Quality methodology for a transformation process to Linked Open Data from structured data sources: the role of Ontology for the definition of the Semantic Data Model.

Abstract

The emergence of Linked Open Data has allowed structured data scattered on the Web to be interconnected for the development of new knowledge. It means that this scattered data can be transformed to a new type of data, linked data, which allows for linking with the rest of the related data available on the Web. However, for the results of this transformation to really achieve the development of this knowledge, it depends to a great extent on the quality of the data once transformed. Our research is focused on making this transformation process include improving quality to ensure that the documents generated are correct. Based on the principles of measurement theory, we propose a quality methodology that describes the quality guidelines present in each design phase of the transformation process, fundamentally those which determine the quality of the design of the semantic data model in question. The identification of these quality guidelines is resolved by discovering their relevance for a set of structured data. Considering that a linked open data model adopts ontological principles, we analyze how such quality guidelines affect this process when we reuse or do not reuse an existing ontology and how to carry out its construction so that the generated data can be integrated with other data sources in an interoperable way. Finally, the scope of the quality methodology is analyzed so that it can be used in the design of an automatic transformation process of structured data to Linked Open Data.

Keywords: quality methodology, quality guideline, quality dimensions, existing ontology, semantic data model.

Introduction

A surprising amount of structure data that is not linked can still be found. It is a problem that the founder of the World Wide Web Tim Berners-Lee highlighted when he released a group of good practices better known as "The Four Rules" that warranted positive use and behavior of Linked Open Data (LOD) [1] and that currently continues in force. Called LOD or semantic data, the data that is available on the web under an open license, they also use different technologies so that a person or machine can explore the web of data to infer information. Currently, researchers dedicate great effort in determining a focus for the construction of semantic data that reflect the structure and complexity of the domain, from non-semantic information. Both aspects are elements that are defined mainly in a process of transformation. Hence the quality of the data obtained is determined by the quality of the transformation process itself. The term of data quality was not new, many of the quality aspects defined for the LOD were analyzed according to the existing literature about quality in structured data. Quality is a concept that is commonly defined as "fit for use" [2]. To make the measurement and quality improvement more organized, quality dimensions have been created. The term dimension refers to a unique aspect of quality such as availability, consistency or understandability.

Suitability for use is a definition of utility rather than quality. It is the knowledge of this utility that allows to link coherently groups of data for its analysis. This research focuses on the importance of creating an approach that helps to obtain correct LOD to ensure its usefulness. For this we define a quality methodology focused on the transformation process of structured data to LOD. The main goal is to make known and describe the quality guidelines involved in this process of conversion and thus ensure that the structure and semantics in the data obtained is as expected. For this it is necessary to (1) identify the quality dimensions that need to improve the original data before the transformation process, to (2) define the characteristics that the LOD must have to inhibit quality problems and to (3) determine the quality dimensions that ensure a correct LOD construction process (by defining a mapping mechanism). Its importance is to consider the improvement of the quality of the data from the transformation process itself and, in addition, that the data are suitable so that, once they gain the characteristics of LOD, they can continue to improve those quality dimensions that could not be improved when they were in a structured format.

For the design of the methodology, several cases of use were selected with the objective of identifying the need for improvement that arises in a process of transformation from structured data to LOD. These are the World Bank's Data Source[1] and Territorial Statistics Data Source[2]. Both provide statistics from different sectors in XML format, for example, the World Bank offers CO2 emissions by country.

Latest situation

The use of Methodology in Data Quality. The literature reflects that the use of a quality methodology is very efficient for the publication of Linked Open Data (LOD) in different society sectors [3-5]. It is characterized by being more detailed on how to select appropriate assessment techniques, providing more examples and related contextual knowledge. In addition, they identify all types of problems, regardless of which improvement techniques can or should be applied [6].

Efforts to obtain Accessible, Interoperable and Reusable LOD. There is a greater effort to incorporate, during the construction of RDF documents, some aspects of quality; mainly those that guaranteed their linking with the rest of the data on the Web. This is due to the extensive analysis of the LOD quality conducted in the last decade as a result of the major problems that this data can generate when published on the Web without adequate quality. DCAT (*Data Catalog Vocabulary*) [7] is one of the quality tools that is currently working to achieve quality in the LOD; is a vocabulary designed to facilitate interoperability between the catalogues of data published on the web, a very important aspect of quality that this data must meet. It is a specification that makes use of a standard model, which guarantees interoperability, but it remains a challenge to include ontological commitments, such as avoiding the use of blank nodes, a requirement that determines the quality because it precisely decreases this interoperability. As far as being able to describe the complexity of the domain is also another of the current problems that this specification shows unresolved even though it provides the inclusion of metadata. It is necessary to define a quality improvement mechanism that does not contain the impossibility of including the ontological commitments and generate LOD as close as possible to reality. These actions can be fulfilled if the quality characteristics to be considered during their creation are indicated step by step. Another research that currently stands out are the principles of FAIR (*Findable, Accessible, Interoperable, Reusable*) [8]. They are very useful because they define requirements that LOD must meet so that they can be reused. These requirements ensure the inclusion of some of the aspects of quality but are characterized by being relatively undetailed and this sometimes can impede when it comes to describing the complexity of a resource.

