redin Revista Facultad de Ingeniería, No.86, pp. 40-53, 2018

A fuzzy logic system to evaluate levels of trust on linked open data resources

Un sistema de lógica difusa para evaluar niveles de confianza en recursos de datos abiertos vinculados

Paulo Alonso Gaona-García[1], **Jhon Francined Herrera-Cubides**[1*], **Jorge Iván Alonso-Echeverri**[1], **Kevin Alexandre Riaño-Vargas**[1], **Adriana Carolina Gómez-Acosta**[2]

[1]Grupo de Investigación GIIRA, Facultad de Ingeniería, Universidad Distrital Francisco José de Caldas. Carrera 7 No. 40B - 53 Bogotá D.C. C. P. 110231. Bogotá D.C., Colombia.
[2]Fundación San Mateo. Transversal 17 Nº 25 - 25 Bogotá. C. P. 110231. Bogotá D.C., Colombia.

ARTICLE INFO:
Received September 21, 2017
Accepted January 12, 2018

KEYWORDS:
Linked Open Data (LOD), Levels of Trust, Open Data Consumption, Fuzzy Logic, Fuzzy Systems Type I.

Datos enlazados vinculados (LOD), Niveles de confianza, Consumo de datos abiertos, Lógica Difusa, Sistemas Difusos tipo I.

ABSTRACT: Linked Data is a way of using the network by creating links among data from different information sources in order to improve search processes, semantic interoperability among other functionalities. One of the strategies used to perform linked resources queries is through the consumption of SPARQL Endpoints. However, to determine existent trust levels within and outside the knowledge domains where such resources are, the operation of these Endpoints is one of the critical factors for both to realize the state of these resources and their subsequent consumption. To recognize these states, the present article aims at exposing the description, modeling, implementation and analysis of a type-I fuzzy system based on logical rules. It also addresses decision-making regarding the uncertainty that presents the definition of levels of trust to determine the consumption obtained over a set of LOD Datasets Located in several Endpoints. Finally, it presents results, conclusions and further work from a case study performed through the postulation of parameters obtained at runtime over several Endpoints.

RESUMEN: Linked Data es una forma de utilizar la red creando links entre datos de diferentes fuentes de información con el propósito de enriquecer los procesos de búsqueda, la interoperabilidad semántica, entre otras funcionalidades. Una de las estrategias utilizadas para llevar a cabo consultas de recursos vinculados es a través del consumo de Endpoints SPARQL. No obstante, el funcionamiento de estos Endpoints es uno de los factores críticos para conocer el estado de estos recursos y su posterior consumo para determinar niveles de confianza que presentan de estos recursos dentro y fuera de los dominios de conocimiento desde donde se encuentran alojados. Para conocer estos estados, el presente artículo tiene como propósito plantear la descripción, modelamiento, implementación y análisis de un sistema difuso de tipo-I basado en reglas lógicas, está asimismo orientado a la toma de decisiones sobre la incertidumbre que presenta la definición de niveles de confianza para determinar el consumo que se tiene sobre un conjunto de Datasets LOD alojados en diversos Endpoints. Finalmente, presenta resultados, conclusiones y trabajo futuro a partir de un caso de estudio realizado a través de la postulación de parámetros obtenidos en tiempo de ejecución sobre varios Endpoints.

1. Introduction

The evolution of the Semantic Web [1] has been experiencing a growing production of content on the Web, a scenario where a user can generate and publish resources that have or not a certain degree of validity, reliability and usefulness. Likewise, they can assume the role of consumer of content published by another person called the "prosumer" role [2]. In order to carry out a successful process of linked data, a series of rules on the publication of web-based data has been proposed [1]. These rules specify that the data has to be available in RDF formats, and that to be queried prosumers need to use a standard SPARQL query language on a SPARQL Endpoint.

Despite the existence of recommendations for the publication of data, there have been significant challenges

* Corresponding author: Jhon Francined Herrera Cubides
E-mail: jfherrerac@udistrital.edu.co
ISSN 0120-6230
e-ISSN 2422-2844

DOI: 10.17533/udea.redin.n86a06

[3] in linking processes, both at the metadata level and at the data level itself. From the study of these challenges, one of the motivations that started this research is oriented to answer the following questions: How do you add confidence to the published resources? How do you measure those rules of data publication in the Web from mechanisms that allow you to evaluate their effectiveness? How do you determine trust parameters for evaluating the veracity of the sources linked under principles of Linked Data? A scenario where the application of these dimensions is important, for example, it is in the education domain, where there are many available educational resources published on the Web; as a result, there is an increasing need of linking and discovering them. Moreover, due to this huge amount of educational resources, the need of establishing levels of confidence, trust and quality on such contents is evident.

