Improving the Linked Open Data creation from tabulated data: Quality Guide

Any Navarro Camero[1] UO272835@uniovi.es and

José Emilio Labra Gayo[1] labra@uniovi.es

[1] Department of Computer Science, University of Oviedo, Oviedo, Asturias, Spain

Abstract. The tabulated data has shown to be an important resource to improve the completeness of the existing knowledge graph (KG). The most current research has developed very advanced algorithms that manage to automate this task which consists mainly in inferring the types, relationships and facts of the table and then integrating all this information into the KG. However, they do not conscientiously analyze some quality problems, such as the incompleteness itself that presents these KG, which opens a gap to generate Linked Open Data (LOD) that inadequately reflect the domain of the data. This, in addition to intruding on the reuse of LOD, trigger quality testing tasks that are as cumbersome as complex is the domain. In this research we design a quality guide for the construction of accurate LOD when it comes to new tabulated data that cannot be integrated into a KG due to such incompleteness. It aims to anticipate possible errors in the generated data, report on the implications of its non-eradications and provide, based on the principles of quality that are expected to be fulfilled in these data, which are the tasks that must necessarily be executed to obtain accurate LOD.

Keywords: Linked Open Data, Tabulated Data, Knowledge Base, Knowledge Graph, Quality Guide.

1 Introduction

The most current methods for the LOD creation from tabulated data are characterized for being automatic processes. They consist of the definition of an algorithm that is capable of inferring entities, types and relations in the tabulated data and integrating them with bases existing knowledge, Freebase DBPedia, YAGO or Wikipedia. This inference or extraction of new knowledge of the table seek to complete these knowledge bases with new obtained information of the tabulated data; the most current are characterized for being very precise in finding the relations between the information and because they re-use trustworthy information (originated from the knowledge bases). Nevertheless, due to this uncompletedness, they cannot be effective for any set of tabulated data. On having been processes merely automatic unable to understand the data context, they do not detect certain quality problems, increasing the possibility of unleashing problems of in consistency in some points of the knowledge graph (KG). When the construction of this data is realized it is necessary that they are as precise as possible, it is not enough to integrate correctly the tabulated data with the knowledge bases. We consider this argument because they exist a group of quality dimensions that can turn affected in the obtained LOD if they are not analyzed earlier or during its creation. The fundamental target of this research work consists of developing a quality guide that helps to the conversion of tabulated data in accurate LOD, warns of the errors of saying LOD during its construction, at the time that it simplifies the whole process.

2 Related Work

The most predominant algorithms provide a solution for the construction and increase of knowledge graph (KG) making use of the information of the existing knowledge bases (KB), Freebase, DBepedia or YAGO. They consist of two basic targets (1) to predict the type (classes) of every column of a table and (2) later the column peer relationship. T2K Match [5], generative models' language [11], Mentor [12], TableMiner+ [15] and ColNet [7] constitute the most advanced algorithms in order of, lower to higher efficiency. They range from the use of an evidence locally and later globally [12], by the classification of the columns and the creation of relations between the columns using more complete KB [5,15], using generative models' language to represent the relations in one KG with relational phrases extracted from a corpus web scale [11] and using machine learning and a set of search results, candidates to identify the classes and they infer its entities to achieve a note of more precise type of column [7]. All these algorithms are characterized by using specific and incomplete KB, means that they cannot be used to transform any group of tabulated data. A recent study replicated two of these advanced methods [4], TableMiner and T2K Match, and founded that, since the table contains certain redundant information, the new relations that are created between the information of the table and the concepts of KG also are redundant and also slightly novel. In general, we think that the problem of these algorithms is that, on having been based on skills of automatic learning for its functioning, they can generate not ideal results when they work with novel information, of which previous record is not had in the bases of knowledge used for the inference of types and relations. In these cases, we believe that, at least until the KB reach a major completeness level, the manual achievement will be necessary on the part of users expert in the mastery of the information with which it is about to work.