Mapping Mechanism Definition regarding with Data Quality. The ability to describe the complexity of the domain and its linkage with other data are the most important properties expected in the LOD for its reuse. Its obtaining from structured data constituted the first investigations carried out. In these investigations the researchers developed approaches that aimed at the definition of a mapping mechanism capable of detecting the complexity domain [9]. It is also observed that in order to detect this complexity, it makes use of an existing ontology or by means of the execution of queries that link entities once transformed. The various mapping initiatives range from the elaboration of proof-of-concept projects [10, 11], to the creation of domain-specific projects [12, 13] and the creation of applications [14-16]. These woks are fundamentally focused on the efficient creation of assignments in a mapping mechanism for the design of an ontology, but they do not include data quality tasks that ensure accurate LOD generation. This means that until the data is obtained it is not possible to determine its real utility, for this it is necessary to use an evaluation tool. Other investigations provide quality dimensions that help the correct design of ontologies. The most recent research [17] made a systematic survey on quality dimensions applicable to ontologies in the Semantic Web. However, we consider it important to investigate in another approach, the role of an ontology in the analysis and improvement of quality during the creation of LOD considering that it is the reuse of an existing ontology that determines in many cases the quality in the LOD. A correct transformation process to LOD will allow this analysis because it highlights the importance of reusing an existing ontology. In turn, this task provides information on how to design a mapping mechanism in accordance with the quality dimensions that must be met by the LOD, considering that it is this

[1] https://datos.bancomundial.org/indicador

[2] https://www.bcn.cl/siit/estadisticasterritoriales/

mechanism that defines the structure of this data. Therefore, an interesting objective to develop is to discover what are the qualities expected in this data and then define a mapping mechanism to achieve it.

Lessons to ensure quality in the LOD. In terms of how to ensure that LOD are correct we have seen that the reuse of an existing ontology is a good solution for quality compliance because it is known beforehand that they are already validated. But in some cases, this may not be possible due to the specific characteristics of the domain. On the other hand, although there are good practices to make LOD interoperable and reusable, the inclusion in the data of other qualities to improve other aspects of quality will give them greater reliability and will even help them have more resources to be able to describe the complexity of the domain. In addition, there is a strong dependence between the different dimensions which indicates that in order to consider one dimension as correct it is necessary that the rest of the dimensions related to it are also correct [18].

Contribution. Our research is based on experience on which quality dimensions need to be improved in a process of transformation from structured data to LOD. The main goal is to ensure that the data obtained (LOD) are correct and that they contain the necessary characteristics so that they can adequately describe the complexity of the domain and therefore ensure that they are reusable. For this, based on this experience, we define a quality methodology that describes in each step of the transformation, the quality guidelines that must be met. This allows us to discover how it affects the identified dimensions, the reuse or not of an existing ontology, a task that has been observed to be fundamental in the determination of the quality of the LOD. Also, part of this research is the analysis of the qualitative characteristics that are expected in a mapping mechanism because it constitutes the nucleus of a process of design of LOD.

1 Quality analysis in the data sources: Identifying needs for improvement

Quality tasks performed on a data source are solved by discovering what you must improve. An analysis of the problem consists of discovering the phases of the transformation process and incorporating the quality tasks that proceed in each one of them. The selected data is in XML (*Extensible Markup Language*) format. When we analyze their structure, we observe that they do not use a specific standard for the domain of the problem, but a generic standard for expressing tabulated data. Although this structure constitutes a specific language, it is not defined correctly because it does not use an XSD (*XML Schema Definition*) that allows its validation. Another way to carry out this validation is by defining a namespace where the corresponding XSD file is associated. Documents identify a name space, but it is not associated to a valid XSD. The existence of an XSD is important because it defines an XML-based standard that can be used to check if the source data is as expected. Although the information is provided by a reliable source, we cannot ensure that the structure and data are completely correct because there is a risk of generating erroneous conclusions at some point. This tells us that two extra validation jobs should be included in the methodology, these are: the validation of the structure and the validation of the data. Therefore, the first challenge is:

(a) *Validation of the original documents:*

The absence of a validator does not necessarily indicate that the data is wrong. It indicates that it is not possible to verify automatically that the data is correct when the objective is to generate reliable data.

