the increasing size and availability of web data made its quality principles become a core component in many applications; these principles play a critical role in the success of many systems and organizations. We think that data quality is subjective and that "beauty is the eye of the beholder", we think that data quality cannot be assessed independently by the user easily, but it is related directly to the satisfaction of the end customers. It was found out that many data quality problems are in fact “data misinterpretations”, or problems with the data semantics [4]. In this paper we try to list some of the most important quality principles within the context of semantic web.

The rise of Semantic Web was accompanied by a tremendous increase in the amount of data; everyone has the ability to publish and retrieve information to be consumed or integrated into his applications. The Semantic Web has significantly changed lots of people vision about the Internet. However, the huge growth of data and its heterogeneous sources is accompanied by the need of creating tools and measures to help data consumers identify their needs and evaluate the quality of these sources.

Lots of organizations are trying to leverage external data sources like social media feeds, weblogs, sensor data or data published by governments or organizations [1], doing that efficiently will lead to produce more informed business decisions. These external sources exhibit heterogeneous models, formats and terminologies, thus lots of work is being done in order to improve the quality of this structured knowledge. The Semantic Web can be looked at as a “global database” [2] that machines can directly access and naturally understand [3]. Data quality involves data management, modeling, analysis, storage and presentation, quality control and assurance [5], the actual value of data is realized when it is used, thus the quality relates directly to the ability of satisfying the users continuous needs; saying that we think it is important to automate the process of controlling and assuring data quality in order to respond to the changing demands in a timely manner.

For lots of users, especially decision makers, finding the accurate necessary information when needed is very important. At a very high level; the ease of finding and retrieving information can be one of the main pillars of data quality. From this point, we tried to list some measures and characteristics to describe the quality of a particular linked dataset.

Semantic Data extends the requirements of classical data as quality is not restricted only to the data or data source only; the quality of the used ontologies and the links should be also taken into consideration. Recently, the development and application of ontologies have been gaining big momentum. Many semantic groups have been contributing in the production of different ontologies; however, most of the newly created ontologies were usually reused from previously existing ones; so it was like gathering bits and pieces from several ontologies to combine one. Given the heterogeneous structure and content of ontologies and the diverse background of the ontologies` makers, deciding what ontology to use is one of the most difficult and important tasks any developer should do.

Ontologists have been using several existing methodologies and design patterns that support the building process; this can guarantee a certain level of quality in the resulting ontology, and would contribute to better quality semantic systems.

Identifying good data resources to simplify the task of consuming high-quality data have been tackled by several approaches. In [6][7] a resource is ranked by the quality of the incoming and outgoing links. Moreover, “Sieve” [8] is a framework that tries to express quality assessment methods as well as fusion methods.

Publishing Linked Open Data has 4 main principles that should be followed:

* Make the Data available on the web: Assign URIs for names of things
* Make the data machine readable: Use HTTP URIs so that looking up these names is easy
* Use publishing standards: when the lookup is done provide useful information using standards like RDF
* Link your data: include links to other resources to enable users discover more things.

By following these guidelines, a certain level of uniformity is achieved, which will increase the usability of data. Certain quality measures should be accompanied with each of these points; we can summarize the quality measures as:

* Quality of the raw data: Identifying noise, duplications, abbreviations … etc. Tools such as Google Refine [ref] have been created to tackle these issues.
* Quality of the semantic conversion: This is the process of transforming “normal” raw data into “rich” data i.e. **input**: [tabular data] 🡪 **output**: [RDF using x Vocabulary]. The availability of high quality vocabularies and the efficiency of the discovery process is a major factor.
* Quality of the linking process: The quality of linking a dataset with others, Silk [ref] is a popular framework that tackles this issue.
* Quality of reasoning: How to make the data more explicit, or how to turn the data into useful information [9].

These quality measures can be highly automated; however, the heterogeneous nature of Linked Data makes it very hard to identify and point out data quality indicators without the introduction of social mechanisms i.e. crowdsourcing. Even if there are methods that can “objectively” rank resources we strongly think that human interaction (reviews and rankings) is still needed to review and adjust the ratings.