For this reason, one of the relevant aspects is to determine the concept of quality, understood as the *"fitness for use"* based on a proposal made by [4]. Therefore, the validation of the model proposed in this research will allow, among other things, to know the real state of the linked data in the Web. In addition, to determine aspects that allow evaluating and defining if a Dataset (or data within a Dataset) can be discovered automatically, and the possibility of creating links between the data resources (Datasets and data elements).

There are some proposals aimed at adding confidence to these resources with very variable information quality. In fact, different techniques have been used, such as:

- PageRank [5], a method for objectively and mechanically rating webpages, and thus effectively measuring the human interest and attention devoted to them.
- Content-based trust [6], a trust judgement on a particular piece of information or some specific content provided by an entity in a given context.

Many of these techniques focus on the user's contribution, trust links, content of their metadata, etc. However, there is a concern about how to assess the levels of confidence offered by current resources, specifically about how to assess the levels of trust offered by the current SPARQL endpoints that are available in the repositories where several of them present problems as:

- Broken links [7] that do not allow reaching other linked resources.
- The resources have been relocated from the server, but their references are not updated.
- Whenever an URI returns an error, it is because non-reference issues arise.
- The server or SPARQL Endpoint is delayed or cannot respond to a SPARQL query.

This is how the SPARQL Endpoints that currently exist should be able to execute any SPARQL query. However, some Endpoints are not able to resolve queries whose execution time or response cardinality exceeds a certain value, while others simply stop execution without producing any response. In addition, some limitations have been found by Vidal *et al.* [8] where the data accessible on the Web are usually characterized by: i) absence of statistics, ii) unpredictable conditions when executing remote queries, and iii) changing characteristics.

Due to these situations, this research seeks to address the SPARQL Endpoints consumption to evaluate the trust they offer, and it is especially focused on two characteristics of Accessibility: Availability and Response Time. Consequently, this research is based on fuzzy logic techniques given that this technique:

- Plays an important role so that the margin of the classification is more precise.
- Allows managing the level of uncertainty associated with the aggregate information of LOD.
- Allows optimizing the process of consumption, recovery and evaluation of the trust offered by the consulted Endpoints on these two measurement criteria.

In order to carry it out, this article presents six sections. In section 2, a presentation of the background of the main references that define the research area is revealed. Section 3 presents the proposed framework for the evaluation of trust levels and their articulation with Fuzzy Logic. Section 4 presents the results obtained. Section 5 discusses the results. Finally, section 6 presents conclusions and further work.

2. Background

The following section is a description of the main concepts used in this research regarding levels of trust in the consumption of SPARQL Endpoints.

2.1 Linked open data, data and metadata

Linked Open Data refers to a set of the best practices for the publication and interconnection of structured data on the Web in order to create a global interconnected data space called Web of Data [9].

This process of publication and interconnection of data requires an ensemble of information expressed in

data models. As described by Ahmad *et al.* [10], data repositories expose information across multiple models. A model must contain:

- The minimum amount of information transmitted to the consumer, the nature and content of their resources.
- Information enabling the discovery, exploration and exploitation of data.

This information is classified into the following types:

- General information: General information about the data set (e.g. title, description, ID). The data set owner manually fills out this information. In addition to that, labels and domain information are required to classify and improve Dataset detection.
- Access to information: Information on access and use of the Dataset. This includes the Dataset URL, some license information (e.g. license title and URL), and information about Dataset Resources. Each resource generally has a set of attached metadata (for example, resource name, URL, format, size).
- Property information: Information about the Dataset property (e.g. details of the organization, details of who provides support, of the author). The existence of this information is important to identify the authority in which the generated report will be sent and the newly corrected profile.
- Source information: Temporary and historical information about the Dataset and its resources (for example, creation and update dates, version information, version number). Most of this information can be automatically filled out and tracked.

Data repositories are Dataset access points (endpoints) that provide tools to facilitate publication, exchange, search, and display of data. CKAN is the world's leading platform for managing open source data repositories, which makes websites more powerful such as the Datahub.io that hosts the LOD cloud metadata.

However, what is a Dataset? As described in [9], a Dataset is any web resource that provides structured data:

- Tabular data: data tables that can normally be downloaded as CSV files or spreadsheets;
- Object Collections: XML documents or JSON files;
- Linked data: the standard web to describe the data. You can find linked data in RDF resources or in SPARQL Endpoints.

As Datasets are web resources, they must have a URL. Generally, Datasets URLs are discovered in data repositories or other data sources published by a data provider. Often, data sources publish links to Datasets along with their metadata, which is structured data about the Dataset itself (for example, the Dataset author, the last Dataset update time, etc.).

However, as the goal of this research is oriented to the data published, and beyond the metadata itself, it is necessary to take into account that the principles of Linked Data suggest the fulfillment of a series of levels where their last instances are oriented to use standard open formats such as RDF and SPARQL, in addition to link their data with the data of other people, thus forming graphs of knowledge [11].