Continuing with the analysis of the data, we discover that there is information that cannot be read by computer. For example, in data relating to CO2 Emissions, the unit of measurement is defined as follows:

Fig. 1. Record of CO2 Emissions by country and year.

```
<record>

    <field name="Country or Area" key="ABW">Aruba</field>

    <field name="Item" key="EN.ATM.CO2E.PC">Emisiones de CO2 (toneladas métricas per cápita)</field>

    <field name="Year">1964</field>
```

```
  <field name="Value" />

</record>
```

The unit of measure is there, but in a text. The information cannot be converted to RDF the way it is because it is information that is lost during a query or during an automatic analysis and, above all, it is not possible for you to do an integration task with other related data in the rest of the world, such as being able to make a CO2 comparison between two countries. This problem gives us another challenge:

***(b)** Extend the original data.*

This step consists of looking for uninterpretable information. We define noninterpretable information as: the existence of non-automatable fields, as happened with the unit of measurement in the previous example (Figure 1.), the existence of information of interest in the documentation, and the existence of information that is understood information that can only be deduced by a human being.

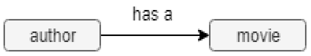
All these tasks are focused on achieving, as far as possible, that the data are correctly interpreted by computer. It means that, once this phase has been completed, the data is ready to be transformed. Therefore, the next challenges are those related to the creation of RDF documents.

***(c)** Construction of RDF documents.*

The semantic data model using RDF is based on the definition of *subject-predicted-object* triple, as shown in Figure 2. From a structural point of view, this semantic data model represents a graph of information. Both the *subject* and the *object* are nodes in the network, while the *predicate* will be an edge in the network. The *subject* is the element that identifies a described resource; this described resource constitutes the *object*, and the *predicate* indicates the relationship that exists between the *subject* and the *object*. A mapping mechanism is used in order to define the structure of the semantic data model, which is nothing more than establishing which elements are candidates to be a node and the relations between them.

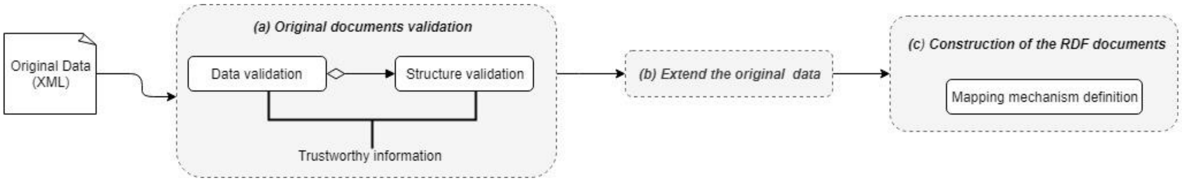
This phase fundamentally consists of determining the optimal way to choose which elements are candidates to be a node; the relations between them are imposed by the data domain itself. It should be borne in mind that to make the construction of documents requires of all the information, both the data of the XML document as well as extended data (step ***(b)***).

Fig. 2. Example of RDF triple.



The challenges identified above form the basis for the design of the quality methodology because they reflect the tasks to be carried out in a transformation process and the needs for improvement in each of them. Figure 3 shows how these challenges shape the transformation process for the generation of correct RDF documents.

Fig. 3. Quality tasks to develop during the transformation process.


Original Data (XML)
(a) Original documents validation
Data validation
Structure validation
Trustworthy information
(b) Extend the original data
(c) Construction of the RDF documents
Mapping mechanism definition

2 Resolution challenges: Identifying Quality Dimensions

The selection and definition of quality dimensions is the first task in a data quality improvement process. Each of these dimensions are just characteristics of the information that need to be improved. The following table shows the characteristics of the data that will be improved in the transformation process.

Table 1. Quality Dimensions identified during the transformation process.

Improvement phase/Challenges	**Dimensions**	**Needs for improvement**
(a) V*alidation of the original documents:* Ensure data reliability	**Syntactic Validation** *It consists in detecting whether the data complies with the syntactic rules defined for the domain to which they belong, and the rules of the format used to represent it.* **Semantic Accuracy** *Measures whether ether the data is real for the concept it represents and the degree of accuracy for the domain.*	Avoiding anomalous interpretations by the analysis tool that implements the quality methodology to prevent an incorrectly generated document.
(b) E*xtend the original information:* Ensure that the generated documents contain more complete information and is correctly represented so that it can be processed automatically.	**Completeness** *A data source is complete when the information covers all the needs for a specific task.* **Interpretability** *It measures the degree to which the data can be processed by machine.*	There is information on data source of interest that cannot be interpreted automatically. The data of the source must reflect the complexity of the domain. Improve the usefulness of the data source.
(c) Construction of RDF documents*:* Ensure that the information uses a correct semantic data model. This will allow that data source can be scanned and processed by a computer and by a human in an efficient way.	**Interlinking** *It measures how effective the union between the nodes of a graph is. It is known as "mapping coherence".* **Interoperability** *Refers to the extent to which data from a source allows it to be able to exchange information with other data source.*	Avoiding the loss of information in case a node is unable to be linked. All source information is obtained. The generated documents need to be linked to other data source to infer new knowledge and convey information.

This type of tasks essentially helps to prepare processable data, it does not mean that some of the quality aspects or dimensions should not be improved or verified once transformed. This is because the obtained data contains a new semantically more complex structure and because of the limitations of structured data for automatic comparison with other data sources. Table 2 shows the quality aspects that need to be verified in the documents obtained. Below is a diagram that reflects the process of improvement. These make up all the dimensions identified.

