Undefined 1 (2009) 1–5
IOS Press

Systematic Review on Data Quality Assessment Methodologies in Linked Open Data

Amrapali Zaveri [a], Anisa Rula [b], Andrea Maurino [b], Ricardo Pietrobon [c] and 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.

Keywords: data quality, assessment, survey, Linked Data

1. Introduction

The advent of semantic web technologies, as a enabler of Linked Open Data (LOD), has swept the world with an unprecedented data volume with close to 50 billion facts represented as triples. Although accumulating massive amounts of data is certainly a step in the right direction, data is only as good as it's quality. On the Data Web we have very varying quality of information and from various domains. Biological and health care data are no exception, with widespread availability of data in a wealth of areas including drugs, clinical trials, medicine, proteins, diseases all published as Linked Data amounting to 2.3 billion triples.

Data quality is commonly defined as the *fitness of use* 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. In the case of DBpedia, for example, the data quality is perfectly sufficient for enriching Web search with facts or suggestions about common sense information, such as entertainment topics (similarly to Google's recently launched knowledge graph). In such a scenario, where the DBpedia background knowledge can be, for example, used to show the movies Franka Potente was starring in and actors she played with it is rather neglectable if, in relatively few cases, a movie or an actor is missing. For developing a medical application, on the other hand, the quality of DBpedia is probably completely insufficient. It is to be noted that inspite of varying quality on the traditional document-oriented Web, it is still perceived to be extremely useful by most people. Consequently, the key challenge is to determine the quality of datasets published on the Web and make this quality information explicit. Other than on the document Web, where information quality can be only indirectly (e.g. page rank) or vaguely defined, we can have much more concrete and measurable data quality indicators for structured information (i.e. data), such as correctness of facts, adequacy of semantic representation or degree of coverage.

There are already many methodologies and frameworks available for assessing data quality which address different aspects of this task by proposing appro-

priate tools. Despite quality in LOD being an essential concept, few efforts are currently in place to standardize how quality tracking and assurance should be implemented. Therefore, in this article we present a survey of existing approaches, that have been published for assessing the data quality of Linked Data datasets. We attempt to gather the most notable approaches proposed so far in the literature, present them concisely in a tabular format and group them under a classification scheme. In particular, we formalise the commonly used terminologies across papers related to data quality. Additionally, a generalisation of the dimensions, criteria and indicators is presented along with an overview of the steps involved in the quality assessment. This is done to help researchers and implementors have a clearer view of existing work, thereby encouraging further experimentation.

2. Survey Methodology

This systematic review was conducted by two reviewers from different institutions following the procedures as described in [27,32]. 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 comprises of all the above mentioned reasons, in that, it summarises and compares various data quality assessment methodologies as well as identifies open problems focused on Linked Open Data. Moreover, it contributes a conceptualization of the data quality assessment field and thereafter proposes a new method for data quality assessment for LOD.

Related surveys. In order to justify the need of conducting the systematic review, we first conducted a search for related surveys and literature reviews. We did not come across any study focused on data quality assessment methodologies and tools for Linked Data. However, there is a comprehensive review [1], which surveys 13 methodologies for assessing the data quality of datasets available on the web in structured or semi-structured formats.

Research question. The goal of this review is to analyse 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:

What are the existing approaches for assessing the quality of Linked Data?

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

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

Define eligibility criteria. The eligibility criteria is an important element of any systematic review. First, each member created a set of inclusion and exclusion criteria on their own. Second, as a result of a discussion between both members a list of eligible criteria was obtained as follows:

- Inclusion criteria:
  - Studies published in English between 2002 and 2012.
  - Studies focused on data quality assessment in 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
  - Methodologies that were published as a poster
  - Studies that were focused on data quality management
  - Studies that did not focus neither on Linked Data nor on structured data
  - Studies that did non propose any methodology or framework about the assessment of quality in Linked Data

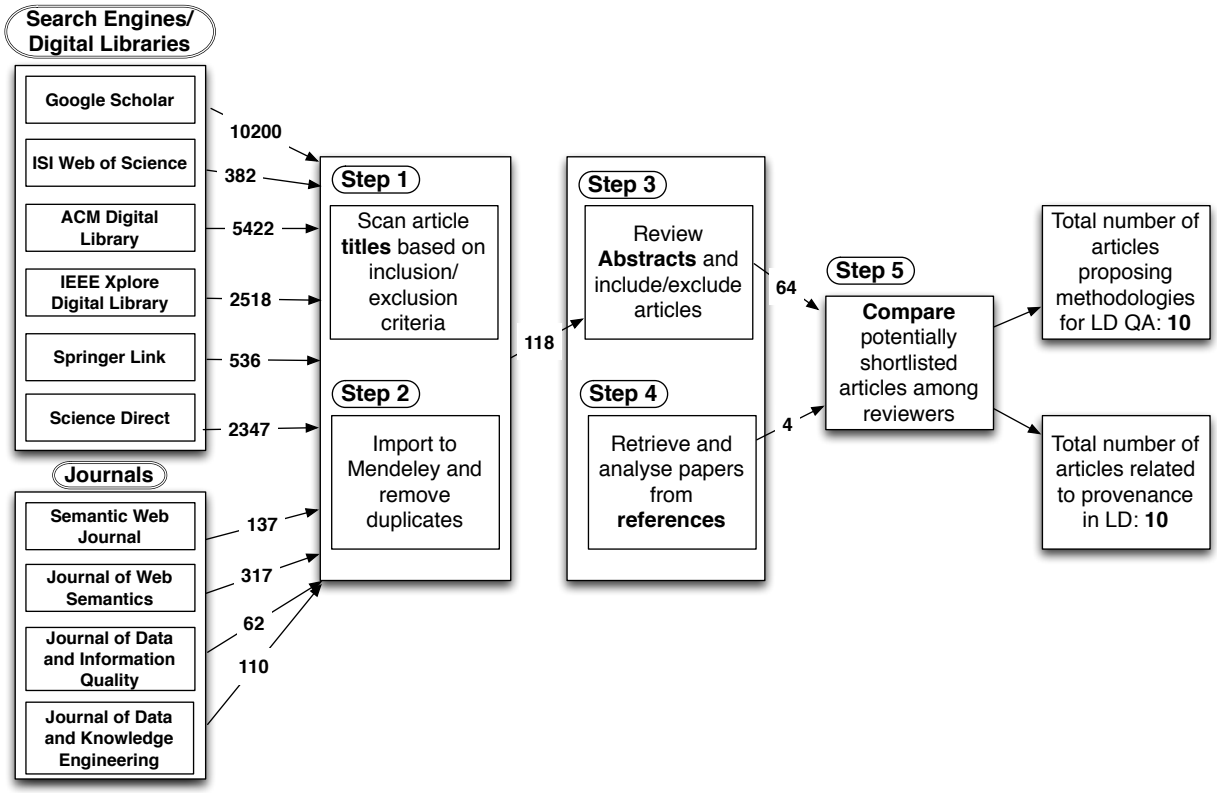

Fig. 1. Number of articles retrieved during literature search.