Finally, the target of this work will be therefore the construction of a quality guide that simplifies to these users expert in the mastery of the information the construction of more accurate LOD at the time that we avoid the complex evaluations that often need from a rigorous study on, for example, who there are the people in charge of solving the errors and how to solve [9].

3 Quality Guide

The document conversion work will be divided into four phases. We defined a preliminary phase that aims to explain the preparatory tasks that are necessary before starting with the different phases of work in which the conversion will be done.

The definition of these phases of work includes in turn several sections:
Target: It is a description of what we seek to achieve in the phase and what will be the status of the data once completed.
Problems & Implications: It indicates possible problems that may arise in a phase and explain and justifies the consequences derived from the problems. In the problems it will be indicated a value of gravity, for, if this cannot be resolved, one has a guide of the implications that this will have in the quality of the LOD generated.
Execution Tasks: At this point the different tasks that will carry out the conversion are detailed. Each phase includes the tasks necessary to correct the problems described in that phase (a task can solve one or several problems). Each task sets a series of steps that, if completed correctly, would correct the associated problem. It may happen that if you fail to achieve some step, you will be prompted to re-perform some steps from a previous phase to try to continue the process correctly. It means that each phase clarifies when it comes or not, to continue the conversion process. If it happens that the characteristics of the domain of the data are imposed on the characteristics that are expected to be fulfilled in the obtained KG, it means that the data handler is forced to move to the next phase in breach of a task of the process, you will be able to verify what problem is unsolved and what is the severity of this. The problems are separated by commas, if not, they are grouped, for example: 1-5 (problems from 1 to 5).
The Figure 1 shows a schema that defines the working logic of the guide.

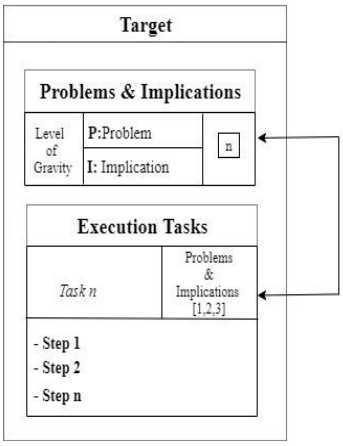

Fig. 1. Work logic of the quality guide: The Level of Gravity is defined as high, medium or low; *"n"* refers to the problem number. The heading Problems & Implications of the Execution Tasks section indicates which problems and implications (*n*) the task resolves.

Table 1. Quality Guide: target, problems and implications and execution tasks of the conversion phases of tabulated data to accurate LOD.

Preliminary Phase
Target: The objective of this phase is to inform about the resources we must have to start with the subsequent task related to the conversion process.

- Check whether the data source provides any validation mechanism for the data. For example, if it is an XML file, check whether the source provides an XSD or DTD file or a valid namespace. If it does not exist, create it taking into account the characteristics of the domain. - Located all information of interest in the documentation provided by the data source. It is important to have all the information of the domain to get more accurate LOD.		
Phase I: *Input data validation*		
Target: To make a correct validation of the structure and format of the input data to increase its reliability and ensure its correct interpretation during the process of conversion to LOD. This task ensures compliance with the Syntactic Validation & Semantic Accuracy dimensions.		
Problems & Implications		
High	**P:** The structure or document data is incorrect: incorrect or unplaced data in the corresponding site.	1
	I: Anomalous interpretations during the conversion process because current algorithms do not detect this type of problem.	
Medium	**P:** Non-correct data types for some fields.	2
	I: Impossibility of automated treatment in the future for those fields.	
Medium	**P:** The restrictions expected in some fields are not met.	3
	I: These values would be very possibly erroneous and would affect the reliability of the generated data.	
Medium	**P:** Unexpected values base on the properties of the parsed field.	4
	I: These values would be very possibly erroneous and would affect the reliability of the generated data.	
High	**P:** There are properties and values that contradict other data in your domain.	5
	I: Distorts the domain of the generated semantic data and can even create irrational information after linking properties and values.	
Execution Tasks		
Check document structure: - Verify that the number of columns and cells of all rows in the table matches the one expected. If any fails: Add the missing information or remove the surplus. Add null values if necessary. - Verify that there are all cells or headers or that there are no displaced cells or errors in the format of the data, for example, clarification at the time of putting real data (if '; ' or ', ' is used when separating the data), a mixture of data representation format, for example that have numeric value, also in text. - Find text information in tables that cannot be processed by computer and change it to a computer readable format. For example, the World Bank CO2 emissions are expressed in metric tonnes per capita[1], a text.		1,5

