundancy of entities, be it at the schema or the data level. Thus, conciseness can be classified into (i) intensional conciseness (schema level) which refers to the redundant attributes and (ii) extensional conciseness (data level) which refers to the redundant objects.*

Metrics. As conciseness is classified in two categories, it can be measured by 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, since data is fused from different datasets, an example of intensional conciseness would be a particular flight, say `A123`, being represented by two different identifiers in different datasets, such as `http://airlines.org/A123` and `http://flights.org/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 repre-

sent the data, then the intensional conciseness is high. However, if the integration leads to the duplication of values, that is the same information is stored in different ways, this leads to extensional conciseness. This may lead to contradictory values and can be solved by fusing duplicate entries and merging common properties.

Relations between dimensions. Both the dimensions accuracy and objectivity focus towards the correctness of representing the real world data. Thus, objectivity overlaps with the concept of accuracy but differs from it because the concept of accuracy does not heavily depend on the consumers' preference. On the other hand, objectivity is influenced by the user's preferences and by the type of information (eg. height of a building vs. product description). Objectivity is also related to the verifiability dimension, that is, the more verifiable a source is, the more objective it will be. Accuracy, in particular the syntactic accuracy of the documents, is related to the validity-of-documents dimension. Although the interlinking dimension is not directly related to the other dimensions, it is included in this group since it is independent of the user's context.

3.2.4. Accessibility

The dimensions belonging to this category involve aspects related to the way data can be accessed and retrieved. There are four dimensions part of this group, which are *availability*, *performance*, *security* and *response-time* as displayed along with their corresponding metrics in Table 5. The reference for each metric is provided in the table.

3.2.4.1. Availability. Availability refers to "the extent to which information is available, or easily and quickly retrievable" [5]. In [14], on the other hand, availability is expressed as the proper functioning of all access methods. In the former definition, availability is more related to the measurement of available information rather than to the method of accessing the information as implied in the latter definition.

Definition 15 (Availability)**.** *Availability of a dataset is the extent to which information is present, obtainable and ready for use.*

Metrics. Availability of a dataset can be measured in terms of accessibility of the server, SPARQL endpoints or RDF dumps and also by the dereferencability of the URIs.

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 is presented with 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 system is less likely to be used, in which case the user may not book flights 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.

3.2.4.2. Performance. Performance is denoted as a quality indicator which "comprises aspects of enhancing the performance of a source as well as measuring of the actual values" [14]. However, in [27], performance is associated with issues such as avoiding 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. In the aforementioned references we can notice that there is no such a formal definition provided for performance. In the former reference the authors give a general description of performance without explaining what is meant by performance. In the latter reference, the authors describe the issue related to performance.

Definition 16 (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, that is a query should be answered in a reasonable amount of time. Also, detection of the usage of prolix RDF features or usage of slash-URIs can help determine the performance of a dataset. Additional metrics are low latency and high throughput of the services provided for the dataset

Example. In our use case, the target performance may depend on the number of users, i.e. it may be required to be able to server 100 simultaneous users. Our flight search engine will not be scalable if the time re-

Dimension	Metric	Description	Type
Availability	accessibility of the server	checking whether the server responds to a SPARQL query [14, 26]	O
	accessibility of the SPARQL endpoint	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 URIs returns an error (4xx client error/ 5xx server error) response code or detection of broken links [14,26]	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
	misreported content types	detection of whether the content is suitable for consumption, and whether the content should be accessed [26]	S
	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
Performance	no usage of slash-URIs	checking for usage of slash-URIs where large amounts of data is provided [14]	O
	low latency	if an HTTP-request is not answered within an average time of one second, the latency of the data source is considered too low [14]	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	O
	no use of prolix RDF features	detect use of RDF primitives i.e. RDF reification, RDF containers and RDF collections [27]	O
Security	access to data is secure	use of login credentials or use of SSL or SSH	O
	data is of proprietary nature	data owner allows access only to certain users	O
Response-time	delay in response time	delay between submission of a request by the user and reception of the response from the system [5]	O

Table 5

Comprehensive list of data quality metrics of the accessibility dimensions, how it can be measured and it's type - "S"ubjective or "O"bjective

quired to answer to all queries is similar to the time required when querying the individual datasets. In that case, satisfying performance needs requires caching mechanisms.

Latency is the amount of time from issuing the query until the first information reaches the user. Achieving 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 and (iii) avoiding the use of prolix RDF features. Since Linked Data may involve 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 accessibility.

3.2.4.3. Security. Security refers to "the possibility to restrict access to the data and to guarantee the confidentiality of the communication between a source and its consumers" [14].

Definition 17 (Security)**.** *Security can be defined as the extent to which access to data can be restricted and hence protected against its illegal alteration and misuse. It refers to the degree to which information is passed securely from users to the information source and back.*

Metrics. Security can be 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, we consider a user that 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. 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 can be used as a means to keep the information secure especially in cases of medical and 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 can be applied [54]. Although security is an important dimension, it does not often apply to open data. The importance of the security depends on whether the data needs to be protected.

3.2.4.4. Response-time. Response-time is defined as "that which measures the delay between submission of a request by the user and reception of the response from the system" [5].

Definition 18 (Response-time)**.** *Response-time measures the delay, usually in seconds, between submission of a query by the user and reception of the complete response from the dataset.*

Metrics. Response-time can be assessed by measuring the delay between submission of a request by the user and reception of the response from the dataset

Example: A user is looking for a flight which includes multiple destinations. She wants to fly from `Milan` to `Boston`, `Boston` to `New York` and `New York` to `Milan`. In spite of the complexity of the query the search engine should respond quickly with all the flight details (including connections) for all the destinations.