Search strategy. Search strategies in a systematic review are usually iterative and are ran 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*

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

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

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

Journals:

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

Conferences and their Respective Workshops:

- International Semantic Web Conference (ISWC)

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

Thereafter the bibliographic metadata about the 118 potentially relevant primary study 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 reference in the selected articles
- Looking up the article title in Google Scholar and retrieving the "Cited By" papers to check against the eligibility criteria
- Taking each data quality dimension individually and perform a related article search

After performing these search strategies, we further retrieved 4 additional articles.

Extracting data for quantitative and qualitative analysis. An overview of the search methodology and the number of retrieved articles at each step is shown in Figure 1. The result of the above described methodology is 21 papers from 2002 to 2012 that are reported in Table 1 which are the core of our survey. The next step was then to extract data from each of the articles to perform quantitative and qualitative analysis.

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 automatization of supported tools
- the phases/steps that compose the assessment methods

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

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

3. Conceptualization

There exist a number of discrepancies in the definition of many concepts in data quality due to the contextual nature of quality [1] . 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) in more 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 it may support multiple methods of access. The RDF triples, RDF graph and the RDF datasets have been adopted by the W3C Data Access Working Group [2,18,8].

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 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. The concept of data quality is a domain-specific subconcept of the general concept of quality. A popular definition for quality is the "fitness for use" [25]. Data quality is commonly conceived as a multidimensional construct, as the "fitness for use" may depend on various factors such as accuracy, timeliness, completeness, relevancy, objectivity, believability, understandability, consistency, conciseness, availability, and verifiability [36].

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 [28]. On the other hand,

Table 1
List of the selected papers.

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

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 [16]. Also, it is to be noted that *data* and *information* are interchangeably used in the literature.

Data Quality Problems. The data quality problem refers to a set of issues that can affect the potentiality of the applications that use data. Bizer et al. [4] defines data quality problems as the choice of web-based information systems design which integrate information from different providers. In [35] 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 [10] the author does not provide a definition of it but implicitly explains the problems in terms of *data diversity*. In [20] the authors discuss about *errors* or *noise* or *difficulties* and in [21] the author discuss about *modelling issues* which are prone of the non exploitations of those data from the applications.

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 [4]. These metrics are heuristics that are designed to fit a specific assessment situation [30]. Since the dimensions are rather abstract concepts, the assessment metrics rely on quality *indicators* that allow for the assessment of the quality of a data source w.r.t the criteria [10]. An assessment score is computed from these indicators using a scoring function.

In [4], 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) Provenance (3) Intrinsic (4) Accessibility (5) Representation and (6) Dataset Dynamicity, as depicted in Figure 3.

Data Quality Assessment Method. A data quality assessment methodology is defined as the process of evaluation if a piece of data meets in the information consumers need in a specific use case [4]. The process involves measuring the quality dimensions that are rel-

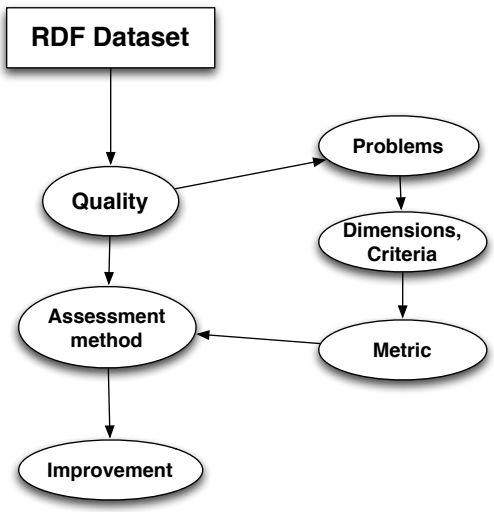

Fig. 2. Conceptualization of the data quality domain.

evant to the user and comparing the assessment results with the users quality requirements.

3.2. Linked Data quality dimensions

After analyzing the 20 selected approaches, we identified 25 different data quality dimensions that can be applied to assess the quality of Linked Data. In this section, we thus attempt to formalize and adapt the definitions for each of the dimensions to be applicable in the Linked Data context. Additionally, we also classify the dimensions into six different groups, such as

- *Contextual dimensions* are those that highly depend on the context of the task at hand as well as the subjective preferences of the data consumer.
- *Provenance* related dimensions are those that focus on the provenance and trustworthiness of the data.
- *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.
- *Accessibility dimensions* involves aspects related to the accessing of information.
- *Representational* dimensions capture aspects related to the design of the data like the conciseness, consistency as well as the interpretability of information.
- *Dataset dynamicity* contains dimensions related to the timeliness of the data. In particular, it focuses on two aspects: the currency (if the data is up-to-date) and volatility (time period of validity) of the data.

Figure 3 shows the classification of the dimensions in the above mentioned groups as well as the inter and intra relations between them.

3.2.1. Provenance

The ability to track the origin of data is a key component in building trustworthy, reliable applications.

3.2.2. Consistency

Consistency implies that *two or more values do not conflict with each other* [3]. Similarly Flemming et al. [10] and Hogan et al. [20] define consistency as *when there is no contradictions in the data*. Mendes [35] gives a more generic definition by determining a consistent dataset if it is free of conflicting information. Consistency is considered a sub-dimension of representational consistency and is marked as an intrinsic dimension.

An example of inconsistency can be the usage of disjoint classes [20]: the classes `foaf:Person` and `foaf:Document` are defined as being disjoint therefore an instance cannot belong to both classes. Lei et al. [29] also determine the issue of inconsistent use of terms in a vocabulary as a problem of deficiency, which denotes the situation of an inconsistent instance defined in an ontology. For example, an organization ontology may define that there should be only one director for an organization. The inconsistency problem occurs when there are two directors in the dataset. In [35] an example of inconsistency is given with the value of the total population of a city in two resources of DBpedia from two languages, namely the English and Portuguese language editions. If both resources contain different values for the population, in the process of integration of the dataset, this will lead to the problem of inconsistency of the value.

Definition 1 (Consistency)**.** *Consistency is always related to techniques for identifying contradictions. Depending on the available techniques, the assessment of consistency can vary. 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. Hence, our definition of 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 reasoner, which supports the respective expressivity of the underlying knowledge representation formalism. In practice, RDF-Schema inference and reasoning with regard to the different OWL profiles can be used. For domain specific applications, consistency rules can be defined, for example, according to the SWRL [22] or RIF standards [26] and processed using a rule engine.

3.2.3. Timeliness

An important aspect of data is their update over time. The main time-related dimension proposed in the literature is timeliness. However, there are two more dimensions related to time aspect such as currency and volatility. Timeliness is defined as the degree to which information is up-to-date [3] while in [10] the authors defines timeliness criterion as the currentness of the data provided by a source. The definition given in Mendes et al. corresponds to the distance between the input date from the provenance graph to the current date. The recent data receives scores closer to 1 [35]. Similarly Rula et al. describes currency defined as the age of data given as a difference between the current date and the date when the information is last modified in the system and provides two types of currency based on document and graph level [38].