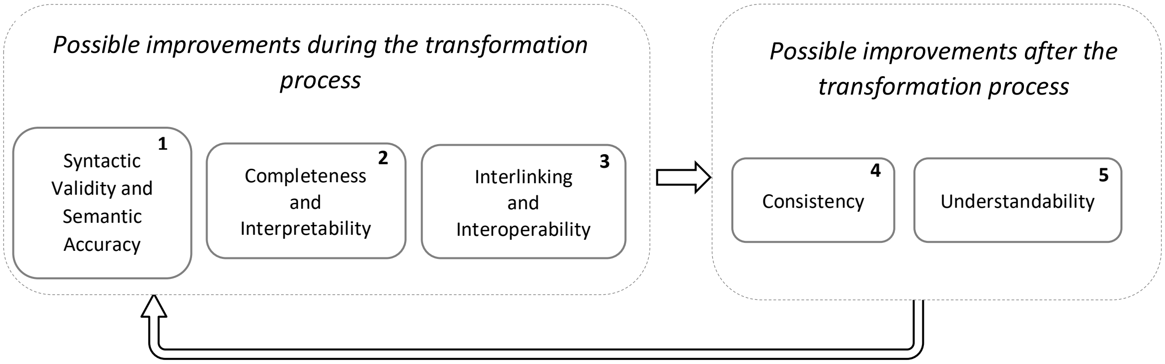
Table 2. Quality Dimensions that must checked once the transformation process has been completed.

Quality Dimensions	Needs for improvement
Syntactic and Semantic Accuracy	Documents comply with the RDF standard. The information contained in the triple is correct for the domain they represent.
Consistency *It is to verify that the data of the source does not contradict itself and is consistent with other data in its domain.*	Prevent a situation in which there is no information that is contradicted in different points of the graph.
Understandability *It refers to the capacity of the characteristics of a data source to represent information in an orderly, clear and readable manner for data consumers.*	Ensure that the information can be recoverable and analyzed efficiently.

As mentioned above, some dimensions can be improved once the data has been transformed. The dimensions regarding completeness may be improved by the ability to link them to other data on the Web, a task that is not possible in a structured format. In addition, errors can locate in this phase that are present in the original document which would have gone unnoticed, therefore being unable to be compared with other domains.

Figure 4 shows how all these quality dimensions form a cycle of improvement of the quality of data in this process. It indicates an order of execution of the quality tasks related to each quality dimension to ensure that the generated data is correct and adequately reflect the domain to each they belong.

Fig 4. Quality Dimensions that ensure obtaining accurate LOD.


Possible improvements during the transformation process
1
Syntactic Validity and Semantic Accuracy
2
Completeness and Interpretability
3
Interlinking and Interoperability
Possible improvements after the transformation process
4
Consistency
5
Understandability

It is important to clarify that the dimensions which have been identified are the most important in the transformation process. There are others such as relevance, timelines, trustworthiness, licensing, performance and security that have not been included because it makes no sense to evaluate them in this process except for trustworthiness, but in this case, it is not necessary to check this considering that the data is known to come from a reliable source.

Each of these quality dimensions are connected to a group of quality metrics and in turn each metrics represent a different quality problem. This relationship suggests that after the identification of the dimensions, the next step should be to specify which characteristics inhibit the associated quality problems in each dimension, considering the main objective of this research. The design of the methodology is the effect that the accomplishment of this task produces.

3 Quality Methodology

Its design is based on the following assumption: the need to add more qualities to the data during a process of transformation from structured data into LOD so that the documents generated are correct; in other words, that the documents generated adequately describe the domain to which they belong and are reusable. The fundamental objective is, firstly, to resolve the quality problems that may exist in a structured data source and which prevent them from being transformed into correct semantic data; and, secondly, to specify the qualities that each element of the RDF documents. It allows the organization of different phases of improvement previously analyzed (see Fig. 3.), in terms of a set of basic steps as part of the generation of accessible and reusable RDF documents with reliable information from structured data. We incorporated the *Remarks* section with the purpose of specifying the scope of application of these qualities. Sometimes these specifications are general, that is, for the whole phase, and in other cases they are specific for one point of the phase and even for both cases (general and specific).

Phase 1: *Analysis of the case of use*

Remarks: It is to assess the current status of XML documents.

Check using a standard justifiable in XML documents: (1)

Locate all the relevant information in the documentation provided by the data source: (2)

Phase 2: *Validation of input data*

Remarks: The data that the application will work with represents, in XML documents, information in tabular format: (3) y (4)

The content of each cell of the previous table and the number of these depends on the type of specific information in the data source that is being converted: (5) y (6)

Syntactic Validity

1. Check the document structure. Some mistakes we might find are:
   - The number of cells in some rows of the table does not match with what is expected.
   - The XML element used to designate a cell does not have the expected name.
2. Each XML element in the document has several restrictions depending on their type that should be properly verified:
   - Required or optional attributes that are present in each element according to its type.
   - Content type of each attribute.
   - Type the content of each XML element.