Most providers that reach these levels provide a SPARQL Endpoint that in general terms is a SPARQL protocol service as defined in the SPROT (The SPARQL Protocol for RDF) specification. A SPARQL Endpoint allows users (human or otherwise) to query a knowledge base through the SPARQL language. Usually, results return in one or more machine-processable formats. Therefore, an Endpoint is mainly a machine-friendly interface to a knowledge base [12].

SPARQL Endpoints consumption can generate different types of perceptions in the users, given the existing trust level in that Endpoint. If the respective links are not stored in these Datasets [8]:

- The query responses undertaken may be incomplete, and important data may not have been included in the response.
- If it is required to evaluate the relationships between two Datasets in order to determine possible new associations, these queries could be affected by reducing the level of trust in that Endpoint.

Therefore, the work of evaluating and verifying the information retrieved from a query can lead to two scenarios: the expected one, contributing to obtain the highest level of trust, and the unexpected one: inconsistent or null, decreasing the level of trust that the user can perceive on the Endpoint that is consuming. EuropeanaLabs [13] describes an example of access and consumption of SPARQL Endpoint and Kabutoya *et al.* [14] provides an example of how to manage trust levels.

2.2 Trust and data quality

As described by Gil *et al.* [15], many authors have studied the concept of trust in various areas of computing and in the context of the Web and the Semantic Web.

Previous work on trust has focused on aspects such as reputation and authentication. Trust is an important issue in distributed systems and security. Many authentication mechanisms have been developed to verify if an entity is which it claims to be, usually using public and private keys. Many access control mechanisms, which are based on policies and rules, have been developed to be certain that an entity can access specific resources (information, hosts, etc.) or perform specific operations. Other authors have also studied the detection of malicious or untrusted entities in a network, traditionally in security, and more recently, in P2P networks and e-commerce transactions.

In the context of Linked Data, Endpoints handle links to other resources. The Accessibility dimension (through its metrics of availability and response times), allows evaluating the trust offered by such Endpoint. These metrics allow us to obtain arguments to answer questions such as: What SPARQL Endpoints can be trusted when you need to consult them? [16] Question that guides the research process presented in this paper. To address this question, the following aspects were raised:

a. Data quality is a multidimensional construction with a popular definition as the *"fitness for use"*. Data quality may depend on several factors such as accuracy, timeliness, completeness, relevance, objectivity, credibility, comprehension, consistency, conciseness, availability and verifiability. In terms of the Semantic Web, there are several concepts of data quality. Semantic metadata, for example, is an important concept that must be taken into account when evaluating the datasets quality. But, on the other hand, the notion of link quality is another important aspect presented in Linked Data in which it is automatically detected if a link is useful or not [4].

b. As stated by Zaveri *et al.* [4], within the accessibility dimension, which involves aspects related to the way data can be accessed and retrieved, the following metrics can be found:

    - Availability. The availability of a Dataset is the extent to which the information is present, obtainable and ready to be used.
    - Response time. It measures the delay, usually in seconds, between the display of a query by the user and the receipt of the complete response from the data set.

c. Since each of the SPARQL Endpoints that are available have their own navigation form and data exploration, it was considered important to use a graphical notation for the development of this process.

The LD-VOWL (Linked Data - Visual Notation for OWL Ontologies) was used to perform the graphical representation based on the extraction of information from the Endpoint schema of linked data and its corresponding display of the information using a specific notation [17]. In [18–20] different studies are presented on the characteristics of VOWL and its comparison with other tools. In brief, there are different standpoints on VOWL:

- The VOWL notation allows a more compact representation of ontology.
- Coloring the concepts, property types, and instances have a positive impact on task solution. It helped to identify elements and find information in the graphical representation.
- In Linked Data, the LD-VOWL extracts schema information from Linked Data endpoints and displays the extracted information using the VOWL notation.
- SPARQL queries are used to infer the schema information from the endpoint data, which is then gradually added to an interactive VOWL graph visualization.

The VOWL notation apparently provides only one possible way to visualize OWL ontologies using node-link diagrams. As OWL is not an inherently visual language, other types of visualizations would also be possible and could be more appropriate in certain cases. For instance, if users are mainly interested in the class hierarchy contained in an OWL ontology, they might prefer a visualization that uses an intended tree or a treemap to depict the ontology.

Some of the purposes for which this LD-VOWL proposal was used are:

- To unify queries and simplify the process of accessing different Endpoints from the same place in addition to being able to include the variables under investigation.
- To work with a tool that manages a visual language implementation to allow the expression of the different relationships and links between the entities involved in the queries of each user.
- This tool offers a friendly and understandable navigation experience for the majority of Internet users without knowledge in this field of work.