Definition 2 (Timeliness)**.** *Timeliness is related to measurements of data freshness. Measuring timeliness, requires available temporal meta-information attached with data which can employ different representation approach. On the Linked Data Web, the approach employed can be metadata-based representation such as Dublin Core, reification-based representation or n-ary based representation. Hence, our definition of timeliness is relative to a particular set of temporal information, which are represented based on one of the above approaches, for measuring the distance between the current value and the last modified value of the data.*

3.2.4. Accuracy

Bizer et al. defined accuracy as the *degree of correctness and precision with which information in an information system represents states of the real world* [3]. Accuracy is an intrinsic dimension i.e. it is independent of the user's context. We found no formal definition of accuracy in any of the other articles.

However, Lei et al. describe inaccurate annotation such as *inaccurate labeling* and *inaccurate classification* [29] as an example of inaccuracy. Inaccurate labeling is when the mapping from the instance to the object is correct but is not properly labeled. For example, the person instance Trevor Collins might be correctly identified but wrongly marked as Both Trevor Collins. Inaccurate classification is when the knowledge of the source object has been correctly identified but not accurately classified. For example, the person Enrico Motta is classified as an instance of the high level class Person rather than the more precise class Professor. An additional problem identified is spurious annotation where there is no object to be mapped back to for an instance.

Based on the definition by [3], we identified the problem of detecting poor attributes described in Böhm as an example of inaccuracy. Poor attributed are those that do not contain useful values for data entries [5].

3.2.5. Completeness

Bizer defines completeness as the degree to which information is not missing. Schema completeness which is the degree to which entities and attributes are not missing in a schema; column completeness which is a function of the missing values in a column; and population completeness which refers to the ratio of entities represented in an information system to the complete population [3]. In other works completeness is not properly defined but introduce problems related to it and the approach of how to identify if there is a problem. Therefore, in [20] completeness is defined as a problem of incompleteness which includes equatable to a dead-link in the current HTML web, a software agent will not be able to retrieve data relevant to a particular task [20]. In [35] define completeness 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(instance) level, a dataset is complete if it contains all of the necessary objects for a given task.

3.2.6. Amount of data

Amount of Data is defined as the extent to which the volume of data is appropriate for the task at hand [3]. An alternative definition says that the amount of data provided by a data source influences its usability [10] and is appropriate to approximate a true scenario precisely without getting false negative [9]. It can be observed that there is a substantial agreement on the abstract definition of the amount of data. Although a

main advantage of the Web of Data compared to the traditional web is the possibility to aggregate data from several sources, the necessity to match the underlying vocabularies puts that advantage into perspective.

3.2.7. Availability

Availability is defined as the extent to which information is available, or easily and quickly retrievable [3] and refers to the proper functioning of all access methods [10].

3.2.8. Understandibility

Understandability is defined as the extent to which data is easily comprehended by the information consumer [3]. Similarly Flemming relates understandibility to the comprehensibility of data i.e. the ease with which human consumers can understand and utilize the data [10].

3.2.9. Relevancy

Relevancy is defined as the extent to which information is applicable and helpful for the task at hand [3].

3.2.10. Reputation

Since there is no definition for reputation we provide one which applies in the Web of Data. Reputation can be associated with a data publisher, a person, organisation, group of people, community of practice. The data publisher should be identifiable for a certain (part of a) dataset. Reputation is usually a score, for example, a real value between 0 and 1. There are different possibilities to determine reputation and can be classified into direct/indirect approaches. Direct - survey in a community questioned about other members. Indirect - use links/references/page rank.

3.2.11. Verifiability

Verifiability is defined as the degree and ease with which the information can be checked for correctness [3]. Similarly Flemming refers to verifiability criterion as to the means a consumer is provided with, which can be used to examine the data for correctness. Without such means, the only way of having a certain assurance of the correctness of the data is the consumer's trust in the source [10].

3.2.12. Interpretability

Interpretability is defined as the extent to which information is in appropriate languages, symbols, and units, and the definitions are clear [3]. Issues related to meaning (part of semantic), e.g. currency in dollars.

3.2.13. Representational Conciseness

In [3], representational conciseness is defined as *the extent to which information is compactly represented.* This dimension is related to the format or Representation of the data.

For example, in [21], 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.

Definition 3 (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. It can be also seen as the degree to which the structure of the available data corresponds to the data itself, rather than being too verbose.*

Representation of RDF data in N3 format is considered to be more compact than RDF/XML [10]. The concise representation not only contributes to the human readability of the data but also influences the performance of data when queried.

3.2.14. Representational Consistency

In [3], representational consistency is defined as *the extent to which information is represented in the same format.* This dimension is related to the format or Representation of the data.

As stated in [21], 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. In practice, for instance, when a data provider needs to describe information about people, FOAF [2] should be the vocabulary of choice.

Definition 4 (Representational Consistency)**.** *Representational consistency is the degree to which the format and structure of the information conforms 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.*

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 sup-

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

ports data integration and management tasks. Moreover, it maximises the probability that data can be consumed by applications that may be tuned to well-known vocabularies, without requiring further preprocessing 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 [3], SchemaCache [4] and Swoogle [5].

3.2.15. 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 [21,10]. Additionally, the existence of a machine-readable indication (by including it in a voiD description) as well as a human-readable indication of a licence is also important. Licensing is part of the meta-information and Provenance of a dataset.

An empirical analysis done in [21] showed that 27 PLDs (Pay Level Domains) i.e. 14.4% returned some licensing information for some local document. They found that, on average, providers gave licensing information for 3.4% of local documents. They also identified the need for an agreed-upon licensing property and an agreed set of common license URIs.

Definition 5 (Licensing)**.** *One of the main aims of Linked Data is 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. Indication of a license grants a consumer the right to (re-)use that data along with a condition to this (re-)use. A source can choose a licence 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 [31].*

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.

[3] http://www.schemaweb.info/
[4] http://schemacache.com/
[5] http://swoogle.umbc.edu/

3.2.16. Performance

In [10], performance is denoted as that which *comprises aspects of enhancing the performance of a source as well as measuring of the actual values.* However, in [21], performance is associated with issues such as avoiding prolix RDF features such as (i) RDF reification, (ii) RDF containers and (iii) RDF 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. Performance is related to the Accessibility of a dataset.

For example, [21] identifies the hosting of unstable URIs as a problem associated with the performance of agents and applications. The problem could occur when remote resources are removed, changed, updated or moved and the links are not updated and thus affects the performance of a dataset.

Definition 6 (Performance)**.** *Performance refers to the accessibility of a dataset, that is, the more performant a data source the more efficiently a consumer can access or process the data. Provision of Linked Data as an RDF dump, use of hash-URIs, are means of improving the performance of a dataset. Additionally, high performance can be achieved through low latency and high throughput. Latency is the amount of time in seconds from issuing the query until the first information reaches the user. Achieving high performance should be the aim of a dataset, however large it may be.*

Since Linked Data involves the aggregation of several large datasets, they should be easily and quickly retrievable. Also, the performance should be maintained even while executing complex queries over large amounts of data.