[1] https://datos.bancomundial.org/indicador/EN.ATM.CO2E.PC?locations=ES

<table>
<tr><td colspan="2">All these found errors must be corrected to continue with the next phase of the conversion. Since we cannot control all possible format, it is the responsibility of the data transformer to validate the structure of the file.</td><td></td></tr>
<tr><td colspan="2"><em>Detect outliers:</em>
To detect if a value is atypical the Tukey test will be used. For each field of information, we collect their values and calculate the first and third quartiles (Q1 and Q3 respectively). All values that do not meet the following formulas are outliers:
$Value >= Q1 - 3 * (Q3 - Q1)$
$Value <= Q3 + 3 * (Q3 - Q1)$
The atypical values found, try to correct them because they are possibly incorrect values and if it is not possible to leave them with null value.</td><td>4,5</td></tr>
<tr><td colspan="2"><em>Verify that there are not inaccurate values:</em>
- Use of functional dependencies between the values of two or more different properties. If functional dependencies are found in the fields in the data table, it must be verified that the values for each record for those fields meet them. If it is not fulfilled these values will be considered inaccurate.
- Verify that the type of each data cell is adjusted to the corresponding domain, both in the type and in the range of values that must be supported. For example, we cannot give valid a column in which it is shown as environmental humidity value 150%, since this can never be greater than 100%.
- Review that for each property and its corresponding values comply with the established rules of the type to which they belong. Some rules are, for example:
<strong>Identifier:</strong> The values of a property are of type identifier if they are responsible for representing the identity of each records. It should be verified that they are unique to each record and that they can obtain related data (they are used in the same format in other types related to this). Convert them to a format that if it meets the above if necessary.
- <strong>Combination of valid values:</strong> a property has this type if there is only one set of valid values possible. In this case it should be verified that each value is one of the possible.
The rules for the rest of the source data types must be identifies and defined in the same way. To rely on the opinion of experts or on other sources of similar data that are known are correct. If there are inaccurate values, correct them; if this is not possible, replace them with null values.</td><td>2-5</td></tr>
<tr><td colspan="3"><strong><em>Phase II:</em></strong> <em>Extend the input data</em></td></tr>
<tr><td colspan="3"><strong>Target:</strong> To include all the necessary information according to the domain to which the data belong, to ensure that it will be possible later the correct inference of information from this data by the consumers LOD generated. Task to accomplish this objective improve the dimensions of Completeness, Interoperability, Trustworthiness and Consistency.</td></tr>
<tr><td colspan="3"><strong>Problems & Implications</strong></td></tr>
<tr><td></td><td><strong>P:</strong> There is non-processable information by computer.</td><td></td></tr>
</table>