* **Data source measures:**
* **Accessibility**: This measure relates to the proper functioning of all access methods and protocols, the de-referencability of the URIs and the ingoing and outgoing links.
* **Authority & Sustainability**: Is the data source provider a known credible source or is he sponsored by well-known associations or providers? Are there are credible basis for believing the data will be maintained?
* **License**: Is the data source license clearly defined?
* **Trustworthiness & verifiability**: a data consumer should be provided with a mean to examine the correctness and accuracy of the data source, the consumer should be also sure that the data he receives is the same data he has vouched for and from the same resource; this can be ensured using digital signatures thus verifying all possible serialization of that data.
* **Performance, low latency resource, high throughput**
* **Core data measure:**
* **Accuracy**: Are the nodes referring to factually and lexically correct information?
  + **Referential correspondence**: Is the data described using accurate labels without duplications? The goal is to have one-to-one references between data and real world references.
  + **Boundness**: is the data that I have clean and not polluted with irrelevant or outdated data?
  + **Consistency**: does the data contradict itself? I.e. is the population of Europe the same as the sum of the population of the European countries? To achieve that we need to validate the underlying vocabulary and syntax of the document with other resources; data inconsistencies can be caused by the usage of homogeneous data types or when a source redefines a vocabulary that in his opinion is not expressive enough.
* **Comprehensibility**: This measure is very important when it comes to humans interactions with the data, whether it was for evaluation or consumption purposes. The data (things) labels should be understandable to humans, conveys logical meaning to the described entity and allow easy consumption and utilization of the data. Moreover, if a thing is described using multiple labels (a set of owl:sameAs things have multiple labels) how can we specify which label is canonical?
* **Completeness:** Do we have all the data needed that represents all the information related to a real world entity? Moreover, is the data related or linked to this set complete as well? I.e. All European countries, all the French cities, all the street addresses, all the postal codes … etc.
* **Typing**: are the nodes properly typed as resources or just as string literals? Having the node properly typed will allow users to ask more questions and infer facts about a certain entity.
* **Provenance:** provenance in the Semantic Web is considered as one of the most important indicators of "quality." Data sets can be used or rejected depending on the availability of sufficient and/or relevant metadata attached.
* **Versatility**: Can the data provided be presented using alternative representations, this can be achieved by conversion into various formats and that the data source enables content negotiation.
* **Traceability**: can we identify the different sources of my data? Can I tell what portions of my data came from where?
* **Modeling related measures:** a very important aspect that affects the quality of a dataset is the **quality and suitability of its data model** with the intended usage, a model`s quality strongly depends on the following aspects:
* **Modeling correctness**: Is the data structure properly modeled and presented?
* **Modeling granularity:** does the model capture enough information to be useful?
* **Modeling consistency:** do I have to ask for many permutations of the same query or is the direction of relations consistently done?
* **Datasets linkage measures:**
* **Connectedness**: is the combination of datasets done at the correct points?
* **Isomorphism**: are the combined datasets modeled in a compatible way? Are the combined models reconciled?
* **Directionality**: is the data consistent in the direction of relations?
* **Global quality measure (I donno about the name here but it’s a measure that is applied for the source, the core data, the model and the links):** These quality measures are applicable to all different aspects of a Semantic system (the data source, the core data, the model and the links).
* **Timeliness**: Is the information up-to date, does the data source contain the latest core data presented with the last updated model? Are the links from and to the data source updated to the latest references? To achieve that a source can state the updates and validations frequency in the *changefreq* attribute or by clearly identifying the last update or modification timestamp. Failing in updating the source data will increase the change that the referenced URIs have changed, so another aspect of up-to-date information is the de-referencability of all internal and external URIs.
* **History**: Can we keep track of who edited my data and when?
* **Freshness**: The ability to replicate the remote repository into local triple store and maintain the timeliness of the replica.

|  |  |
| --- | --- |
|  | [1] Danah Boyd and Kate Crawford, "Six Provocations for Big Data," Computer and Information Science, vol. 123, no. 1, 2011.  [2] <http://www.w3.org/DesignIssues/Semantic.html>  [3] tim berners-lee weaving the web. Harper, San Francisco, 1999  [4] Improving Data Quality Through Effective Use of Data Semantics  [5] [Principles of Data Quality](http://www2.gbif.org/DataQuality.pdf) [1] Arthur D. Chapman  [6] Hierarchical Link Analysis for Ranking Web Data  [7] Sindice at SemSearch 2010  [8] Sieve: Linked Data Quality Assessment and Fusion  [9] [Scalable Authoritative OWL Reasoning for the Web](http://www.deri.ie/fileadmin/documents/DERI-TR-2009-04-21.pdf)  [10] Executing SPARQL Queries over the Web of Linked Data  [11] Thomas R. Bruce and Diane I. Hillman 'The Continuum of Metadata Quality'  [12] A Framework for Information Quality Assessment |