The response time 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. 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. Locally replicating or caching information are possible means of improving the response time. Another option is by providing low latency, so that the user is provided with a part of the results early on.

Relations between dimensions. The dimensions in this group are inter-related with each other as follows: performance of a system is related to the availability and response-time dimensions. Only if a dataset is available or has low response time, it can perform well. Security is also related to the availability of a dataset because the methods used for restricting users is tied to the way a user can access a dataset.

3.2.5. Representational dimensions

Representational dimensions capture aspects related to the design of the data such as the *representational-conciseness*, *representational-consistency*, *understandability*, *versatility* as well as the *Interpretability* of the data. These dimensions along with their corresponding metrics are listed in Table 6. The reference for each metric is provided in the table.

3.2.5.1. Representational-conciseness. Representational-conciseness is only defined as "the extent to which information is compactly represented" [5].

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

Metrics. Representational-conciseness is measured by qualitatively verifying whether the RDF model that is used to represent the data is concise enough in order to be self-descriptive and unambiguous.

Example. A user, after booking her flight, is interested in additional information about the destination airport such as its location. Our flight search engine should provide only that information related to the location rather than returning a chain of other properties.

Representation of RDF data in N3 format is considered to be more compact than RDF/XML [14]. The concise representation not only contributes to the human readability of the data but also influences the performance of data when queried. For example, in [27], the use of very long URIs or those that contain query parameters is an issue related to the representational-conciseness. Keeping URIs short and human readable is highly recommended for large scale and/or frequent processing of RDF data as well as for efficient indexing and serialisation.

Dimension	Metric	Description	Type
Representational-conciseness	keeping URIs short	detection of long URIs or those that contain query parameters [27]	O
Representational-consistency	re-use existing terms	detecting of whether existing terms from other vocabularies have been reused [27]	O
	re-use existing vocabularies	usage of established vocabularies [14]	S
Understandibility	human-readable labelling of classes, properties and entities by providing rdfs:label	no. of entities described by stating an rdfs:label or rdfs:comment in the dataset / total no. of entities described in the data [14]	O
	indication of metadata about a dataset	checking for the presence of the title, content and URI of the dataset [14,5]	O
	dereferenced representations: giving human readable metadata	detecting 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	interpretability of data	detect the use of appropriate language, symbols, units and clear definitions [5]	S
		detect the use of self-descriptive formats, identifying objects and terms used to define the objects with globally unique identifiers [5]	O
	interpretability of terms	use of various schema languages to provide definitions for terms [5]	S
	misinterpretation of missing values	detecting use of blank nodes [27]	O
	dereferenced representations: giving human readable metadata	detecting use of rdfs:label to attach labels or names to resources [27]	O
	atypical use of collections, containers and reification	detection of the non-standard usage of collections, containers and reification [27]	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 [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
	accessing of data in different ways	checking whether the data is available as a SPARQL endpoint and is available for download as an RDF dump [14]	O

Table 6

Comprehensive list of data quality metrics of the representational dimensions, how it can be measured and it's type - "S"ubjective or "O"bjective

3.2.5.2. Representational-consistency. Representational-consistency is defined as "the extent to which information is represented in the same format" [5]. 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 [14]. As stated in [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 conform to previously returned information. Since Linked Data involves aggregation of data from multiple sources, we extend this definition to not only imply compatibility with previous data but also with data from other sources.*

Metrics. Representational consistency can be assessed by detecting whether the dataset re-uses exist-

ing vocabularies or terms from existing established vocabularies to represent its entities.

Example. In our use case, consider different airlines companies using different notation for representing their data, e.g. some use RDF data and some others use turtle. In order to avoid interoperability issue, we provide data based on the Linked Data principle which is designed to support heterogeneous description models, which is necessary to handle different format of data. The exchange of information in different formats will not be a big deal in our search engine since strong links are created between datasets.

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[15] should be the vocabulary of choice. Moreover, re-using vocabularies maximises 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. Even though there is no central repository of existing vocabularies, suitable terms can be found in SchemaWeb[16], SchemaCache[17] and Swoogle[18]. Additionally, a comprehensive survey done in [52] lists a set of naming conventions that should be used to avoid inconsistencies[19]. Another possibility is to use LODStats [13], which allows to perform a search for frequently used properties and classes in the LOD cloud.

3.2.5.3. Understandibility. Understandability is defined as the "extent to which data is easily comprehended by the information consumer" [5]. Understandability is also related to the comprehensibility of data i.e. the ease with which human consumers can understand and utilize the data [14]. Thus, the dimensions understandability and comprehensibility can be interchangeably used.

Definition 21 (Understandability)**.** *Understandability refers to the ease with which data can be comprehended, without ambiguity, and used by a human consumer. Thus, this dimension can also be referred to as the comprehensibility of the information where the data should be of sufficient clarity in order to be used.*

Metrics. Understandability can be measured by detecting whether human-readable labels for classes, properties and entities are provided. Provision of the metadata of a dataset can also contribute towards assessing its understandability. The dataset should also clearly provide exemplary URIs and SPARQL queries along with the vocabularies used so that can users can understand how it can be used.

Example. Let us assume that in our flight search engine, it allows a user to enter a start and destination address. In that case, strings entered by the user need to be matched to entities in the spatial dataset4, probably via string similarity. Understandable labels for cities, places etc. improve search performance in that case. For instance, when a user looks for a flight to U.S (label), then the search engine should return the flights to the United States or America.

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

3.2.5.4. Interpretability. Interpretability is defined as the "extent to which information is in appropriate languages, symbols, and units, and the definitions are clear" [5].