Medium	**I:** It will cease to include information in the KG and thus precludes the creation of relationships. It decreases the reliability and accuracy of the data generated and that can be linked to other data.	6
High	**P:** Schema incompleteness	7
	I: Incompleteness of the domain in the KG generated precluding the creation of new types and consequently entities and relationships. It decreases the reliability, accuracy of LOD and its linkage with other data.	
Medium	**P:** Incompleteness of the average population.	8
	I: It affects the completeness, reliability, precision and inference of relations in the LOD.	
Medium	**P:** Incompleteness of the instances	9
	I: It affects the completeness reliability, precision in LOD. Increases the risk of creating few clustered nodes.	
High	**P:** Types, attributes and values are not defined with a domain-relevant vocabulary.	10
	I: The LOD obtained may not be linked to other data in your domain.	
Medium	**P:** Lack of evidence in the facts of the average data.	11
	I: data values may be outdated or erroneous and generate inaccurate or incorrect information.	
Execution Tasks		
Identify non-processable fields by computer: - Include information offered through external documents, but that is relevant in the domain of the data. This information must be added through new fields in the data tables, so that it is automatically processable. - Verify and include in the data the clear use of a language, symbols, units, data types and definitions, in such a way that they are automatically processable. - Verify that the theme of the data source has been included in the data.		6-9
Verify the completeness of the schema: Verify that all the concepts and entities with their attributes of the corresponding domain have been defined searching and adding those that are not found, if possible. To rely on the data sources of reference for the domain of the information or by the opinion of experts. If all the required data are not found after analyzing the reference sources or consulting experts, and still continues with the conversion, it should be noted that the LOD will have completeness problems.		7
Verify the completeness of the population: Verify that there is all the information that is expected to be obtained for the correct use of the data. To rely on the data sources of reference for the domain of the information or by the opinion of experts. If all the required data are not found after analyzing the references sources of consulting experts and still continues with the conversion, it shows be noted that the LOD will have completeness problems.		8

	Verify the completeness of the instances: Analyze de extent to which the data is adjusted to the defined schema. It is to find the null values by rows. Null values are considered to an empty cell or when the value of a cell those not match the type expected by the column. For each null value it must be verified that there is no specific value that has been lost and can be recovered, and if so, it will be added. If you continue with null values in the instances, the LOD generated will have problems of completeness.	9
	Verify the use of a relevant vocabulary: Analyze whether the use of classes and properties conform to some formal definition. To accomplish this task, it is important to take as reference, world recognized data sources. If it is not possible because it is a rarely analyzed field, use the terms defined by the experts. Rename those classes and properties that do not conform to the correct uncovered terms.	10
	Verify the provenance of the data: Identify the author of each data group and determine its reliability based on reputation or references that can be verified. The data that cannot be verified will be considered doubtful and therefore must be eliminated from generated data source, even if this can generate problems of completeness.	11
	Phase III: *KG Construction*	
	Target: To ensure that the data uses a correct semantic data model in which the data is correctly interconnected and meets all the characteristics that guarantee its linkage and reuse with other similar data. If the previous phases have been successfully exceeded to this point, the information is considered accurate, reliable and usable. Tasks to meet this objective improve the dimensions of Interlinking, Interoperability and Syntactic Validation.	
	Problems & Implications	
High	**P:** Incorrect definition of classes, attributes or relationships.	12
	I: Creation of the KG with vagueness of the domain being and impediment to its connection with other data in addition, it generates problems of dereferencing and lost of information in the graph.	
Medium	**P:** Blank nodes	13
	I: Decreases the interoperability of the KG. In addition, the current RDF semantics of the blank nodes do not align well with SPARQL, interpret them as names blank nodes [1].	
High	**P:** Poorly defined or unstable KG nodes.	14
	I: Decreases the interoperability of the KG. It generates dereference problems and interconnection with data sources in different locations.	
	Execution Tasks	
	Define the semantic data model:	