Moreover, authors such as Weise *et al.* [21], Lohmann *et al.* [22] and Sánchez [19], define this tool with a high level of usability in aspects such as:

- Other tools can be readily integrated with LD-VOWL as it runs completely on the client's side and it only requests the server through SPARQL, for example, WebVOWL [22] and Protégé [22, 23].

- It is a well-specified visual language for the representation of user-oriented ontologies. This defines the graphical representations for most elements of the Ontology Web Language (OWL) combined with a force-directed graphic design that visualizes the ontology.

- It manages interaction techniques that allow the exploration of the ontology and customization of the visualization.

Bearing this in mind along with the work undertaken with the proposed Visual Data Web notation, and in compliance with the specifications of the VOWL project, the consumption of the classified and selected Endpoints was carried out according to the vocabularies allowed by VOWL, which include RDF, RDFS, OWL and SKOS. Likewise, they were selected according to the strategies that each Endpoint uses to categorize and manage the linked resources. Thus, facilitating the necessary inputs and factors of the conception of the diffuse logic model proposed for this research work, and supporting the construction and improvement of methods and mechanisms for both the collection and evaluation of linked data on the Internet.

The following sections present in detail the aspects that were taken into account for the use of this tool, the process developed for the inclusion of the analysis of variables, and the respective implementation of the Type-I Fuzzy System based on logical rules, which eventually allowed the evaluation of these variables.

2.3 Fuzzy logic and the trust dimension in linked data

Fuzzy logic is a methodology that provides a simple way to get a conclusion from vague, ambiguous, inaccurate, noisy, incomplete or with a high degree of imprecision, unlike conventional logic that works with well-defined and precise information. It also allows to be implemented in hardware and software, or a combination of both [24].

This type of logic is capable of handling several types of ambiguity. Where x is a fuzzy set and u is a related object, the statement "u is a member of A" is not always either exactly true or exactly false. It may be true only to some degree, the degree to which u is in fact a member of x. A crisp set is specified in such a way as to divide everything under discussion into two groups: members and non-members. A fuzzy set can be specified in a mathematical form by assigning to each individual in the universe of discourse a value giving its degree of membership in the fuzzy set [25].

In this context, the use of fuzzy logic will allow abstracting information about the trust offered by the Endpoints, which come to offer information with a degree of inaccuracy. In this domain, different investigations have been identified, such as those proposed by [26–28]. It is important to note that many of these proposals do not consider a more intuitive visualization tool for the deployment of information resulting from queries. On the other hand, its focus is on confidence in information, annotations, content and reputation. In short, in this research, fuzzy logic is proposed to optimize the process of consumption, recovery and evaluation of the trust deployed by Endpoints based on the use of Accessibility/Availability and Response Time dimensions in fuzzy logic rules.

3. Methodology

This research is established in a quasi-experimental type methodology, where this design, as an empirical method, allows performing analysis of properties resulting from the application of the technological process in order to obtain an analysis of the variables to work, that is to say, the accessibility and the response times, as dimensions that allow generating trust in the accessibility of the data on the part of users.

From this perspective, a quasi-experimental design was carried out to implement a strategy that provides information about aspects identified in the problem presented by Vidal *et al.* [8] and Vandenbussche *et al.* [16], information that will allow identifying the level of trust in accessibility that these queries can throw when consuming such Endpoint. In order to carry out this proposal, a structured methodological design is defined in phases which allows us to determine the processes leading to our research proposal (figure 1). In order to carry out the quasi-experimental methodology, the methodological design starts from a Preliminary Stage (inputs) where different exercises were carried out as: a) connection to LOD repositories, b) Endpoint Consumption, c) Theoretical review of the accessibility. All of the foregoing as a strategy to obtain information about the levels of trust offered by an Endpoint.

The following are considered as subsequent stages:

— Download, Configuration and Test of the Selected Visualization Tool (VOWL): This stage considers the wiring process with the VOWL tool to establish the working environment that will allow the interaction with the designed system and the results from the queries made.

— Identification and Design of the Strategy to Address the Trust Levels (Type-I Fuzzy Sets): This stage addresses the resolution of the uncertainty generated