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 can be measured by the use of globally unique identifiers for objects and terms or by the use of appropriate language, symbols, units and clear definitions

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

[15] `http://xmlns.com/foaf/spec/`
[16] `http://www.schemaweb.info/`
[17] `http://schemacache.com/`
[18] `http://swoogle.umbc.edu/`
[19] However, they only restrict themselves to only considering the needs of the OBO foundry community but still can be applied to other domains

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 the information that Freebase contains information about Boston Logan Airport. The third, however, contains a human-readable label, which the user can easily interpret.

The more interpretable a Linked Data source is, the more easy 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.

3.2.5.5. Versatility. In [14], versatility is defined to as the "alternative representations of the data and its handling."

Definition 23 (Versatility)**.** *Versatility mainly refers to the alternative representations of data and its subsequent handling. Additionally, versatility also corresponds to the provision of alternative access methods for a dataset.*

Metrics. Versatility can be measured by the availability of the dataset in different serialisation formats, different languages as well as different access methods.

Example. Consider a user from a non-English speaking country who wants to use our flight search engine. In order to cater to the needs of such users, our flight search engine should be available in different languages so that any user has the capability to understand it.

Provision of Linked Data 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.

Relations between dimensions. Understandability is related to the interpretability dimension as it refers to the subjective capability of the information consumer to comprehend information. Interpretability mainly refers to the technical aspects of the data, that is if it is correctly represented. Versatility is also related to the interpretability of a dataset as the more versatile forms a dataset is represented in (for eg. in different languages), the more interpretable a dataset is. Although representational-consistency is not related to any of the other dimensions in this group, it is part of the representation of the dataset. In fact representational-consistency is related to the validity-of-documents, an intrinsic dimension, because the invalid usage of vocabularies may lead to inconsistency in the documents.

3.2.6. Dataset Dynamicity

An important aspect of data is its update over time. The main dimensions related to the dynamicity of a dataset proposed in the literature are *currency*, *volatility*, and *timeliness*. Table 7 provides these three dimensions with their respective metrics. The reference for each metric is provided in the table.

3.2.6.1. Currency. Currency refers to "the age of data given as a difference between the current date and the date when the information was last modified" by providing both currency of documents and currency of triples [50]. Similarly the definition given in [42] describes currency as "the distance between the input date from the provenance graph to the current date". According to these definitions, currency only measures the time since the last modification, whereas we provide a definition that measures the time between a change in the real world and a change in the knowledge base.

Definition 24 (Currency)**.** *Currency refers to the speed with which the information (state) is updated after the real-world information changes.*

Metrics. The measurement of currency relies on two components: (i) delivery time (the time when the data was last modified) and (ii) the current time, both possibly present in the data models.

Example. Consider a user is looking for the price of a flight from Milan to Boston and she receives the updates via email. The currency of the price information is measured with respect to the last update of the price information. The email service sends the email with a

[20] identify the time when first data have been published in the LOD
[21] elder than the last modification time

Dimension	Metric	Description	Type
Currency	currency of statements	1- [(current time - last modified time)/(current time - start time[20])] [50]	O
	currency of data source	(last modified time of the semantic web source < last modified time of the original source)[21] [15]	O
	age of data	current time - created time [42]	O
Volatility	no timestamp associated with the source	check if there are temporal meta-information associated to a source [38]	O
	no timestamp associated with the source	check if there are temporal meta-information associated to a source	O
Timeliness	timeliness of data	expiry time < current time [15]	O
	stating the recency and frequency of data validation	detection of whether the data published has been validated no more than a month ago [14]	O
	no inclusion of outdated data	no. of triples not stating outdated data in the dataset / total no. of triples in the dataset [14]	O
	time inaccurate representation of data	1 - (time-inaccurate instances / total no. of instances) [14]	O

Table 7

Comprehensive list of data quality metrics of the dataset dynamicity dimensions, how it can be measured and it's type - "S"ubjective or "O"bjective

delay of about 1 hour with respect to the time interval in which the information is determined to hold. In this way, if the currency value exceeds the time interval of the validity of the information, the result is said to be not current. If we suppose validity of the price information (the frequency of change of the price information) to be 2 hours, the price list that is not changed within this time is considered to be out-dated. To determine whether the information is out-dated or not we need to have temporal metadata available and represented by one of the data models proposed in the literature [49].

3.2.6.2. Volatility. Since there is no definition for volatility in the core set of approaches considered in this survey, we provide one which applies in the Web of Data.

Definition 25 (Volatility)**.** *Volatility can be defined as the length of time during which the data remains valid.*

Metrics. Volatility can be measured by 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 published on the Web). Both these metrics are combined together to measure the distance between the expiry time and the input time of the published data.

Example. Let us consider the aforementioned use case where a user wants to book a flight from Milan to Boston and she is interested in the price information. The price is considered as volatile information as it changes frequently, for instance it is estimated that the flight price changes each minute (data remains valid for one minute). In order to have an updated price list, the price should be updated within the time interval pre-defined by the system (one minute). In case the user observes that the value is not re-calculated by the system within the last few minutes, the values are considered to be out-dated. Notice that volatility is estimated based on changes that are observed related to a specific data value.

3.2.6.3. Timeliness. Timeliness is defined as "the degree to which information is up-to-date" in [5], whereas in [14] the author define the timeliness criterion as "the currentness of the data provided by a source".

Definition 26. *Timeliness refers to the time point at which the data is actually used. This can be interpreted as whether the information is available in time to be useful.*

Metrics. Timeliness is measured by combining the two dimensions: currency and volatility. Additionally timeliness states the recency and frequency of data validation and does not include outdated data.