3.2.17. Objectivity

In [3], objectivity is expressed as *the extent to which information is unbiased, unprejudiced and impartial.* Objectivity is an Intrinsic dimension and overlaps with the concept of Accuracy. It is also strong related to the Verifiability dimension, that is, the more verifiable a source is, the more objective it will be.

Objectivity highly depends on the type of information. For example, the height of a building can be measured objectively whereas the objectivity of product descriptions require the preference of the information provider [3].

Definition 7 (Objectivity)**.** *Objectivity refers to the bias or opinion expressed when a data publisher interprets or analyze facts. It is therefore defined as the de-*

gree to which data is unbiased, unprejudiced and impartial. Objectivity, in terms of Linked Data, is a subjective dimension that focuses on how well the data publisher can express or represent the real world data as RDF such that it is unbiased, unprejudiced and impartial.

Search engines may display biased information due to either (i) occurrence of a certain keyword on a webpage indexed by a search engine to be ranked higher for searchers or (ii) paid web sites to purposefully rank their pages higher than others. These kinds of bias lead to error in judgement and decision making and should be avoided.

3.2.18. Believability

In [3], believability is explained as *the extent to which information is regarded as true and credible*. Believability is a subjective dimension, is part of the Provenance of a dataset and can be seen as expected Accuracy. It is strongly related to the Verifiability and Reputation dimensions.

Definition 8 (Believability)**.** *Believability is the degree to which the information is accepted to be correct by a user. Therefore, it can be defined as the extent to which information is regarded or accepted as true, real and credible. It involves the trusting of information that is provided, that is, in other words deciding which information to believe.*

In Linked Data, believability can only be subjectively measured and highly depends on the provenance information of the dataset. Tim Berners-Lee proposed [6] 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. Pressing of such a button for any piece of data or an entire dataset would contribute towards the trustworthiness, that is the believability, of the dataset.

3.2.19. Response Time

In [3], response time is defined that which *measures the delay between submission of a request by the user and reception of the response from the system.* This dimension depends on the type and complexity of the request and is related to the Accessibility of a dataset.

Definition 9 (Response Time)**.** *Response Time measures the delay in seconds between submission of a query by the user and reception of the complete response from the information system. It depends on several factors such as network traffic, sever workload, server capabilities and/or complexity of the user query, which affect the quality of query processing.*

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.

3.2.20. Security

Security is a dimension which points to *the possibility to restrict access to the data and to guarantee the confidentiality of the communication between a source and its consumers.* This dimension is related to the Accessibility of a dataset.

Definition 10 (Security)**.** *Security can be defined as the extent to which access to data can be restricted and hence kept secure. It refers to the degree with which information is passed securely from users to the information source and back. Security covers technical aspects of the accessibility of a dataset, such as secure login and the authentication of a information source by a trusted organization.*

However, this dimension does not have widespread applicability in Linked Data since it mandates that the data should be open and not restricted for use, albeit with an indication of a license. On the other hand, adequate protection of a dataset is an important aspect to be considered against its alteration or misuse and therefore a reliable and secure infrastructure of storing a dataset is important.

3.2.21. Uniformity

In [10], uniformity is defined as *the usage of established techniques in order to increase the usability of the data.*, similar to the Representational Consistency dimension. Uniformity is part of the Representation of a dataset as it recommends the usage of established formats and vocabularies, referencing URIs of established datasets and stating the content-types as specifically as possible.

For example, in [20], it was observed that RDF/XML content content was commonly returned with a content-type other than `application/rdf+xml`, such as `text/xml` (9.5%) or `application/xml` (5.9%) or `text/plain` (1%) or `text/html` (0.4%).

Definition 11 (Uniformity)**.** *The uniformity dimension refers to the usage of established techniques in order to represent data in a consistent format . Standardized*

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

formats such as RDF/XML, N3 or RDFa for data representation should be used to represent the data. Additionally, specification of a correct content-type, usage of established vocabularies and adding references of established URIs contribute towards the uniformity of the data.

The presence of uniformity in a dataset increases the usability of the data as it enables the consumer to process the data easily and also increases the chances of connecting the dataset with others. Moreover, it helps to better understand the data as the semantics of a particular entity may be better represented using an established URI.

3.2.22. Versatility

In [10], versatility is referred to as the *alternative representations of the data and its handling*. Versatility is part of the Representation of a dataset.

Definition 12 (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. Furthermore, handling of these versatile representations and access methods is an important aspect to be considered. Thus, versatility is defined as the alternative representations and also alternative access methods of a dataset as well as its subsequent handling.*

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 is an indication of the versatility of the dataset. Provision of Linked Data in the HTML format 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 a HTTP request.

3.2.23. 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 [10]. This is an Intrinsic dimension and related to the Accuracy, Consistency and Objectivity dimensions.

An example of the invalid usage of the underlying vocabulary is the usage of `foaf:image` instead of the defined `foaf:img`. Also, the use of deprecated classes and properties results in the usage of undefined terms in the future versions of these vocabularies [20]. Another example of the invalid syntax is the assignment of improper datatypes to literals [20].

Definition 13 (Validity of documents)**.** *Validity of documents, in Linked Data, can be defined as the syntactic accuracy of documents, in that, it refers to not only the valid usage of the underlying vocabularies but also the valid syntax of the documents. Valid usage of the underlying vocabularies refers to the usage of the existing vocabularies as intended, that is, using the right semantics and in a valid context. A syntax validator should be used to ensure the validity of RDF data, that is, the correct parsing of the triples.*

RDF datasets with mistyped descriptions can result in an incorrect interpretation of the data. Moreover, invalid usage of certain vocabularies can result in consumers not able to process data as intended. Syntax errors, typos, not defining any additional properties that are added, use of deprecated classes and properties all add to the problem of the invalidity of the document as the data cannot neither be processedn or a consumer cannot perform reasoning on such data.

3.2.24. Conciseness

The authors in [35] characterizes conciseness as follows: *On the schema level, 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, a dataset is concise if it does not contain redundant objects (two equivalent objects with different identifiers). Similarly, extensional conciseness measures the number of unique objects in relation to the overall number of object representations in the dataset.*

For example, as shown in [35], in the data integration task of the two editions of DBpedia (from two different languages), when the two editions contain identical URIs per object, it is an example of extensional conciseness or redundancy in the dataset.

3.2.25. Coherence

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.

Definition 14 (Coherence)**.** *Coherence refers to the creation and maintenance of links in a (semi-) automated way to facilitate data integration. 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.*

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. For example, the disease named *Tuberculosis* in one dataset is the same as the disease named *TB* in another dataset. Therefore, it is necessary to link both the URIs to enable queries over both the datasets. 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. For example, the creation data of a person may be different than the creation data of a document that describes this person. The ability to distinguish the two through the use of different URIs is critical to the coherence of the Web of Data [19].

4. Comparison of selected approaches

In this section, we compare the selected approaches based on the different perspectives discussed in Section 2. In particular, we analyze and compare each approach based on the dimensions (Section 4.1), their respective metrics (Section 4.2), types of data (Section 4.3), level of automatization (Section 4.4) and the usability of tools (Section 4.5).