The definition of the semantic data model can be made using existing technologies, such as RDF schema or OWL. Considered the following general properties of the ontology: • It must include in the own way for the ontology in question specific metadata, such as version, license and dates, and to refer to other ontologies. • It must include its corresponding URI to avoid using its location as a namespace, as this would make the namespace change when you change its location. • References to the terms of another ontology must be actually defined in the namespace of that ontology. - Indicate which elements will have identity and belong to complex classes (they will not be simple types) and which elements will be simple values: It is advisable to use a previously defined ontology; this will indicate which classes and attributes to use. W3C offers a standardized ontology that can be used like basic element for tabulated data (RDF Dates Cube Vocabulary). If this is the case, it is convenient: o Check the accuracy of the ontology. It is defined as the cardinality of the intersection between classes in the ontology and the classes of the data source, divided by the total number of classes in the ontology. This calculation is done on the Luzzu [6] reference framework. - If on the contrary we need to create a source of semantic data from scratch or extend a defined source with new classes and attributes, we must ensure these requirements: o For the choice of classes, prioritize classes in terms of using a value traying to facilitate their integration with other existing sources. It is advisable to comply with [14]: o Each concept must belong to a single class. o Avoid the creation of superclasses (sheet class of an ontology) if you are not going to associate any instance with it. o Classes cannot be created to capture types without any distinction between them because it lacks precision. o Declare the minimum number of classes possible in which you will only contain a subclass and have no siblings. o Classes cannot be referred by the same identifier. For example, ‘man’ can refer to different but related concepts, such as referring to ‘the human species’ or ‘male person’. o Subclasses of a class that are separated from each other (a subclass can only be of one type), you must specify this separation in the ontology [3]. o All classes must be semantically different and used. - As part of the definition of the attributes we will have to take into account the characteristics of the domain perform the following tasks also trying to facilitate their integration with other existing sources [14]:	12-14

  - Definition of the attributes indicating their name and function. Definition of the restrictions of these attributes.
  - Definition of the restrictions of these attributes.
  - Definition of the cardinalities of these attributes.
  - They should always be accompanied by their dominance and rank.
  - As the classes cannot be referred by the same identifier.
  - Attributes must not contain design information. For example, if it is a button or an image because it is completely irrelevant to existing KB such as DBPedia.
  - It must be related to the rest of the ontology: declared and used.
- Find all possible relationships between classes. Be supported by the related documentation or expert opinion.
- Classes and their relationships and attributes should always use the relevant vocabulary of the domain and make appropriate use of language, symbols and terms.
- To accurately represent the domain, the correct use of the relationship type is required [13]:
  For example, the '*rdfs: subClassOf*' relationship is reserved for the subclass relation, 'rdf: type' for objects that belong to a particular class, and '*owl: sameAs*' is used to indicate that two instances are equivalent [8].
- Once the creation of the ontology has been completed, there is the problem of knowing if the clustering coefficient and centrality achieved is relay the adaptation. To check if so, the following formula should be used:
  $Clusterin\,g_{by_{node}} = \frac{|\{e_{jk}\}|}{\frac{k_i(k_i-1)}{2}}$ $:v_j, v_k, \in N_i,\ e_{jk} \in E$ where, $|\{e_{jk}\}|$ it counts the number of edges between the neighbors of the node and k_i is the coefficient of the node.

  - If a grade of equal grouping has not been reached to at least 90 % of the number of nodes of the graph, it is necessary to return to the previous phase and to try to extend the information of the source with the target to define more relations between the nodes of the graph.
- To obtain a centrality high degree it is necessary that all the nodes have a value as nearby as possible of the centrality:
  $Centrality = \sum_{j,k} \frac{b_{jik}}{b_{jk}}$ where b_{jk} is a number of shorter ways from the node j up to the node k, and b_{jik} is the number of shorter ways from j up to k pass across the node i. The summation will cover all the present nodes in the graph.
  - If the biggest reached centrality grade overcomes 110 % of the minor, the centrality is low, therefore, debit to return to the previous phase.

To detect the nodes of the graph:

- In every node definition to use a universal identifier. The experience of other investigations advises the use of coherent local URI [1]. This will guarantee that the LOD are better subject to the interconnection and recycling. To rename all the identifiers that local URIs does not use in its