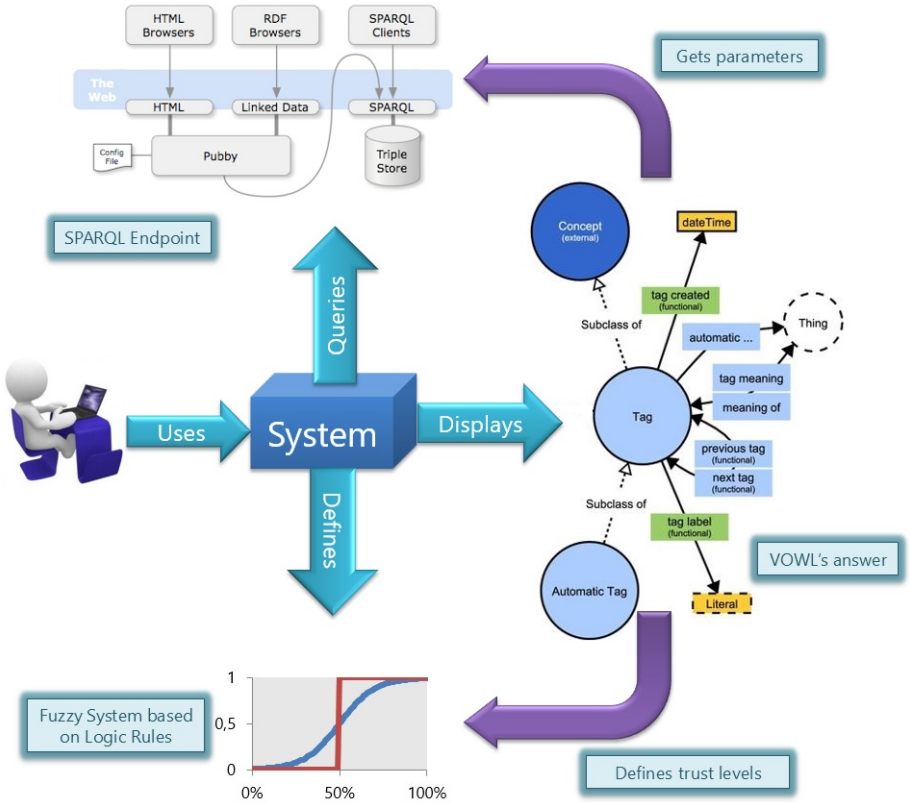


Figure 1 Proposed Methodological Design. Source: Authors

by the definition of levels of trust at the time of consuming linked data.

- Implementation and Testing of the Proposed Fuzzy Model: This stage describes the behavior of the current system through its interaction with elements from the VOWL tool and with the necessary mechanisms to introduce the fuzzy logic model, and defines the definition of concepts on levels of trust.

- Generation and Analysis of Results Obtained: This stage contemplates the process of examining the results of implementing the proposed system, and describes aspects of demonstrable levels of trust during execution.

4. Development of methodological design

4.1 Proposed framework

The proposed model (Figure 2) is composed of two fundamental structures. In the first instance the framework, which allows to obtain a display of the consulting process to an Endpoint from the wiring with the tool VOWL and the available arguments at runtime. In the second instance, it displays the creation of established rules for the declaration of levels of inference regarding data consumption dimensions, this will allow the system to make a series of decisions as a result of the process of treatment of the parameters defined under the design and implementation of the Type-I Fuzzy System. The following sections describe, briefly, the development of the proposed methodological design.

4.2 Visualization tool: VOWL

Based on the approaches defined by Lohmann *et al.* [29], OWL does not have a standardized visual notation like other related modeling languages. However, its visualization is very useful to be able to have a better understanding of its structure, and to be able to identify in a more direct way those key elements to analyze in the process of consuming Endpoints. In this process of searching tools for the visualization of OWL [30], VOWL was identified like a complete tool, which uses a well-specified, easy to understand and to implement notation.

Under these criteria, it is proposed to extend the VOWL functionalities [31] in order to analyze the suggested mechanisms of trust levels to validate the developed inference rules, as well as to determine a mechanism to obtain information about the times consumed in the