4.1. Dimensions

The linked open data paradigm is the fusion of three different research areas namely *semantic web* (for the capability to generate semantic connections among data), *world wide web* (related to the availability and open access to huge amounts of data) and *data management* (owing to the fact that there is the need to mange large set of heterogeneous and distributed data). The selected approaches use quality dimensions taken from any one of these specific areas. In the data management area the literature provides a thorough classification of data quality dimensions. By analyzing the classifications of quality dimensions provided by [41,42,37,24,7,34], it is possible to define a basic set of data quality dimensions as: accuracy, completeness, consistency and timeliness, which constitute the focus of the majority of authors Catarci et al. [39]. However, no general agreement exists either on which set of dimensions defines the quality of data or on the exact meaning of each dimension and the same problem also occurs in LOD.

As mentioned in Section 3, data quality assessment involves the measurement of quality dimensions that are relevant to the consumer. An initial list of data quality dimensions was first obtained from [3]. Thereafter, each approach was analyzed to extract the problem it was tackling and thus mapped to one or more of the quality dimensions. For example, the dereferencability issues, problem of no structured data available and misreported content mentioned in [21] were mapped to the dimensions of Completeness as well as Availability. However, not all the problems could be mapped to the initial set of dimensions such as the problem of incoherency or interlinking between datasets or the problem of the alternative representations of the data and its handling i.e. the versatility of the dataset. As a result , we obtained a further set of quality dimensions, which were particularly relevant for Linked Data. Table 2 shows the complete list of the 25 identified Linked Data quality dimensions along with frequency and occurrence of each dimension in the included approaches.

As can be seen in the table, there are three visible groups: (a) a set of approaches that focus only on the provenance of the datasets [17,11,40,14,13,15,12,23, 6,38]; (b) a set of approaches which use majority of the dimensions (more than 5) [4,10,21,35] and (c) a set of approaches which focus on very few and specific dimensions [5,9,16,20,29,33]. Additionally it can be observed that Provenance, Consistency, Timeliness, Accuracy and Completeness are most frequently used dimensions for the majority of the approaches.

4.2. Metrics

As defined in Section 3, data quality metric is a procedure for measuring an information quality dimension. In general, multiple metrics can be associated with each quality dimension. *In some cases, the metric is unique and the theoretical definition of a dimension coincides with the operational definition of the corresponding metric. For this reason, in the following*

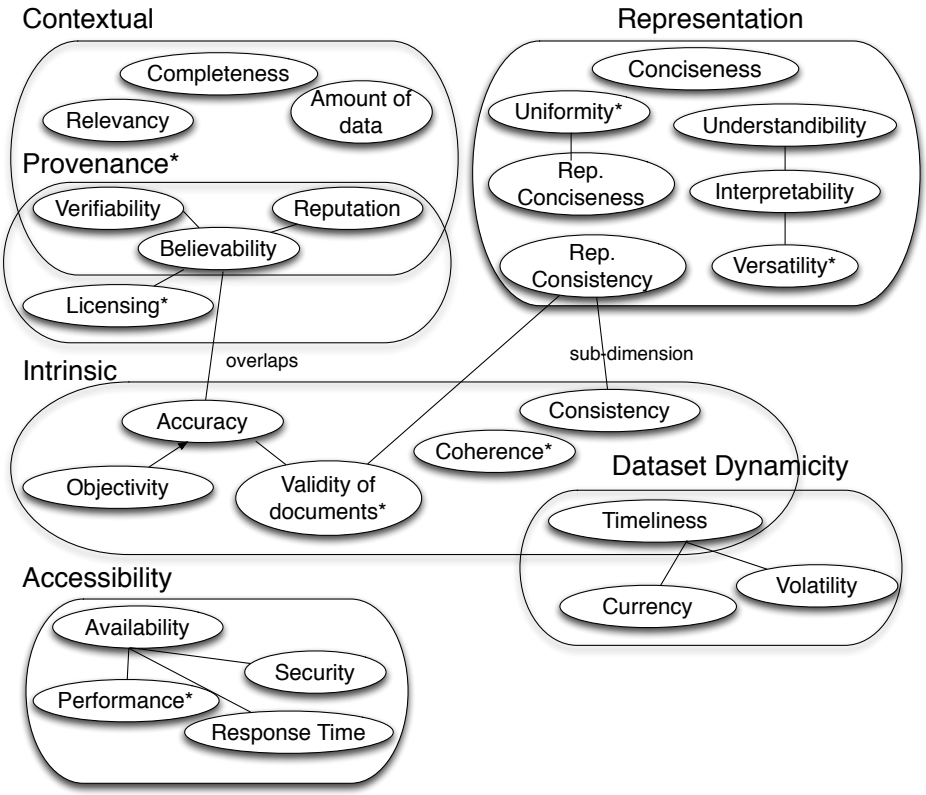

Fig. 3. Linked data quality dimensions and the relations between them.

we make a distinction between theoretical and operational definitions of dimensions only when the literature provides multiple metrics. Table 3 provides a list of the data quality metrics for each of the dimension and classifies it as being either subjective or objective.

In [17] the authors determine the measurement of trustworthiness of RDF statements as a value of trust which is either unknown or a value in the interval [-1,1]. Actually, they do not prescribe and implementation of how to assign a trust value to the statements but instead they provide a data provenance model which includes information about the publisher of the dataset, creation method and creation time of the dataset and also the publisher and publication time of possible original sources. Therefore, these provenance information can be exploited to assess other dimensions such as timeliness or accuracy which in turn can be merged to provide a unified value of provenance. In [11] the authors suggest to have a trusted third party to provide information such as citation count of a publication or global reputation. Even in this case, the meta information should be provided in advance and the trustworthiness is measured by the reputation dimensions. In [40] the author propose a statistical approach to measure a trust propagation rating between two nodes. In [14] the authors infer the trust path from a node A to a node C based on previous trust values given from the path AB and BC. However, a pre-established value of trust is needed. In [13] the authors do not address how the trust value is derived. Even in this case trustworthiness is calculated possibly as a combination of its popularity, reputation, and authority.

4.3. Type of data

The ultimate goal of an assessment activity is the analysis of data that, in general, describes real world objects in a format that can be stored, retrieved, and processed by a software procedure, and communicated through a network. In LOD, most authors either implicitly or explicitly distinguish three types of data:

– RDF triple. 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; s, p, o are called, respectively, the subject, the predicate and the object of the triple.

- Graph. An RDF graph G is a set of RDF triples. A named graph is a pair $\langle G, u\rangle$, where G is called the default graph and $u \in \mathcal{U}$. [ANDREA we need to underline that a graph is a set of datasource provided by different providers]
- Dataset. An RDF dataset is a set of default and named graphs = $\{G, (u_1, G_1), (u_2, G_2), ...(u_n, G_n)\}$.

4.4. Level of automatization

A set of software tools are needed to support the assessment phase. Such tools implement the methodologies and metrics defined in the above described steps. Due to the nature of the quality dimensions and related metrics it is possible that some activities are fully or semi automatic or manually realized. Table 4 shows the level of automatisation for each of the identified tools.

Table 2: Consideration of data quality dimensions in each of the included approaches.