Example. A user wants to book a flight from Milan to Boston and thus our flight search engine will return a list of different flight connections. She picks one of them and follows all the steps to book her flight. The data contained in the airline dataset shows a flight company that is available according to the user requirements. In terms of time-related quality dimension, the information related to the flight is recorded and reported to the user every two minutes which fulfils the requirement decided by our search engine that cor-

responds to the volatility of a flight information. Although the flight values are updated on time, the information received to the user about the flight's availability is not on time. In other words, the user did not perceive that the availability of the flight was depleted because the change was provided to the system a moment after her search.

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 to be relevant, given the rate of change of the source value and the user's domain and purpose of interest.

3.2.6.4. Relations between dimensions. Timeliness, although part of the dataset dynamicity group, is also considered to be part of the intrinsic group because it is independent of the users context. In [2], the authors compare the definitions provided in the literature for these three dimensions and identify that often the definitions given for currency and timeliness can be used interchangeably. Thus, timeliness depends on both the currency and volatility dimensions. Currency is related to volatility since currency of data depends on how volatile it is. Thus, data that is highly volatile must be current, while currency is less important for data with low volatility.

4. Comparison of selected approaches

In this section, we compare the 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 4.1), their respective metrics (Section 4.2), types of data (Section 4.3), level of automation (Section 4.4) and usability of the three specific tools (Section 4.5).

4.1. Dimensions

The Linked Open Data paradigm is the fusion of three different research 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 [55,57,48,30,9,46]. 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 provided by most authors [51]. However, no consensus exists on which set of dimensions might define 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 to the consumer. We therefore gathered all data quality dimensions that have been reported as being relevant for LOD by analyzing the selected approaches. An initial list of data quality dimensions was first obtained from [5]. Thereafter, the problem being addressed in 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 present in 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. We therefore obtained a further set of quality dimensions from [14], which was one of the first studies focusing towards data quality dimensions and metrics applicable to LOD. Yet there were some problems that did not fit in this extended list of dimensions such as the problem of 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 26 Linked Data quality dimensions along with their respective frequency of occurrence in the included approaches. This table presents the information split into three distinct groups: (a) a set of approaches focusing only on dataset provenance [23,16,53,20,18,21,17,29,8]; (b) a set of approaches covering more than five dimensions [6,14,27,42] and (c) a set of approaches focusing on very few and specific dimensions [7,12,22,26, 38,45,15,50]. Overall, it can be observed that the dimensions provenance, consistency, timeliness, accuracy and completeness are the most frequently used.

We can also conclude that none of the approaches cover all data quality dimensions that are relevant for LOD.

4.2. Metrics

As defined in Section 3, a data quality metric is *a procedure for measuring an information quality dimension*. We notice that most of metrics take the form of a ratio, which measures the desired outcomes to the total outcomes [38]. For example, for the representational-consistency dimension, the metric for determining the re-usage of existing vocabularies takes the form of a ratio as:

$$\frac{\textit{no. of established vocabularies used in the dataset}}{\textit{total no. of vocabularies used in the dataset}}$$

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

For some of the included approaches, the problem, its corresponding metric, and a dimension were clearly mentioned [14,5]. However, for the other approaches, we first extracted the problem addressed along with the way in which it was assessed (metric). Thereafter, 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 present 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.

We observed that for a particular dimension there can be several metrics associated with it but one metric is not associated with more than one dimension. Additionally, there are several ways of measuring one dimension either individually or by combining different metrics. As an example, the availability dimension can be measured by a combination of three other metrics, namely accessibility of the (i) server, (ii) SPARQL end-point, and (iii) RDF dumps. Additionally, availability can be individually measured by the availability of structured data, misreported content types, or by the absence of dereferencability issues.

We also classify each metric as being *Objectively* (quantitatively) or *Subjectively* (qualitatively) assessed. Objective metrics are those which can be 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 can be quantified. On the other hand, subjective dimensions are those which cannot be quantified but depend on the users perception of the respective dimension (via surveys). For example, metrics belonging to dimensions such as objectivity, relevancy highly depend on the user and can only be qualitatively measured.

The ratio form of the metrics can be generally applied to those metrics which can be measured quantitatively (objectively). 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 can be measured both objectively as well as subjectively. For example, although completeness is perceived as that which can be measured objectively, it also includes metrics which can be subjectively measured. That is, the schema or ontology completeness can be measured subjectively whereas the property, instance and interlinking completeness can be 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 can be measured objectively but on the other hand, the scope and level of detail can be measured subjectively.

4.3. Type of data

The goal of an 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 can be associated from small-scale units of data such as assessment of RDF triples to the assessment of entire datasets which potentially affect the assessment process. In LOD, we distinguish the assessment process operating on three types of data:

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

Approaches / Dimensions	Bizer et.al.,2009	Flemming et.al.,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 et.al, 2006	Gil et.al., 2002	Golbeck et. al., 2003	Gil et.al., 2007	Jacobi et.al., 2011	Bonatti et. al., 2011
Provenance		✔											✔	✔	✔	✔	✔	✔	✔	✔	✔
Consistency	✔	✔	✔			✔		✔	✔	✔											
Currency									✔		✔	✔									
Volatility								✔													
Timeliness		✔									✔										
Accuracy			✔					✔			✔										
Completeness					✔	✔			✔		✔										
Amount-of-data		✔		✔																	
Availability		✔					✔														
Understandability		✔																			
Relevancy																					
Reputation									✔										✔		
Verifiability		✔																			
Interpretability							✔														
Rep.-conciseness							✔														
Rep.-consistency		✔					✔														
Licensing		✔					✔														
Performance							✔														
Objectivity	✔																				
Believability	✔																				
Response-time	✔																				
Security	✔																				
Versatility	✔	✔																			
Validity-of-documents		✔																			
Conciseness								✔	✔		✔										
Interlinking					✔		✔														

- 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.