Semantic Accuracy

3. The type of each data cell shall conform to its own domain data source, both the type and the range of values that must been admitted. For example, we cannot give a valid column on temperatures of 1000 ºC in a data source for weather information.
4. The number of data present in each row of the table should be adjusted to that required by that data source.

Phase 3: *Extend the original data*

Remarks: All tasks related to the extent of the information will depend entirely on the data source with which you are always working. For this reason, it is impossible to construct a generic algorithm that automatically performs this task for all data sources if the semantic data generator is used.

Completeness and Interoperability

5. Identification of non-automatable fields:
   - Natural language information.
   - Information provided through external documents. In these cases, the content of the cell would be a reference to a different document.
   - No text information in formats such as graphs.
   - The theme of the data source.
   - The origin of the data.
6. Extract new information in the documentation that provides the data source.

Phase 4: *Construction of the RDF documents*

Remarks: Although it is expected that the documents generated always comply with these guidelines, it will be the characteristics of the domain which allow or not their application; data cannot be improved, and neither can relationships that are not present or inferred from the original data source. To employ an existing ontology promotes the integration of semantic data generated within previously existing data sources; it will tell us which classes an attribute used. If it is not the case, we will define them always trying to facilitate integration.

The definition of semantic data model may be achieved using existing technologies, such as RDF Schema or OWL (*Web Ontology Language*)[3]. This will facilitate the development of the application to be able to use already developed tools for managing this definition.

4.1 Define Semantic Data Model

Interlinking

7. Determine classes (types) containing our semantic data model.
   - Follow the most reusable approach possible:
     - Prioritize classes for using a value. Section 5 provides a mapping mechanism that warrants it.
8. Determine the attributes that form each class:
   - Defining attributes indicating their name and function.
   - Defining restrictions of these attributes.
   - Definition of the cardinality of these attributes.
9. Define relationships between classes:
   - Declare all possible relationships (*degree of clustering*).
   - The number of relationships between nodes is as close as possible (*centrality degree*).

   As relations we mean subclass relationships or aggregation. The inclusion of relationships will be imposed in the case of using an existing ontology but should indicate otherwise. These relationships can facilitate the integration of our data source with other external data sources, especially for those types that we defined specifically for this problem.

   The degree of clustering which we refer to is the density of the graph generated from the schema of the data source that we analyze. A very dense graph indicates a strong grouping between the different nodes that form it. If we consider that these nodes are the entities includes in our data source, a high density indicates that the entities are very grouped; in other words, very interconnected.

[3] https://www.w3.org/OWL/

The degree of centrality refers to the capacity that a set of data shows to obtain information. With this we mean that the centrality of a data can helps to determine a situation of the real world that initially was not represented in the data but that can be inferred from the relations between these.

Interoperability

- Use the relevant vocabulary domain to define classes and their relationships and attributes.
- Avoid creating blank nodes.
- Appropriate use of language symbols and terms

4.2 Detecting the nodes of the graphs

Remarks: The different nodes of the graph will be the instances of the classes previously defined in the semantic data model determined in the previous point. These instances later (in the next phases) will be stored in the form of triples in the generated RDF document. This class must get all instances for each of the previously defined classes (upon the definition of the model) and assign each of them the identifier will allow them to be used.

10. Determine the nodes of the graph:
    - Employ a standard of universal identifiers to represent each node.
    - If this is not possible, it is advisable:
      - That each universal identifier must be as far as possible an alias and must be referenced resource. One way to achieve this would be to establish *owl*:*sameAs* to known URI aliases [19]. To find a known URI alias you can rely on Wikidata[4] or DBPedia[5], or any other source of information that has semantic data related to the domain of the data source in question.
      - To make use of terms already defined for that domain.

Identifiers should never change. A node whose identifier changes over time will lose its identity. A change cannot be fixed simply by being renamed in a given data source, because that will continue without updating that identifier in other data sources that can be found in different locations.

4.3 Generate triples

Remarks: This is the most passive step in terms of the quantity of quality guidelines to be analyzed because this step is only to fill in the graph.

11. Collect from the original data source the values of the different attributes of each node:
    - It must be those indicated by the semantic data model and must exist in the original data source.
    - It must be ensured that the way to represent the resources uses the appropriate types classes and their corresponding values comply with the lexical syntax of the assigned type.

Phase 5: Validation the generated RDF document

Remarks: The linked data contains a set of inherent features that must be verified. Although the data has already been transformed considering a group of improvements, it is essential to verify that the generated RDF document complies with the structure defined in the previous semantic model. This is essential to guarantee the correction of both the generation process and the information previously provided by the original data source. During this validation it must be verified that the documents are syntactically and semantically accurate. This is accomplished if it is determined that they are valid RDFs, and for this it is possible to rely on tools that can automate the task, like *Shape Expressions* [20].