Approaches / Dimensions	Provenance	Consistency	Timeliness	Accuracy	Completeness	Amount of Data	Availability	Understandability	Relevancy	Reputation	Verifiability	Interpretability	Rep. Conciseness	Rep. Consistency	Licencing	Performance	Objectivity	Believability	Response Time	Security	Uniformity	Versatility	Validity of documents	Conciseness	Coherence
Bizer et.al.,2009		✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔			✔	✔	✔	✔					
Böhm et.al.,2010		✔		✔																					
Chen et.al.,2010		✔				✔																			
Flemming 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			✔	✔	✔									✔											
Hartig,2008	✔																								
Gamble et.al., 2011	✔																								
Shekarpour et.al., 2008	✔																								
Golbeck et.al., 2006	✔																								
Gil et.al., 2002	✔																								
Golbeck et. al., 2003	✔																								
Gil et.al., 2007	✔																								
Jacobi et.al., 2011	✔																								
Bonatti et. al., 2012	✔																								
Rula et.al., 2012	✔																								

Table 3: Comprehensive list of data quality metrics and the type - subjective or objective

Dimension	**Metric**	**Type of metric ("S"ubjective/ "O"bjective)**
Accuracy	spurious annotation/representation	O
	inaccurate annotation	O
	inacurrate labeling and classification	O
	detecting poor attributes	O
Consistency	no definitions of entities as being members of disjoint classes	O
	valid usage of inverse-functional properties	S
	no redefinition of existing properties	S
	usage of homogeneous datatypes	O
	no stating of inconsistent values for properties	O
	duplicate annotation	O
	ambiguous annotation	S
	duplication representation	O
	atypical use of collections, containers and reification	O
	use of undefined classes and properties	O
	misplaced classes or properties	O
	misuse of owl:datatypeProperty or owl:objectProperty	O
	use of members of deprecated classes or properties	O
	bogus owl:InverseFunctionalProperty values	O
	malformed datatype literals	O
	literals incompatible with datatype range	O
	ontology hijacking	O
	misuse of predicates	O
	negative dependancies (correlation) among predicates (attribute), negative entity correlation	O
Objectivitiy	no bias or opinion expressed when a data provider interprets or analyses facts	S
Timeliness	stating the most recent and frequency of the validation of the data	S
	no inclusion of outdated data	O
	dereferencability of all internal and external URIs	O
	time inaccurate representation of data	O
Believability	meta-information about the identify of information provider	O
	checking source from which information is retrieved	S
Completeness	schema completeness	S
	column completeness	O
	population completeness	O
	data (instance) completeness	S
	URI/HTTP: accessibility and dereferencability	O
	no structured data available	O
	misreported content types	S
	number of interlinks	O
Understandibility	human-readable labelling of classes, properties and entities by providing rdfs:label	O

	human-readable description of classes, properties and entities by providing an rdfs:comment	O
	indication of metadata about a dataset	O
	indication of one or more exemplary URIs	O
	indication of a regular expression that matches the URIs of a dataset	O
	indication of an exemplary SPARQL query	O
	indication of some of the vocabularies used	O
	operability of HTML documents	S
	provision of message boards and mailing lists	O
	meta information about the language of web content	O
Relevancy	using meta-information attributes	S
	count occurrence of relevant terms within attributes	S
	sort documents according to their relevancy for a given query	S
	existence of links to external data providers: use of external URIs, provide owl:sameAs link	S
Verifiability	stating basic provenance information	S
	usage of a dedicated provenance vocabulary	O
	usage of digital signatures	O
Amount of data	no. of triples in a dataset	O
	no. of internal and external links	O
	scope and level of detail - coverage	S
	whether the amount of collected data is appropriate	S
Interpretability	use of appropriate language, symbols, units and clear definitions	S
	use of self-descriptive formats, identifying objects and terms used to define the objects with globally unique identifiers	O
	use of various schema languages to provide definitions for terms	S
	avoiding blank nodes	O
	dereferenced representations: giving human readable metadata	O
Representational Conciseness	keeping URIs short	O
Representational Consistency	re-use existing terms	O
	thorough picking vocabularies	S
	provision of data in different representational formats	O
Availability	accessibility of the server	O
	accessibility of the SPARQL endpoint	O
	accessibility of the RDF dumps	O
	usage of dereferencable URIs	O
	hosting of stable URIs	O
	redirection using the status code 303 See Other	O
	dereferencability issues	O
	no structured data available	O
	misreported content types	S
	dereferenced forward links	O
	dereference back-links	O
Response time	delay between submission of a request by the user and reception of the response from the system	O
Security		
Uniformity	usage of an established format	O
	stating the content-types as specifically as possible	O
	usage of established vocabularies	O

	referencing of established URIs	O
Versatility	provision of the data in various formats	O
	provision of the data in various languages	O
	application of content negotiation	O
	correct interpretation of the accept-headers sent	O
	human-readable indication of a SPARQL-endpoint	O
	machine-readable indication of a SPARQL-endpoint	O
Validity of documents	no syntax errors	
	exclusive usage of defined classes and properties	O
	no usage of deprecated classes and properties	O
	usage of proper datatypes	O
Licensing	machine-readable indication of a licence	O
	human-readable indication of a licence	O
	permitted reproduction of data	O
	permitted distribution of data	O
	permitted modification and redistribution of data	O
	no attribution needed	O
	no Copyleft / Share Alike needed	O
Performance	machine-readable indication of one or more RDF dumps	O
	human-readable indication of one or more RDF dumps	O
	usage of Slash-URIs when providing a large amount of data	O
	low latency	O
	high throughput	O
	maintenance of the performance no matter the load	O
	avoiding use of prolix RDF features	O
Conciseness	schema level : intensional conciseness - does not contain redundant attributes	O
	data level : extensional conciseness - does not contain redundant objects	O
Provenance	using provenance information and trust annotations in SW based social networks	S
	a trust value based on the meta-information about the provenance of the data which can be either unknown or a value in the interval [-1,1]	S
	construction of decision networks informed by provenance graphs	O
	statistical techniques and an aggregation algorithm based a weighting mechanism that utilizes fuzzy logic for modeling	O
	use of annotations by many individuals	O

Table 4

Qualitative evaluation of frameworks

Paper	Application	Goal	Type of data				Degree of automation			Tool support
	General/**S**pecific		**Triple**	**Resource**	**Several resources**	**Entire LOD Cloud**	**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	✔	✔	-	-	-	✔	-	✔ http://trellis.isi.edu/
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	✔	-	-	-	-	✔	-	✔ http://trdf.sourceforge.net/
Bizer et.al., 2009	G	Information filtering	✔	✔	✔	-	✔	-	-	✔ http://www4.wiwiss.fu-berlin.de/bizer/wiqa/
Böhm et.al., 2010	G	Data integration	✔	✔	-	-		✔	-	✔ https://www.hpi.uni-potsdam.de/naumann/sites/lodprof/ProLOD/ProLOD.html
Chen et.al., 2010	G	Generating semantically valid hypothesis	✔	✔	-	-	-	-	-	-
Flemming et.al., 2010	G	Assessment of published data	✔	✔	-	-	-	✔	-	✔ http://linkeddata.informatik.hu-berlin.de/LDSrcAss/
Hogan et.al., 2010	G	Assessment of published data by identifying RDF publishing errors and providing approaches for improvement	✔	✔	✔	✔	-	✔	-	✔ http://www.w3.org/RDF/Validator/
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	-	-	-	✔	-	-	✔	✔ http://qa.linkeddata.org/frontend/
Hogan et.al., 2012	G	Assessment of published data	✔	✔	✔	✔	-	-	-	-
Mendes et.al., 2012	S	Data integration	✔	-	-	-	✔	-	-	✔ http://www4.wiwiss.fu-berlin.de/bizer/sieve
Rula et.al., 2012	G	Assessment of time related quality dimensions	✔	✔	✔	-	-	-	-	-