We can observe that most of the methods are applicable at the triple or graph level and less on the dataset level (Table 9). Additionally, it is seen that 9 approaches assess data both at triple and graph level [18,45,38,7,12,14,15,8,50] , 2 approaches assess data both at graph and dataset level [16,22] and 4 approaches assess data at triple, graph and dataset levels [20,6,26,27]. There are 2 approaches that apply the assessment only at triple level [23,42] and 4 approaches that only apply at the graph level [21,17,53, 29].

However, 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, the assessment can be performed at the graph level which can be further 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) can be propagated to the statements (triple level) that are part of the Web source associated with that trust rating [53].

However, there are no approaches that performs assessment only at the dataset level (cf. Table 9). This possible reason is that the assessment of a dataset should always pass through the assessment of a fine grained level such as triple or entity level and then should be propagated to the dataset level.

4.4. Level of automation

Out of the 21 selected approaches, 9 provide tool support for the assessment of data quality. These tools implement the methodologies and metrics defined in the respective approaches. However, due to the nature of the dimensions and related metrics some of the involved tools are either semi or fully automated or sometimes only manually applicable. Table 9 shows the level of automation for each of the identified tools. The automation level is determined by the amount of user involvement.

Automated. The tool proposed in [22] is fully automated as there is no user involvement. The tool 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 quality assessment reports.

Manual. The WIQA [6] and Sieve [42] tools are entirely manual as they require a high degree of user involvement. The WIQA Information Quality Assessment Framework enables users to filter information using a wide range of quality based information filtering policies. Similarly, using Sieve a user can define relevant metrics and respective scoring functions for their specific quality assessment task. This definition of metrics has to be done by creating an XML file which contains the specific configurations for a quality assessment task. Although it gives the user the flexibility of tweaking the tool to match their needs, it involves a lot of time and understanding of the required XML file structure as well as specification.

Semi-automated. The different tools introduced in [14,26,7,18,23,21] are all semi-automated as they involve a minimum amount of user involvement. Flemming's Data Quality Assessment Tool[22] [14] 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. The RDF Validator[23] used in [26] checks RDF documents and reports any data quality problems that might exist. The tool developed in [7] enables users to explore a set of pre-determined data clusters for further investigation.

The TRELLIS user interface [18] allows several users to express their trust on a data source and their respective data. Decisions made by users on a particular source are stored as annotations, which can be used to analyze conflicting information or handle incomplete information. The tRDF[24] approach [23] provides a framework to deal with the trustworthiness of RDF data with a trust-aware extension tSPARQL along with implementation strategies. In all of these components, a user is able to interact with the data as well as the framework.

The approach proposed by [21] provides two applications, namely, TrustBot and TrustMail. TrustBot is an IRC bot that makes trust recommendations to users

[22] http://linkeddata.informatik.hu-berlin.de/LDSrcAss/

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

[24] http://trdf.sourceforge.net/

Fig. 4. Excerpt of the Flemmings Data Quality Assessment tool showing the result of assessing the quality of LinkedCT with a score of 15 out of 100.

based on the trust network it builds. Users have the flexibility to add their own URIs to the bot at any time while incorporating the data into a graph. TrustMail, on the other hand, is a configurable email client which displays the trust level that can be placed on the sender for each email message, be it at a general level or with regard to a certain topic.

Yet another aspect is the amount of knowledge required from users before they can use a tool. Although the tool proposed in [22] is fully automated, it requires a certain level of user understanding in order to interpret its results. The WIQA and Sieve tools require a deep understanding of not only the resources users want to access, but also how the tools work and their configuration. The above-mentioned semi-automated tools require a fair amount of knowledge from the user regarding the assessed datasets.

4.5. Comparison of tools

In this section, we analyze three tools, namely Flemming's data quality tool [14], *Sieve* [42] and *LODGRefine* to assess their usability for data quality assessment. In particular, we compare them with regard to ease of use, level of user interaction, and applicability in terms of data quality assessment, also discussing their pros and cons.

4.5.1. Flemming's Data Quality Assessment Tool

Flemming's data quality assessment tool [14] is a simple user interface[25], where a user first specifies the name, URI, and three entities for a particular data source. Users are then receive a score ranging from 0 to 100 indicating the quality of the dataset, where 100 represents the best quality.

After specifying dataset details (endpoint, graphs, example URIs), the user is given an option for assigning weights to each of the pre-defined data quality metrics. Two options are available for assigning weights: (a) assigning a weight of 1 to all the metrics or (b) choosing the pre-defined exemplary weight of the metrics defined for a semantic data source. In the next step, the user is asked to answer a series of questions regarding the datasets, which are important indicators of the data quality for Linked Datasets and those which cannot be quantified. These include, for example, questions about the use of stable URIs, the number of obsolete classes and properties, and whether the dataset provides a mailing list. Next, the user is presented with a list of dimensions and metrics, for each of which weights can be specified again. Each metric is pro-

[25] Available in German only at: `http://linkeddata.informatik.hu-berlin.de/LDSrcAss/datenquelle.php`

Table 9
Qualitative evaluation of frameworks

Paper	Application	Goal	Type of data			Degree of automation			Tool support
	General/ **S**pecific		**RDF Triple**	**RDF Graph**	**Dataset**	**Manual**	**Semi-automated**	**Automated**	
Gil et.al., 2002	G	Approach to derive an assessment of a data source based on the annotations of many individuals	✔	✔	-	-	✔	-	✔
Golbeck et.al., 2003	G	Trust networks on the semantic web	-	✔	-	-	✔	-	✔
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	✔	-	-	-	✔	-	✔
Bizer et.al., 2009	G	Information filtering	✔	✔	✔	✔	-	-	✔
Böhm et.al., 2010	G	Data integration	✔	✔	-	-	✔	-	✔
Chen et.al., 2010	G	Generating semantically valid hypothesis	✔	✔	-	-	-	-	-
Flemming et.al., 2010	G	Assessment of published data	✔	✔	-	-	✔	-	✔
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	S	Assessment of quality of links	-	✔	✔	-	-	✔	✔
Hogan et.al., 2012	G	Assessment of published data	✔	✔	✔	-	-	-	-
Mendes et.al., 2012	S	Data integration	✔	-	-	✔	-	-	✔
Rula et.al., 2012	G	Assessment of time related quality dimensions	✔	✔	-	-	-	-	-

vided with two input fields: one showing the assigned weights and another with the calculated value.