<table>
<tr><td colspan="2">definition. Not being possible he advises himself to use an alias and these must be dereferenced. A way of obtaining this would be to establish linkage to alias of well-known URI. You can rely on the existing KB.
- The identifiers must be defined according to that they can never change. A node which identifier changes over the course of time will lose its identity.</td><td>14</td></tr>
<tr><td colspan="2"><i>To generate the triples:</i>
- To gather from the original data source the values of the different attributes of every node:
o There will have to be the indicated ones by the semantic data model and will have to exist in the original data source.
- If the data model and the type and form of the identifiers decided correctly, it is not possible to incur errors by how the information refilled one is realized. The only task that owes to guarantee to do correctly is the creation of quite definite triples. Therefore:
o It is necessary to make sure that the way of representing the resources uses the types (classes) opportune and its corresponding values expire with the lexical syntax of the assigned type.</td><td>12</td></tr>
<tr><td colspan="3"><b><i>Phase IV:</i></b> <i>KG Validation</i></td></tr>
<tr><td colspan="3"><b>Target:</b> To verify that generated KG expires with the structure defined in the semantic model of information previously and represents the information of the domain. This is fundamental to guarantee the correction so much of the process of generation as of the information previously offered by the original data source. The tasks to expire with this target improve the dimensions of Syntactic Validity, Semantic Accuracy and Consistency.</td></tr>
<tr><td colspan="3"><b>Problems & Implications</b></td></tr>
<tr><td rowspan="2">High</td><td><b>P:</b> Syntactic errors in the document RDF: the document RDF was not generated correctly.</td><td rowspan="2">15</td></tr>
<tr><td><b>I:</b> It would disable the use of the generated document.</td></tr>
<tr><td rowspan="2">Medium</td><td><b>P:</b> The value of the object of a triple is wrong.</td><td rowspan="2">16</td></tr>
<tr><td><b>I:</b> Erroneous information is published.</td></tr>
<tr><td rowspan="2">Medium</td><td><b>P:</b> Type of incorrect fact.</td><td rowspan="2">17</td></tr>
<tr><td><b>I:</b> Unexpected results on having consulted the information: unreal information.</td></tr>
<tr><td rowspan="2">High</td><td><b>P:</b> The facts in the information do not represent the domain correctly.</td><td rowspan="2">18</td></tr>
<tr><td><b>I:</b> Unexpected results on having consulted the information: unreal information.</td></tr>
<tr><td colspan="3"><b>Execution Tasks</b></td></tr>
<tr><td colspan="2"><i>Syntactic and semantic ratification of the information:</i>
These tasks will be able to be determined automatically from the ontology (model) in question.</td><td>15-18</td></tr>
</table>

The use of Shape Expressions [8] can be of big help for the achievement of this task. Several demonstrations exist in line to understand the ratification engine: Shex[2]; RDFShape[3]; ShexValidata[4]. The model of the information can be used to shape the ratification expressions since it contains the types, the relations of cardinality and values that we want to verify in the obtained information (LOD). He must construct the shape expressions from the semantic data model to realize the following ratifications: - To define and to verify proper syntactic rules of the format RDF. - To verify that the resources of the graph belong to the class that corresponds to him. - To verify that the values of the attributes are lexically well represented. - To verify that the types of data are the awaited ones. - To verify that every node contains the required attributes. - To 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 the defined in its model. - To verify that all the rules that represent relations between the information defined are fulfilled satisfactorily: o The proper rules of the problem for the search incompatibilities. For example, to declare a rule in which it verifies that the age of a person is major than 16 years if the marital status is married. o Status of possible values for every attribute.	
To use the information inference to find inconsistency in the information: - To determine the implicit relations between the nodes. For it nodes can be related inside the LOD between themselves or make use of other external KB making use of the use of identifiers dereferenceable. - For every relation between nodes determined in the previous step to verify, being helped with hardware of consultation like SPARQL or similar, that do not generate inconsistencies. For example, to verify that the average of CO2 in the certain year for Spain is correct, comparing it with the average calculated with the values of CO2 of every month.	18

4 Conclusions

This document describes a quality guide to convert tabulated data to accurate LOD, a tool that, in contrast to other quality improvement tools, shows how different quality tasks can be applied in the main phases of the conversion process and as an alternative that can be used to convert any group of tabulated data to accurate LOD; fundamentally how to avoid errors in the generated LOD and how to proceed when they exist. There have been used like starting point the quality principles that can be improved in every phase. Then, we identify the possible errors that prevent its fulfillment and the