4.5. Comparison of tools

In this section, we analyze three particular tools, namely, Flemmings Data Quality Assessment Tool, Sieve and LODGRefine to assess their usability for data quality assessment. In particular, we compare them in terms of their ease of use, level of user interaction, applicability in terms of data quality assessment and discuss their pros and cons.

Flemmings Data Quality Assessment Tool. The data quality assessment tool proposed in [10], is a simple user interface[7], where a user first needs to specify the name, URI and three entities of a particular data source. Then, via a series of steps, the user is ultimately provided with a score out of 100 indicating the quality of the dataset.

After specifying the dataset details, the user is given an option of assigning weights to each of the pre-defined data quality metrics. There are two choices for assigning the 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 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 of Linked Open Datasets and those which cannot be quantified. For example, questions such as the use of stable URIs, the number of obsolete classes and properties and whether the datasets provides a mailing list. Next, the user is presented with a list of dimensions and metrics and is allowed to specify yet again a set of weightings for each of them. This step is important especially for those indicators for which no formalization of the quantification exists (indicated against each of the metric). Also, those metrics which are assigned a weight of 0 are are not included in the final assessment. Each metric is provided with 2 input fields: first showing the assigned weights and second showing the calculated value.

At the end, the user is presented with a score, out of 100, based on the answers of all the questions, which is the data quality score. Additionally, the rating of each dimension and the total weight (out of 11 on account of 11 dimensions used in the assessment) is presented based on the user input from the previous step. Figure 4 shows an excerpt of the tool showing the result of assessing the quality of DBpedia with a score of 64 out of 100.

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

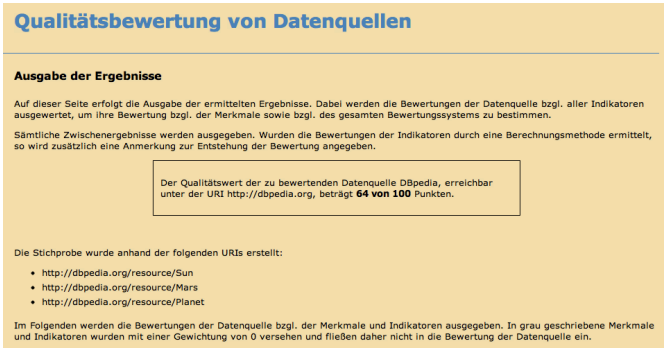

Fig. 4. Excerpt of the Flemmings Data Quality Assessment tool showing the result of assessing the quality of DBpedia with a score of 64 out of 100.

On the 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, on the other hand, 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 proposed 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.

Sieve.

LODGRefine. LODGRefine[8] is a LOD-enabled version of Google Refine, which is an open source tool for working with messy data. Although this tool is not focused towards data quality assessment per-say, it is very 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 over it. In particular, it helps in detecting duplicates, discovering patterns (e.g. alternative forms of abbreviations), spotting inconsistencies (e.g. trailing white spaces), finding blank cells and similar errors.

Additionally, this tool allows users to reconcile data, that is to connect a dataset to existing vocabularies such that it gives meaning to the field values. This

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

feature thus assists in assessing as well as improving the data quality, in particular the interpretability, of a dataset. Reconciliations to Freebase[9] 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[10]. Moreover, it is also possible to extend the reconciled data with DBpedia as well as exporting the data as RDF using this tool, which adds to the uniformity of the dataset.

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, exporting the data as RDF add to the usability, interpretability and uniformity of the dataset. 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.

4.6. Summary and comparison of selected approaches

5. Proposed Linked Data quality assessment steps

As defined in Section 3, a data quality assessment methodology is defined as the process of evaluation if a piece of data meets in the information consumers need in a specific use case [4]. In each of the identified approaches, we extracted the methodology that they followed to assess the quality of a dataset. Based on our analysis of the existing approaches as well as the evaluated tools, we identified a series of steps followed by majority of the approaches. We then adapted and revised the steps to align them with the data quality assessment process performed particularly for LOD as follows:

1. Requirements analysis (optional)
2. Data Quality Checklist
3. Statistics and low-level analysis
4. Aggregated and higher level metrics
5. Comparison (optional)
6. Interpretation

Table 5 comprises of the steps involved in the data quality assessment process. We also provide the input and output for each step along with a list of tools that would support at each step. Moreover, we identify the user involvement for each of the steps, that is, we identify whether the tool is automated, semi-automated or manual.

We now describe each of the steps in detail.

Requirements analysis (optional). The multidimensionality of the information quality makes it dependent on a number of factors that can be achieved by the analyses of the user requirements. Thus, the use case in question is highly important when assessing the quality of a dataset. This step is optional since it is not always provided in LOD related approaches and not all users necessarily have a use case in mind when assessing the quality of the dataset. In other words, a user might just want to find out the completeness of a dataset and will not have any particular use case for that particular dataset.

Input. As the input is the use case specified by the user with the aim of assessing the data quality of a particular dataset.

In this step, we identify two types of users: (a) those who know the problem with their dataset and (b) those who don't. Both users, however, are interested in finding the *fitness of use* for their dataset.

Data Quality Checklist. After specifying a use case or after deciding to assess the quality of a dataset, the user then is presented with a list of data quality dimensions but only those that do not have a specified statistical metric available. That is those dimensions which can perhaps be measured qualitatively.

Input. In this step, a checklist of data quality dimensions are presented to the user, but only those dimensions which call upon a boolean answer. That is, the user has to either tick which dimensions are present or provide a 0 (no) if it not present or 1(yes) if it is present.

Output. The output of this step is the result of the evaluation of these boolean dimensions, that is, a sum of 0's(no) or 1's(yes) which add to the final data quality assessment report.

Tool Support. Flemmings Data Quality Assessment tool[11] includes such questions in the process of data quality assessment. For example, questions such as whether the datasets provides a message board or a

[9] http://www.freebase.com/

[10] http://refine.deri.ie/reconciliationDocs

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

mailing list pointing to the Comprehensibility dimension are asked to the user. Thus, the user involvement is entirely manual where the user must have knowledge about the details of the dataset to answer these questions.

Statistics and low-level analysis. This step performs basic statistical and low-level analysis on the dataset. Generic statistics on the dataset are calculated.

Input. The input for this step is the dataset itself.