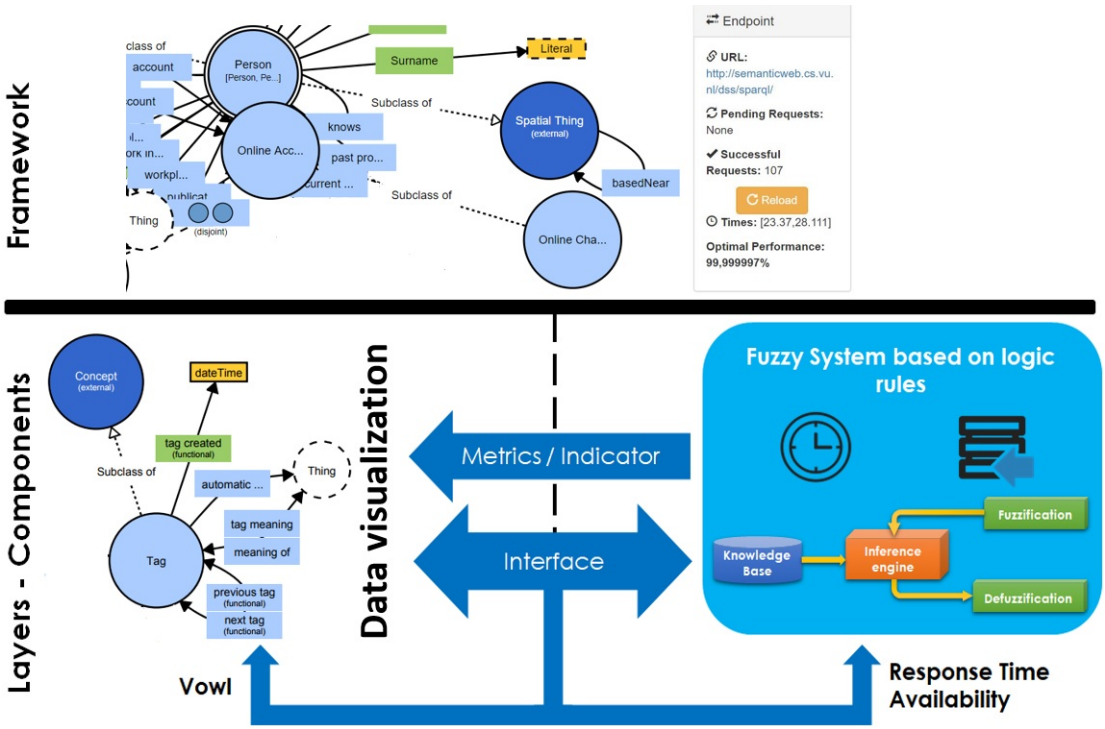


Figure 2 Proposed Framework

process of batch query of the resources exposed in the Endpoint. These aspects were considered key to know the state of the Endpoints. Therefore, from the VOWL visualization strategies, it is proposed to build a model based on fuzzy logic, as regards VOWL, that allows determining the degree of performance that the application reaches for each of the queries made in terms of satisfactory requests and the number of seconds it takes to obtain. The above with the aim of improving the current functionality provided by the tool and generating more elements of interest to Linked Data consumers.

From these exposed functionalities according to the levels of trust, the architecture proposed on VOWL presents two main factors:

- First, the response time in each query and/or consumption made by a Linked Data user, which is measured from the moment a request is executed until there is a response from the Endpoint whether it contains or not associated data, or until time runs out and no response has been obtained.

- Second, it represents the return rate adjusted to the overall performance of consumption taking advantage of the positive and negative results of the consumption. In this way, it was possible to establish the relationship with the dimensions and metrics that are presented and worked in Zaveri *et al.* [4]. In addition, the necessary inputs were provided for the proposed Fussy Logic Model that is explained in the later sections.

4.3 Proposed model of fuzzy logic

As part of the proposed design, it is established to work with Type-I Fuzzy Systems [32] with a design of logical rules based on two variables obtained from the query component to the Endpoint: the number of successful requests on the total requests in a query, and the average response time taken by each consultation round in resolving data consumption.

These variables are structured in the form of antecedents, which represent sets used to obtain a decision contemplated in the consequent sets. The choice of this approach is based on two premises: 1) in the environment of uncertainty generated when querying Datasets, and 2) the need to solve this uncertainty with control over the operation of the system. The above through obtaining arguments at runtime, and from the application to conceive the levels of trust associated with each phase of implementation.

For the implementation of the proposed model, the following procedures were defined:

a. Definition of variables. The variables to be considered for the problem presented correspond to:

- NSR (Number of Successful Request): The value of the proportion of requests that are successful from the total number of requests at the time of the query.
- RT (Response Time): The time taken in seconds for the application to respond to an associated query.

b. Declaration of fuzzy sets. The fuzzy sets that are associated with membership functions are established (Table 1), for the antecedents and the consequents of the system, through the definition of concepts that will adjust the ambiguity with which the system counts to make decisions about performance in execution time.

c. *Attributes for fuzzy sets.* The attributes used (Table 2) are defined for the definition of sigmoidal sets in terms of antecedents, and Gaussian in terms of the consequent ones, taking advantage of their characteristics of continuity and derivability by analyzing the most representative values in each scenario of ambiguity and the determination of the same ones for each function of belonging.

d. *Definition of fuzzy sets.* The fuzzy sets are defined (Figure **??** and **??**) in the dimensions of the antecedents and the consequent ones, by means of the conventions and attributes previously defined in order to be able to implement the system based on logical rules.

e. *Design of logical rules.* The logical system design is established (Figure e) taking into account the use of Mamdani's mathematical model [33] for the material implication of the logical conditional as given in Eq. e.

$$p \rightarrow q \equiv p \wedge q \tag{1}$$

Eq. e Mamdani's Implication.

From equation e, where p is the antecedent and q is the consequent, figure e presents the rules of the system.

f. *Design of the fuzzy system based on logical rules.* In terms of the antecedents and consequents proposed (figure 5), in a fuzzy logic environment the system receives these variables as arguments and makes the decision where appropriate in the sets described in the consequent.