In the end, the user is presented with a score ranging from 0 to 100 representing the final data quality score. Additionally, the rating of each dimension and the total weight (based on 11 dimensions) is calculated using the user input from the previous steps. Figure 4 shows an excerpt of the tool showing the result of assessing the quality of LinkedCT with a score of 15 out of 100.

On one hand, the tool is easy to use with the form-based questions and adequate explanation for each step. Also, the assigning of the weights for each metric and the calculation of the score is straightforward and easy to adjust for each of the metrics. However, this tool has a few drawbacks: (1) the user needs to have adequate knowledge about the dataset in order to correctly assign weights for each of the metrics; (2) it does not drill down to the root cause of the data quality problem and (3) some of the main quality dimensions are missing from the analysis such as accuracy, completeness, provenance, consistency, conciseness and relevancy as some could not be quantified and were not perceived to be true quality indicators.

4.5.2. Sieve

Sieve is a component of the *Linked Data Integration Framework* (LDIF)[26] used first to assess the quality between two or more data sources and second to fuse (integrate) the data from the data sources based on their quality assessment. In order to use this tool, a user needs to be conversant with programming.

The input of Sieve is an LDIF provenance metadata graph generated from a data source. Based on this information the user needs to set the configuration property in an XML file known as `integration properties`. The quality assessment procedure relies on the measurement of metrics chosen by the user where each metric applies a scoring function having a value from 0 to 1.

Sieve implements only a few scoring functions such as `TimeCloseness`, `Preference`, `SetMembership`, `Threshold` and `Interval Membership` which are calculated based on the metadata provided as input along with the original data source. The configuration file is in XML format which should be modified based on the use case, as shown in Listing 1. The output scores are then used to fuse the data sources by applying one of the fusion functions, which are: `Filter`, `Average`, `Max`, `Min`, `First`, `KeepSingleValue ByQualityScore`, `Last`, `Random`, `PickMostFrequent`.

```
1  <Sieve>
2   <QualityAssessment>
3    <AssessmentMetric id="sieve:recency">
4    <ScoringFunction class="TimeCloseness">
5      <Param name="timeSpan" value="7"/>
6      <Input path="?GRAPH/provenance:lasUpdated"/>
7    </ScoringFunction>
8    </AssessmentMetric>
9    <AssessmentMetric id="sieve:reputation">
10   <ScoringFunction class="ScoredList">
11     <Param name="priority" value="http://pt.
           wikipedia.org http://en.wikipedia.org"/>
12     <Input path="?GRAPH/provenance:lasUpdated"/>
13   </ScoringFunction>
14   </AssessmentMetric>
15 </Sieve>
```

Listing 1: A configuration of Sieve: a Data Quality Assessment and Data Fusion tool.

In addition, users should specify the DataSource folder, the `homepage` element that refers to the data source from which the entities are going to be fused. Second, the XML file of the `ImportJobs` that downloads the data to the server should also be modified. In particular, the user should set up the `dumpLocation` element as the location of the dump file.

Although the tool is very useful overall, there are some drawbacks that decreases its usability: (1) the tool is not mainly used to assess data quality for a source, but instead to perform data fusion (integration) based on quality assessment. Therefore, the quality assessment can be considered as an accessory that leverages the process of data fusion through evaluation of few quality indicators; (2) it does not provide a user interface, ultimately limiting its usage to end-users with programming skills; (3) its usage is limited to domains providing provenance metadata associated with the data source.

4.5.3. LODGRefine.

LODGRefine[27] is a LOD-enabled version of Google Refine, which is an open-source tool for refining messy data. Although this tool is not focused on data quality assessment per se, it is powerful in performing preliminary cleaning or refining of raw data.

Using this tool, one is able to import several different file types of data (CSV, Excel, XML, RDF/XML, N-Triples or even JSON) and then perform cleaning action via a browser-based interface. By using a diverse set of filters and facets on individual columns,

[26] `http://ldif.wbsg.de/`

[27] `http://code.zemanta.com/sparkica/`

LODGRefine can help a user to semi-automate the cleaning of her data. For example, this tool can help to detect duplicates, discover patterns (e.g. alternative forms of an abbreviation), spot inconsistencies (e.g. trailing white spaces) or find and replace blank cells. Additionally, this tool allows users to reconcile data, that is to connect a dataset to existing vocabularies such that it gives meaning to the values. Reconciliations to *Freebase*[28] helps mapping ambiguous textual values to precisely identified Freebase entities. Reconciling using *Sindice* or based on standard SPARQL or SPARQL with full-text search is also possible[29] using this tool. Moreover, it is also possible to extend the reconciled data with DBpedia as well as export the data as RDF, which adds to the uniformity and usability of the dataset.