[2] http://rawgit.com/shecSpec/shex.js/master/doc/shex-simple.html

[3] http:/shaclex.herokuapp.com

[4] https://www.we.org/2015/03/ShExValidata/

implications that it bears to avoid them being alert, also, of the gravity of the published LOD containing these errors. All these actions shape a logical sequence of steps that allow to guide the data handler and offer a knowledge of the state of the quality in which the information is in real-time. As future work we consider to be very interesting the integration of some of the automatic tool of generation of more ideal LOD inside the flow of work explained in this guide, so that without losing quality in the generated LOD a major efficiency should be obtained in the conversion process.

References

1. Aidan, H., Jürgen, U., Andreas, H., Richard, C., Axel, P., Stefan, D.: An empirical survey of linked data conformance. Web Semantics: Science, Services and Agents on the World Wide Web, 14, 14-44 (2014).
2. Amrapali, Z., Anisa, R., Andrea, M., Ricardo, P., Jens, L., Sören, A.: Quality assessment for linked data: A survey. Semantic Web, 7(1), 63-93 (2016).
3. Asuncion, G., Mariano, F., Oscar, C.: Ontological Engineering: with examples from the areas of Knowledge Management, e-Commerce and the Semantic Web. Springer Science & Business Media, (2006).
4. Benno, K., Peter, B., Jacopo, U.: Extracting New Knowledge from Web Tables: Novelty or Confidence?, pp. 1-5. (2018). ResearchGate, (2018).
5. Dominique, R., Oliver, L., Christian, B.: Matching HTML tables to DBpedia. In Proceedings of the 5th International Conference on Web Intelligence, Mining and Semantics, pp. 10, (2015).
6. Jeremy, D., Christoph, A., Santiago, L., Sören, A.: Luzzu-A Framework for Linked Data Quality Assessment. In 2016 IEEE Tenth International Conference on Semantic Computing (ICSC), pp. 124-131. IEEE, (2016).
7. Jiaoyan, C., Ernesto, J., Ian, H., Charles, Sutton.: ColNet: Embedding the Semantics of Web Tables for Column Type Prediction. arXiv preprint arXiv:1811.01304, (2018).
8. Jose E., L., Eric, P., Iovka, B., Dimitris, K.: Validating RDF Data, Synthesis Lectures on the Semantic Web: Theory and Technology, Vol. 7, No. 1, 1-328, DOI: 10.2200/S00786ED1V01Y201707WBE016, (2018).
9. María, P., Meri C., Z., Asuncion, G.: OOPS! (OntOlogy Pit-fall Scanner!): International Journal on Semantic Web and Information Systems. 10, 7-34 (2014).
10. Maribel, A., Amrapali, Z., A., Elena, S., Dimitris, K., Fabian, F., Jens, L.: Detecting linked data quality issues via crowdsourcing: A dbpedia study. Semantic Web, 9(3), 303-335 (2018).
11. Matteo, C., Denilson, B., Paolo, M.: Towards Annotating Relational Data on the Web with Language Models. In Proceedings of the 2018 World Wide Web Conference on World Wide Web, pp. 1307-1316. International World Wide Web Conferences Steering Committee. ACM (2018).
12. Matteo, C.: Leveraging Wikipedia Table Schemas for Knowledge Graph Augmentation. Proceedings of the 21st International Workshop on the Web and Databases. ACM, (2018).
13. Natalya F., N., Deborah L., M.: Ontology Development 101: A Guide to Creating Your First Ontology. Stanford Knowledge Systems Laboratory. 25 (2001).
14. Silvio, M., Charlie, A., Jeremy, D.: Towards Ontology Quality Assessment. In MEPDaW/LDQ@ ESWC, 94-106 (2017).
15. Zhang, Z.: Effective and efficient semantic table interpretation using TableMiner+. Semantic Web 8(6), 921-957 (2017).