4.4 Implementation of the proposed model

As regards the design and implementation of the experiments proposed for the model, the following aspects of the system were defined:

- Discretization or postulation of values.
- Fuzzification using singleton.
- Aggregation of rules using the Mamdani's Model.
- Defuzzification with weighted center-average or centroid.
- Link to functional components.

However, taking into account the use of the antecedents in the system to grant a decision, which in this case is to provide an associated performance value, aspects of the consequent that are being handled from the design were considered during the implementation phase:

- Optimal Performance (OP): It occurs in the best possible execution scenario (from 70% to 100%). This performance is the ideal scenario in an LOD resource consumption environment considering that the query result has the expected values of the defined variables. This index may not reach full performance, given the proposal to work with fuzzy systems, but it allows contemplating all those scenarios in which the query has an excellent level of trust.
- Average Performance (AP): It corresponds to the most standard scenario of the implementation within an environment of uncertainty (from 30% to 70%). This performance scenario is the most known when consuming LOD resources. It denotes the need to define levels of trust for consumption, in addition to some characteristics of the uncertain environment that this environment presents.
- Deficient Performance (DP): It is a frequent scenario, but it shows an unwanted situation of a query (from 0% to 30%). This scenario represents the average performance since its results are unattractive when consuming LOD resources, hence the implementation of the Fuzzy Logic Model from this point could consider representative faults on the resources, and it could make suggestions on the organizations that are in the provision of these queries.
- Critical Performance (CP): It is the worst possible scenario of execution (0%). This is the worst case obtained from a consumption of LOD resources since its results do not produce any synonym of success. Consequently, a series of imminent failures have to be fixed insofar as the decisions are made from inference rules to ensure a successful consumption of resources.

After carrying out the corresponding implementation of the fuzzy sets and the respective rules concerning the VOWL visualization tool [31], services were used in order to: first, carry out a count of requests on Endpoints, second, take a count of the number of relationships, and third, calculate the response time of these Endpoints from the average time between each query request. These aspects are integrated for determining the consumption of linked data from an available Endpoint.

Table 1 Membership Functions

mNSR-HIGH	Membership Function of the NSR Concept of High
mNSR-LOW	Membership Function of the NSR Concept of High
mRT-HIGH	Membership Function of the RT Concept of High
mRT-LOW	Membership Function of the RT Concept of Low
mOP	Membership Function of the Optimal Performance
mAP	Membership Function of the Average Performance
mDP	Membership Function of the Deficient Performance
mCP	Membership Function of the Critical Performance

Table 2 Fuzzy Sets Attributes

mNSR-HIGH Center	50%	**mNSR-HIGH Dispersion**	60
mNSR-LOW Center	50%	**mNSR-LOW Dispersion**	60
mRT-HIGH Center	40	**mRT-HIGH Dispersion**	1
mRT-LOW Center	30	**mRT-LOW Dispersion**	1
mOP Center	1	**mOP Dispersion**	0.1
mAP Center	0.7	**mAP Dispersion**	0.1
mDP Center	0.3	**mDP Dispersion**	0.1
mCP Center	0	**mCP Dispersion**	0.1

5. Results analysis

There are a number of research related to the formulation and determination of methodologies aimed at ensuring the quality of the data consumed, through which it is possible to postulate dimensions such as accessibility, and within which the availability and response time of the data are identified [4]. Thus, the development of a system based on inference rules, by defining fuzzy sets with concept formulation under the main dimensions of quality data, is a research with practical implementation that transcends the methodologies oriented to the analysis of data consumption. In the concept of Open Data, this also provides a dedicated development to the resolution of conflicts when acquiring information.

Evaluating the availability criteria and response time of the Endpoints, in a practical scenario, allows observing the behavior of the linked open data through an automated development that works with methodologies of quality assurance, in addition to offering new characteristics in the context of the consumption of the information in Endpoints of a dataset.

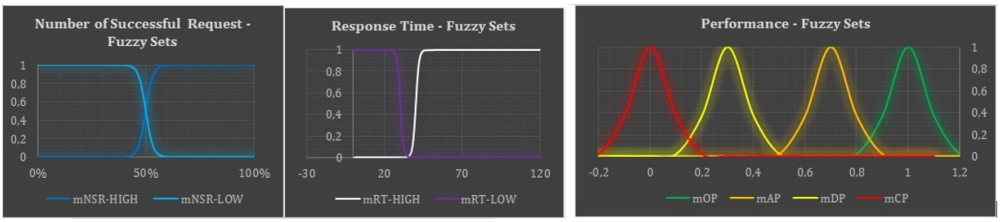
During the system execution phase, derived from these consumptions, figure 6 shows the obtained results, which shows the display of regular VOWL visualization along with the collection of the necessary information for the operation of the designed fussy system (right side of the image). Figure 7 shows the components offered by VOWL:

- *URL:* it corresponds to the endpoint consumed.
- *Pending Requests:* it shows the number of requests pending for completion within the data consumption process.
- *Successful Requests:* it displays the number of completed requests within the data consumption process of the current query.
- *Failed Requests:* it is the number of failed requests within the data consumption process of the current query.