These feature thus assists in assessing as well as improving the data quality of a dataset. Moreover, by providing external links, the interlinking of the dataset is considerably improved. LODGRefine is easy to download and install as well as to upload and perform basic cleansing steps on raw data. The features of reconciliation, extending the data with DBpedia, transforming and exporting the data as RDF are added advantages. However, this tool has a few drawbacks: (1) the user is not able to perform detailed high level data quality analysis utilizing the various quality dimensions using this tool; (2) performing cleansing over a large dataset is time consuming as the tool follows a column data model and thus the user must perform transformations per column.

5. 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 level of automation.

We survey 21 approaches and extracted 26 data quality dimensions along with their definitions and corresponding metrics. We also identified tools proposed for each approach and classified them in relation to data type, the degree of automation and required level of user knowledge. Finally, we evaluated three specific tools in terms of usability for data quality assessment.

As we can observe, most of the publications focusing on data quality assessment in Linked Data are presented at either conferences or workshops. As our literature review reveals, the number of 21 publications published in the span of 10 years is rather low. This can be attributed to the infancy of this research area, which is currently emerging.

In most of the surveyed literature, the metrics were often not explicitly defined or did not consist of precise statistical measures. Also, there was no formal validation of the methodologies that were implemented as tools. Moreover, only few approaches were actually accompanied by an implemented tool. That is, out of the 21 approaches, only 9 provided tool support. 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 revealed, 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 and others were easy-to-use but provided only results with limited usefulness or required a high-level of interpretation.

Meanwhile, there is quite much research on data quality being done and guidelines as well as recommendations on how to publish "good" data are 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.

[28] `http://www.freebase.com/`
[29] `http://refine.deri.ie/reconciliationDocs`

References

[1] AGRAWAL, R., AND SRIKANT, R. Fast algorithms for mining association rules in large databases. In *International Conference on Very Large Data Bases* (1994), pp. 487–499.

[2] BATINI, C., CAPPIELLO, C., FRANCALANCI, C., AND MAURINO, A. Methodologies for data quality assessment and improvement. *ACM Computing Surveys 41*, 3 (2009).

[3] BATINI, C., AND SCANNAPIECO, M. *Data Quality: Concepts, Methodologies and Techniques (Data-Centric Systems and Applications)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.

[4] BECKETT, D. RDF/XML Syntax Specification (Revised). Tech. rep., World Wide Web Consortium, 2004. http://www.w3.org/TR/2004/REC-rdf-syntax-grammar-20040210/.

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

[6] BIZER, C., AND CYGANIAK, R. Quality-driven information filtering using the wiqa policy framework. *Web Semantics 7*, 1 (Jan 2009), 1 – 10.

[7] BÖHM, C., NAUMANN, F., ABEDJAN, Z., FENZ, D., GRÜTZE, T., HEFENBROCK, D., POHL, M., AND SONNABEND, D. Profiling linked open data with prolod. In *ICDE Workshops* (2010), IEEE, pp. 175–178.

[8] BONATTI, P. A., HOGAN, A., POLLERES, A., AND SAURO, L. Robust and scalable linked data reasoning incorporating provenance and trust annotations. *Journal of Web Semantics 9*, 2 (2011), 165 – 201.

[9] BOVEE, M., SRIVASTAVA, R. P., AND MAK, B. A conceptual framework and belief-function approach to assessing overall information quality. *International Journal of Intelligent Systems 18*, 1 (2003), 51–74.

[10] BRICKLEY, D., AND GUHA, R. V. Rdf vocabulary description language 1.0: Rdf schema. Tech. rep., W3C, 2004. http://www.w3.org/TR/2004/REC-rdf-schema-20040210/.

[11] CARROLL, J. Signing rdf graphs. Tech. rep., HPL-2003-142, HP Labs, 2003.

[12] CHEN, P., AND GARCIA, W. Hypothesis generation and data quality assessment through association mining. In *IEEE ICCI* (2010), IEEE, pp. 659–666.

[13] DEMTER, J., AUER, S., MARTIN, M., AND LEHMANN, J. Lodstats – an extensible framework for high-performance dataset analytics. In *EKAW* (2012), LNCS, Springer.

[14] FLEMMING, A. Quality characteristics of linked data publishing datasources. Master's thesis, Humboldt-Universität zu Berlin, 2010.

[15] FÜRBER, C., AND HEPP, M. Swiqa - a semantic web information quality assessment framework. In *ECIS* (2011).

[16] GAMBLE, M., AND GOBLE, C. Quality, trust, and utility of scientific data on the web: Towards a joint model. In *ACM WebSci* (June 2011), pp. 1–8.

[17] GIL, Y., AND ARTZ, D. Towards content trust of web resources. *Web Semantics 5*, 4 (December 2007), 227 – 239.

[18] GIL, Y., AND RATNAKAR, V. Trusting information sources one citizen at a time. In *ISWC* (2002), Springer-Verlag, pp. 162 – 176.

[19] GOLBECK, J. Inferring reputation on the semantic web. In *WWW* (2004).

[20] GOLBECK, J. 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] GOLBECK, J., PARSIA, B., AND HENDLER, J. Trust networks on the semantic web. In *Cooperative Intelligent Agents* (2003).

[22] GUÉRET, C., GROTH, P., STADLER, C., AND LEHMANN, J. Assessing linked data mappings using network measures. In *ESWC* (2012).

[23] HARTIG, O. Trustworthiness of data on the web. In *STI Berlin and CSW PhD Workshop, Berlin, Germany* (2008).

[24] HAYES, P. Rdf semantics. Recommendation, World Wide Web Consortium, 2004. http://www.w3.org/TR/2004/REC-rdf-mt-20040210.

[25] HEATH, T., AND BIZER, C. *Linked Data: Evolving the Web into a Global Data Space*, 1st ed. No. 1:1 in Synthesis Lectures on the Semantic Web: Theory and Technology. Morgan and Claypool, 2011, ch. 2, pp. 1 – 136.