[4] https://www.wikidata.org/wiki/Wikidata:Main_Page

[5] http://es.dbpedia.org/

Syntactic Validation

12. Syntax validation:
    - Define and verify syntactic rules of the RDF format.
    - Verify that the graph's resources belong to the class to which it corresponds.
    - Verify that the values of the attributes are lexically well represented.

Semantic Accuracy

13. Semantic data model validation:
    - Verify that the data types are valid.
    - Verify that each node contains the required attributes.
    - Verify that the cardinalities of the attributes are correct: There cannot be a number of instances or values associated with another instance that is incompatible with what is defined in your model.

Consistency

14. All rules that represent relationships between defined data are satisfactorily fulfilled:
    - The rules of the problem for the search for incompatibilities. For example, declaring a rule in which to verify that a person's age is greater than 16 years if the marital status is married.
    - Range of possible values for each attribute.

Understandability

15. The labeling of classes, properties and entities must be readable. At this point it may be interesting to use a standard for the definition of information. This would also allow the validation of entity values through automated tools (*Shape Expressions* [20] for example).
16. The document will have metadata to facilitate its use.
17. The URI that must be included will have a well-defined format [19].

Validations related to the first three points of the above list may be automatically determined from the ontology. The use of Shape Expressions [20] can be of great help to accomplish this task. On the other hand, it would be very interesting to have used technologies existing for the definition of ontology, as this will facilitate the possibility that there are tools that automate the task of building the validator that checks the ontology used.

However, for validation of the fourth point of the list, it will be necessary to manually write the validation expressions that are in charge of verifying that the relationships between the data are correct. However, although the validation expressions must be written manually, we will also be able to rely on existing technologies and tools to accomplish this task. Note that at this point the RDF document is already built and validated syntactically and semantically. We therefore have a source of semantic data on which we can use query languages as SPARQL to simplify the writing and execution of these expressions. Moreover, from this point, due to its ability to be linked with other data, it will be possible to obtain more information and to carry out a check of relations which was not possible to evaluate when the information was isolated. Regarding this point it is advisable to make comparison with external sources, such as Wikidata to verify the reliability of the information through the relations.

The characteristics that the LOD must meet have been specified. A mapping mechanism that indicates the structure that RDF documents must fulfil (as indicated in "Section 1") has not yet been defined. This is because, we consider it important to first analyze the quality dimensions that are affected when using an existing ontology or not in the construction of RDF documents. This will allow the development of an approach centered on the quality dimensions that a semantic data model needs to satisfy regardless of whether an existing ontology is reused or not.

4.1 Implications of the use or no-use of existing Ontology for the definition of the Semantic Data Model

The design of the semantic data model consists of using an existing ontology if the characteristics of the domain allow it or otherwise, creating a new one. It would be more convenient to use a serial ontology to represent the information, and if necessary, add the data of interest. W3C offers a standardized ontology that can be used as a basic element for information (*RDF Data Cube Vocabulary*[6]). This reuse ensures the use of a well-defined ontology that has been validated and therefore meets the quality standards expected for semantic data. If it is decided to reuse an existing ontology, the quality dimensions that involved in this conversion process are analyzed with these benefits in mind.

The reuse or creation of an existing ontology may or may not affect a set of quality dimensions, mainly those that determine the quality of the semantic data model. The quality dimensions are completeness, interconnection and interoperability: ontologies help to integrate data (*Completeness*) from multiple sources in an interoperable way (*Interoperability*) and help to consult and retrieve the specific information required (*Interconnection*) [21]. The following table describes how this decision affects the improvement of data.

Table 4. Implications of the use or no-use of existing Ontology.

Quality Dimension	Existing Ontology selection	Extend or Creation of Ontology	Recommendations
Completeness	Allows to verify that all the necessary elements or individuals are provided to represent all the data of the source. It helps to detect data that has not been included in the extension phase. This means that it identifies errors in previous phases.	It is not affected because the model is made according to the domain analysis.	Not all completeness errors detected by the ontology are to be remedied in the extension phase. It is advisable to take note of the deficiencies found to try to solve them the data has been transformed.
Interlinking	The semantic data model will be responsible for indicating which elements will have identity and belong to complex classes (they will not be simple types) and which elements will be simple values, and how the nodes should be grouped defined the structure that must be fulfilled by the documents to be generated.	It helps to build more cohesive graphs by having a greater number of elements directly or indirectly related. The selection of a correct structure of relationships between the different nodes of the graph will help to build easily reusable links.	Strong clustering degree. Each node has a value of similar centrality.

[6] https://www.w3.org/TR/vocab-data-cube/

Interoperability	Is certain that the elements are used in other sites.	It is important to verify that the defined classes and attributes of the data are standard and are related to the data of its domain.	Use names and references usually used in the context on which you work and that are standardized.