Output. The output is a statistical overview of the dataset. That is, it provides generic statistics on the dataset based on certain pre-defined heuristics.

Tool Support. LODStats[12] and LODGRefine[13] are two tools identified to provide such basic analysis on a dataset. LODStats gathers comprehensive statistics about a dataset available as RDF. It helps to know the structure, coverage and coherence of the data. It provides 32 different statistics such as property usage, vocabulary usage, datatypes used, average length of string literals, number of interlinks, to name a few. Thus, it helps to provide important insights with regard to the expected quality.

LODGRefine, on the other hand, helps users import their dataset and presents them an overview of the possible short-comings of the dataset. For example, characteristics such as Completeness, Interpretability of the column headers or Consistency is evident to the user. Moreover, it allows easy data cleansing mechanisms by performing transformations on a large amount of data at once. Additionally, LODGRefine also provides the functionality of reconciling data with Freebase[14] or DBpedia, thus increasing the Coherency of the dataset. Both these tools work automatically once a user uploads their dataset.

Aggregated and higher level metrics. In this step,

Input. This step includes all the dimensions that are not included in the Data Quality Checklist step, that is those dimensions which can be measured quantitatively.

Output. As an output, a single metric or a combination of them produces a value within the range [0;1].

Tool Support. There are a number of tools we identified that perform the tasks involved in this step. WIQA [4] is a information quality assessment framework which allows users to apply a wide range of policies to filter information. However, it requires a user to manually

Comparison (optional). Used when the resulted measurements provided in step "Aggregated and higher level metrics" are compared to reference values such as previous values from dataset in the same domain or gold standard values, in order to enable a diagnosis of quality.

Input. "Target/derived dataset (eg. Geonames) Assessment results from step 2 and 4 Original dataset (eg. OpenStreetMap)"

Output. evaluation of the representation evaluation between datasets in the same domain"

Tool Support.

Interpretation. Gives an interpretation to the results obtained from step Data Quality Checklist.

Input. Assessment results from Step 2 and 4

Output. Explanation of the results

Tool Support.

6. Conclusions and open issues

6.1. Open issues

References

[1] 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.

[2] BECKETT, D. RDF/XML Syntax Specification (Revised). W3c recommendation, World Wide Web Consortium, 2004. http://www.w3.org/TR/2004/REC-rdf-syntax-grammar-20040210/.

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

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

[5] 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.

[6] 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.

[7] 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.

[12] http://stats.lod2.eu/

[13] http://code.zemanta.com/sparkica/

[14] http://www.freebase.com/

Table 5
Data quality assessment steps

Steps	Input	Output	Tool support	User involvement		
				Automated	Semi-automated	Manual
Requirements analysis (optional)	Assessment of data quality, 2 types of users: – who know the problem with their dataset – who do not know the problem with their dataset	-	-	-	-	-
Data quality checklist	Checklist of dimensions which have a binary evaluation	Results of the evaluation - 0 (no) or 1 (yes)	Flemming et al., 2010	-	-	-
Statistics and low level analysis	Dataset	Overview statistics of the dataset	LODStats	✔	-	-
			LODGRefine	✔	-	-
Aggregated and higher level metrics	Dimensions not included in Step 2	Results of the evaluation of these dimensions in a range from 0 to 1	WIQA	-	✔	-
			ProLOD	-	✔	-
			Flemming et al., 2010	-	✔	-
			LinkQA	✔	-	-
			W3C's RDF Validator	-	✔	-
			Sieve	-	-	✔
			EvoPat	-	-	✔
			ORE	-	✔	-
Comparison (optional)	– Target/derived dataset – Assessment results from step 2 and 4 – Original dataset	– Evaluation of the representation – Evaluation between datasets in the same domain	-	-	-	-
Interpretation	Assessment results from Step 2 and 4	Explanation of the results	WIQA	-	-	✔

[8] 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/.

[9] CHEN, P., AND GARCIA, W. Hypothesis generation and data quality assessment through association mining. In *IEEE ICCI* (2010), F. Sun, Y. Wang, J. Lu, B. Zhang, W. Kinsner, and L. A. Zadeh, Eds., IEEE, pp. 659–666.

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

[11] GAMBLE, M., AND GOBLE, C. Quality, trust, and utility of scientific data on the web: Towards a joint model. In *ACM WebSci'11, Koblenz, Germany.* (June 2011), pp. 1–8.

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

[13] GIL, Y., AND RATNAKAR, V. Trusting information sources one citizen at a time. In *Proceedings of the First International Semantic Web Conference on The Semantic Web* (2002), Springer-Verlag, pp. 162 – 176.

[14] GOLBECK, J. Using trust and provenance for content filtering on the semantic web. In *Proceedings of the Workshop on Models of Trust on the Web, at the 15th World Wide Web conference* (2006).

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

[16] GUÉRET, C., GROTH, P., STADLER, C., AND LEHMANN, J. Linked data quality assessment through network analysis. In *ISWC* (2011).

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

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

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

[20] HOGAN, A., HARTH, A., PASSANT, A., DECKER, S., AND POLLERES, A. Weaving the pedantic web. In *3rd International Workshop on Linked Data on the Web (LDOW2010), in conjunction with 19th International World Wide Web Conference, CEUR, 2010.* (2010).

[21] 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).

[22] 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.

[23] JACOBI, I., KAGAL, L., AND KHANDELWAL, A. Rule-based trust assessment on the semantic web. In *Proceedings of the 5th international conference on Rule-based reasoning, programming, and applications series* (2011), pp. 227 – 241.

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

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

[26] KIFER, M., AND BOLEY, H. Rif overview. Tech. rep., W3C, June 2010.

[27] KITCHENHAM, B. Procedures for performing systematic reviews. *Keele, UK, Keele University 33* (2004), 2004.

[28] LEI, Y., NIKOLOV, A., UREN, V., AND MOTTA, E. Detecting quality problems in semantic metadata without the presence of a gold standard. In *EON* (2007), pp. 51–60.

[29] LEI, Y., UREN, V., AND MOTTA, E. A framework for evaluating semantic metadata. In *Proceedings of the 4th international conference on Knowledge capture* (2007), no. 8 in K-CAP '07, ACM, pp. 135 – 142.

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

[31] MILLER, P., STYLES, R., AND HEATH, T. Open data commons, a license for open data. In *WWW2008 Workshop on Linked Data on the Web* (2008).

[32] 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 Med 6*, 7 (Jul 2009), e1000097.

[33] MOSTAFAVI, M., G., E., AND JEANSOULIN, R. Ontology-based method for quality assessment of spatial data bases. In *ISSDQ'04* (2004), vol. 4 of *GeoInfo Series*, pp. 49–66.

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

[35] P.N., M., H., M., AND C., B. Sieve: Linked data quality assessment and fusion. In *Invited paper at the 1st International Workshop on Linked Web Data Management (LWDM 2011) at the 15th International Conference on Extending Database Technology, EDBT 2012* (March 2012).

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

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

[38] RULA, A., PALMONARI, M., AND MAURINO, A. Capturing the age of linked open data: Towards a dataset-independent framework.

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

[40] 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.

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

[42] 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.