[26] HOGAN, A., HARTH, A., PASSANT, A., DECKER, S., AND POLLERES, A. Weaving the pedantic web. In *LDOW* (2010).

[27] HOGAN, A., UMBRICH, J., HARTH, A., CYGANIAK, R., POLLERES, A., AND DECKER, S. An empirical survey of linked data conformance. *Journal of Web Semantics* (2012).

[28] HORROCKS, I., PATEL-SCHNEIDER, P., BOLEY, H., TABET, S., GROSOF, B., AND DEAN, M. Swrl: A semantic web rule language combining owl and ruleml. Tech. rep., W3C, May 2004.

[29] JACOBI, I., KAGAL, L., AND KHANDELWAL, A. Rule-based trust assessment on the semantic web. In *International conference on Rule-based reasoning, programming, and applications series* (2011), pp. 227 – 241.

[30] JARKE, M., LENZERINI, M., VASSILIOU, Y., AND VASSILIADIS, P. *Fundamentals of Data Warehouses*, 2nd ed. Springer Publishing Company, 2010.

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

[32] KIFER, M., AND BOLEY, H. Rif overview. Tech. rep., W3C, June 2010. http://www.w3.org/TR/2012/NOTE-rif-overview-20121211/.

[33] KITCHENHAM, B. Procedures for performing systematic reviews.

[34] KNIGHT, S., AND BURN, J. Developing a framework for assessing information quality on the world wide web. *Information Science 8* (2005), 159 – 172.

[35] LEE, Y. W., STRONG, D. M., KAHN, B. K., AND WANG, R. Y. Aimq: a methodology for information quality assessment. *Information Management 40*, 2 (2002), 133 – 146.

[36] LEHMANN, J., GERBER, D., MORSEY, M., AND NGONGA NGOMO, A.-C. DeFacto - Deep Fact Validation. In *ISWC* (2012), Springer Berlin / Heidelberg.

[37] LEI, Y., NIKOLOV, A., UREN, V., AND MOTTA, E. 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* (2007), pp. 51–60.

[38] LEI, Y., UREN, V., AND MOTTA, E. A framework for evaluating semantic metadata. In *4th International Conference on Knowledge Capture* (2007), no. 8 in K-CAP '07, ACM, pp. 135 – 142.

[39] LEO PIPINO, RICAHRD WANG, D. K., AND RYBOLD, W. *Developing Measurement Scales for Data-Quality Dimensions*, vol. 1. M.E. Sharpe, New York, 2005.

[40] MAYNARD, D., PETERS, W., AND LI, Y. Metrics for evalua-

tion of ontology-based information extraction. In *Workshop on "Evaluation of Ontologies for the Web" (EON) at WWW* (May 2006).

[41] MEILICKE, C., AND STUCKENSCHMIDT, H. Incoherence as a basis for measuring the quality of ontology mappings. In *3rd International Workshop on Ontology Matching (OM) at the ISWC* (2008).

[42] MENDES, P., MÜHLEISEN, H., AND BIZER, C. Sieve: Linked data quality assessment and fusion. In *LWDM* (March 2012).

[43] MILLER, P., STYLES, R., AND HEATH, T. Open data commons, a license for open data. In *LDOW* (2008).

[44] MOHER, D., LIBERATI, A., TETZLAFF, J., ALTMAN, D. G., AND PRISMA GROUP. Preferred reporting items for systematic reviews and meta-analyses: the prisma statement. *PLoS medicine 6*, 7 (2009).

[45] MOSTAFAVI, M., G., E., AND JEANSOULIN, R. Ontology-based method for quality assessment of spatial data bases. In *International Symposium on Spatial Data Quality* (2004), vol. 4, pp. 49–66.

[46] NAUMANN, F. *Quality-Driven Query Answering for Integrated Information Systems*, vol. 2261 of *Lecture Notes in Computer Science*. Springer-Verlag, 2002.

[47] PIPINO, L., LEE, Y. W., AND WANG, R. Y. Data quality assessment. *Communications of the ACM 45*, 4 (2002).

[48] REDMAN, T. C. *Data Quality for the Information Age*, 1st ed. Artech House, 1997.

[49] RULA, A., PALMONARI, M., HARTH, A., STADTMÜLLER, S., AND MAURINO, A. On the diversity and availability of temporal information in linked open data. In *ISWC* (2012).

[50] RULA, A., PALMONARI, M., AND MAURINO, A. Capturing the age of linked open data: Towards a dataset-independent framework. In *IEEE International Conference on Semantic Computing* (2012).

[51] SCANNAPIECO, M., AND CATARCI, T. Data quality under a computer science perspective. *Archivi & Computer 2* (2002), 1–15.

[52] SCHOBER, D., BARRY, S., LEWIS, E. S., KUSNIERCZYK, W., LOMAX, J., MUNGALL, C., TAYLOR, F. C., ROCCA-SERRA, P., AND SANSONE, S.-A. Survey-based naming conventions for use in obo foundry ontology development. *BMC Bioinformatics 10*, 125 (2009).

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

[54] SUCHANEK, F. M., GROSS-AMBLARD, D., AND ABITEBOUL, S. Watermarking for ontologies. In *ISWC* (2011).

[55] WAND, Y., AND WANG, R. Y. Anchoring data quality dimensions in ontological foundations. *Communications of the ACM 39*, 11 (1996), 86–95.

[56] WANG, R. Y. A product perspective on total data quality management. *Communications of the ACM 41*, 2 (Feb 1998), 58 – 65.

[57] WANG, R. Y., AND STRONG, D. M. Beyond accuracy: what data quality means to data consumers. *Journal of Management Information Systems 12*, 4 (1996), 5–33.