In summary, Table 4 shows that it is always advisable to analyze completeness, interconnection and interoperability regardless of whether an ontology is reused or not. For this we rely on quality guidelines defined in the quality methodology.

The study of the quality guidelines that must be fulfilled in the RDF documents and implications of the use or no-use of existing ontology (just analyzed in Table 4) allows to define a mapping mechanism that ensures the correct design of the semantic data model because it reveals the characteristics expected of it: *allows us to link the data with other data of your domain and to obtain information that we did not have before*. Mapping mechanism proposal is defined below.

5 Mapping Mechanism proposal: A designing process of reusable Semantic Data Model

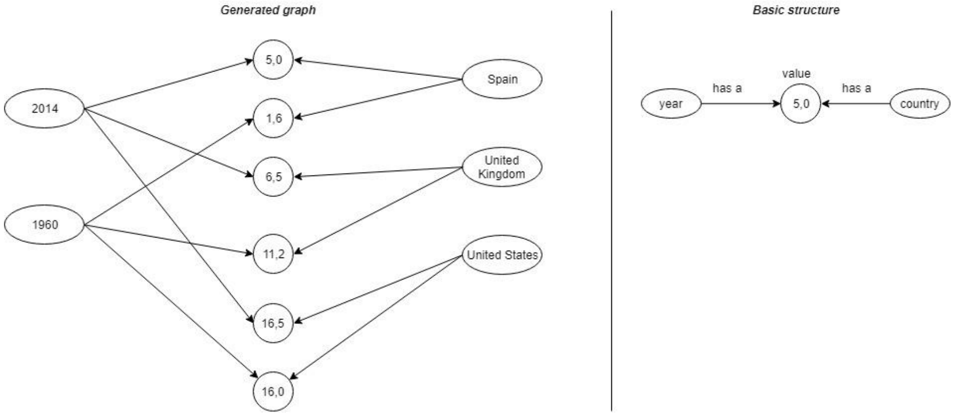
We need to ensure that the way in which the graph is designed complies as far as possible with the quality dimensions expected in the documents. We propose the following, in a table, explaining what type of information is suitable to be a node so that it can be compared with other data internal or external to the sources. There can be no other approach because the main function for which semantic data was created was its link with other data on the web. This method leads us to define an approach that consists of extracting one node per identifiable element. We define an identifiable element as the type of information that can be compared with other data internal or external to the source. Principally, it consists of searching by record the type of data that is a candidate to be an identifier, which does not necessarily have to appear literally in the original table but can be inferred from it. Below is an example of how each of the quality guidelines analyzed above behave in the design of the methodology (see Phase 4). It can be observed how the data are interconnected in such a way that the centrality of the nodes is similar, and how they reflect a strong grouping among them. It is a mechanism that ensures data interoperability by choosing a node that represents relevant domain information, and therefore the possibility that the data can be checked and continue to improve. For example, Table 3 shows the CO2 values per year for several countries:

Table 5. CO2 levels by country and year.

Country name	**1960**	**2014**
Spain	1,6	5,0
United Kingdom	11,2	6,5
United States	16,0	16,5

Cities and years are candidates to be nodes because they allow themselves to be compared with other data. Finally, it represents the description of the resources that in this case are the values of CO2. The relationships between each of these elements are imposed by the very domain of the data. Figure 5 shows the graph obtained after using this approach.

Fig. 5. Nodes per identifiable elements.



It is not the objective of our research to demonstrate which is the optimal mapping mechanism, but to propose a mapping mechanism that includes the correct use of quality guidelines so that it can be used in the transformation process and ensure the obtaining of accurate RDF documents.

6 Quality Methodology Scope

It is a methodology that can be used in a transformation process from structured data to semantic data. It is characterized by being adaptable to a transformation tool. It means that it does not affect the definition of a transformation tool as validations that adapt to the different types of data on which the tool is going to be used. This is possible if we take into account that the original documents are not developed following any standard. On the other hand, it must be considered that quality tasks are determined by the improvement needs of the selected data. The result of using this methodology to develop a tool for transforming structured data into LOD are mainly the following functionalities:

A. Validation methods to validate the original data.
B. To provide functionalities to simplify the processes of obtaining new information.
C. Provide functionalities to decide which will be the nodes and to establish the relations between them and the relation of attributes to complete the construction of the information graph.
D. To be able to specify semantic rules depending on the data domain and syntactic rules by which the triples of the documents must be governed.
E. It must include compliance with the quality guidelines defined in the quality methodology.

Conclusion and Next Steps

A quality methodology has been developed focused on the improvement of semantic data from before its transformation, in a structured format, until it is obtained. During the analysis of the problem, we observed the quality dimensions that needed to be improved so that the data generated reflected the complexity of the domain as close as possible to reality. This analysis allowed to determine the quality guidelines that the LOD must meet in a transformation process to LOD from structured data. Subsequently, it was possible to reveal which quality dimensions should be analyzed when reusing or not an existing ontology to cover all the possible challenges that exist in the creation of LOD and help in the determination of a mapping mechanism that takes into account these quality dimensions. This task indicated that, in any case, whether a data scheme is reused, a consistent quality treatment is necessary in this transformation process. The importance of orienting the design of a mapping mechanism towards the reuse of the data was observed because it contributes to the design of an interoperable, complete and linked semantic data model, that is, it contributes to the fulfillment of the quality dimensions that must be met. As future work, we consider it very important to gain more experience of the quality dimensions that need to be improved during the transformation process by analyzing other data sources.

This will allow us to improve our research. We strongly encourage the empirical validation of the quality methodology to demonstrate both the applicability and the practical benefit of the approach.

Bibliography

1. Linked Data, https://www.w3.org/DesignIssues/LinkedData (2006), last accessed 2019/01/17.
2. Loshin, D. Enterprise Knowledge Management. The Data Quality Approach California: Academic Press. p. 493 (2001).
3. Petrou, Irene, Marios Meimaris, and George Papastefanatos. "Towards a methodology for publishing linked open statistical data." JeDEM-eJournal of eDemocracy and Open Government 6.1 (2014): 97-105.
4. García, Ander, et al. "Methodology for the publication of linked open data from small and medium size DMOs." Information and Communication Technologies in Tourism 2015. Springer, Cham, 2015. 183-195.
5. Debattista, Jeremy, SÖren Auer, and Christoph Lange. "Luzzu—A methodology and framework for Linked Data quality assessment." *Journal of Data and Information Quality (JDIQ)* 8.1 (2016): 4.
6. C. Batini, C. Cappiello, C. Francalanci, and A. Maurino, Methodologies for data quality assessment and improvement, ACM Computing Surveys (CSUR), vol. 41, p. 16 (2009).
7. F. Maali, J. Erickson, and P. Archer. *Data catalog vocabulary (DCAT).* W3C Dataset Exchange Working Group (DXWG), https://github.com/w3c/dxwg, (2018), last accessed 2019/02/03.
8. Wilkinson, Mark D., et al. "The FAIR Guiding Principles for scientific data management and stewardship." Scientific data 3 (2016).
9. Sahoo, Satya S., et al. "A survey of current approaches for mapping of relational databases to RDF." W3C RDB2RDF Incubator Group Report 1 (2009): 113-130.
10. Ougouti, N. S., Belbachir, H., & Amghar, Y. A new owl2 based approach for relational database description. International Journal of Information Technology and Computer Science, 7(1), 48-53 (2015).
11. Rozeva, A. Approach for ontological modeling of database schema for the generation of semantic knowledge on the web. In AIP Conference Proceedings (Vol. 1690, No. 1, p. 060003). AIP Publishing. (2015)
12. Jiménez-Ruiz, E., Xiao, G., Soylu, A., Lanti, D., & Rezk, M. Ontology Based Data Access in Statoil. Web Semantics: Science, Services and Agents on the World Wide Web, 44, 3-36 (2017).
13. Kasrin, N., Qureshi, M., Steuer, S., & Nicklas, D. Semantic Data Management for Experimental Manufacturing Technologies. Datenbank-Spektrum, 18(1), 27-37 (2018).
14. Soylu, A., Kharlamov, E., Zheleznyakov, D., Jimenez-Ruiz, E., Giese, M., Skjæveland, M. G., ... & Horrocks, I. Optiquevqs: a visual query system over ontologies for industry. Semantic Web, 1-34 (2017).
15. Kharlamov, E., Mailis, T., Mehdi, G., Neuenstadt, C., Özçep, Ö., Roshchin, M., ... & Giese, M. Semantic access to streaming and static data at Siemens. Web Semantics: Science, Services and Agents on the World Wide Web, 44, 54-74 (2017).
16. Elbashir, M. K., Aboelhassan, M. A., & Ali, A. A. Mapping Relational Database to Resource Description Framework using Direct Mapping Method. Gezira Journal of Engineering and Applied Sciences, 11(2) (2018).
17. Mc Gurk, Silvio, Charlie Abela, and Jeremy Debattista. "Towards Ontology Quality Assessment." MEPDaW/LDQ@ ESWC (2017).
18. Zaveri, A., Rula, A., Maurino, A., Pietrobon, R., Lehmann, J., & Auer, S. Quality assessment for linked data: A survey. Semantic Web, 7(1), 63-93 (2016).
19. How to Publish Linked Data on the Web, http://wifo5-03.informatik.uni-mannheim.de/bizer/pub/LinkedDataTutorial/ (2011), last accessed 2019/01/17.
20. J. E. L. Gayo, Eric Prud'hommeaux, Iovka Boneva, Dimitris Kontokostas (2018) Validating RDF Data, Synthesis Lectures on the Semantic Web: Theory and Technology, Vol. 7, No. 1, 1-328.
21. Zaveri, Amrapali, and Gökhan Ertaylan. *Linked Data for Life Sciences*. Algorithms 10.4 (2017).