Together with the results obtained by the insertion of fuzzy sets (Figure 7):

- *Times:* it displays the average time among query rounds. In the business layer, this time establishes the metric the system manages to adjust the concept under the corresponding set to the response times, and describes the mechanism the argument uses to interact with the inference rules to produce a result, which translates into a performance index of the level of trust obtained.
- *Performance:* it displays the performance index of each query, which is the result of the system based on inference rules after having worked with the arguments obtained from the number of successful requests and the response time, in a presentation layer format that translates levels of trust processed during data consumption into a metric indicator on quality data.

As seen in the queries performed, where Endpoints with a small amount of data were consumed, the Performance indicator (as analysis between complete requests versus time) varies, which establishes that the higher the



(a) Definition of diffuse set: antecedents (b) Definition of diffuse set: Consequents

Figure 3 Definition of diffuse set

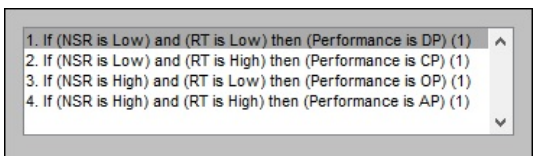


Figure 4 Design of logical rules

percentage value is, the higher is also the level of trust offered by the Endpoint. In this section, it is important to highlight some aspects to consider because these influence the performance of the query:

- Endpoints with large volumes of data, their response time may be high but reliable.
- Performance is influenced by how the platform is implemented where the Endpoint is, and by the Content Provider being part or not of its entire data-binding platform.

To validate the performance, additional executions were carried out with different SPARQL endpoints in a random way (Figure 8) to obtain detailed information in percentage terms about the requests completed (return rate). This allows defining the trust levels in each of them according to the criteria defined in the model, in the sense that now the system recognizes the trace of ambiguity which exists when consuming data on each Endpoint, and can determine the state of the Endpoint based on the dimensions described by the concepts of fuzzy sets. Finally, according to the implementation of the fuzzy system designed for the application, the surface graph (Figure 9) shows how the interaction between the two input arguments of the dataset query emits a result that is translated in terms of associated performance to the consultation rounds.

It is thus that Figure 9 allows a more in-depth understanding of the interior of the proposed fuzzy system, and how at runtime the possible arguments used would yield a contemplated response in the design of the ensuing sets which is reflected in terms of the performance of the query to the requested dataset.

6. Discussions

As can be seen in the results obtained, the management of response time and availability metrics provide information on how the trust offered by published SPARQL endpoints can be evaluated. The performance obtained in the Endpoint consumption, applying the fuzzy sets, allows to classify them in levels of trust proposed in this research. This process contributes to obtain a differentiation in the experience that the user lives in the consumption of linked resources.

In this context, the response time and availability attributes can be used [4] to determine levels of trust in the Endpoint consumption. The main objective of our framework contributes to the aim proposed in [31]: *"The need for new developments, based on specific analyzes of datasets, to determine the most appropriate and practical basis for creating a comprehensive conceptual framework for data quality".*

Similarly, these authors [31] state:

- Availability, such as the ease of locating and obtaining an information object (for example, datasets), a dimension for which metrics such as server accessibility, SPARQL Endpoint accessibility, dereferenciability problems, and others can be used.
- Response time, such as the amount of time until the complete response reaches the user. Dimension for which the percentage of response in time (per second) can be measured.

Thus, determining the tacit importance of relating the number of successful and/or completed requests along with the time spent for each of them, making possible the creation of the Type-I fuzzy logic system presented in this research.

Therefore, the obtained results (Figure 10) are aligned with the indicators and metrics proposed in [34], insofar as the performance achieved by each SPARQL Endpoint is understood as optimal as a function of: a) a time of fast response, and b) the resolution of a high number

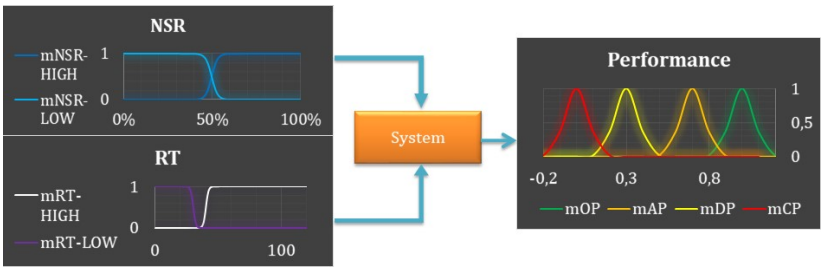


Figure 5 Design of the fuzzy system based on logical rules

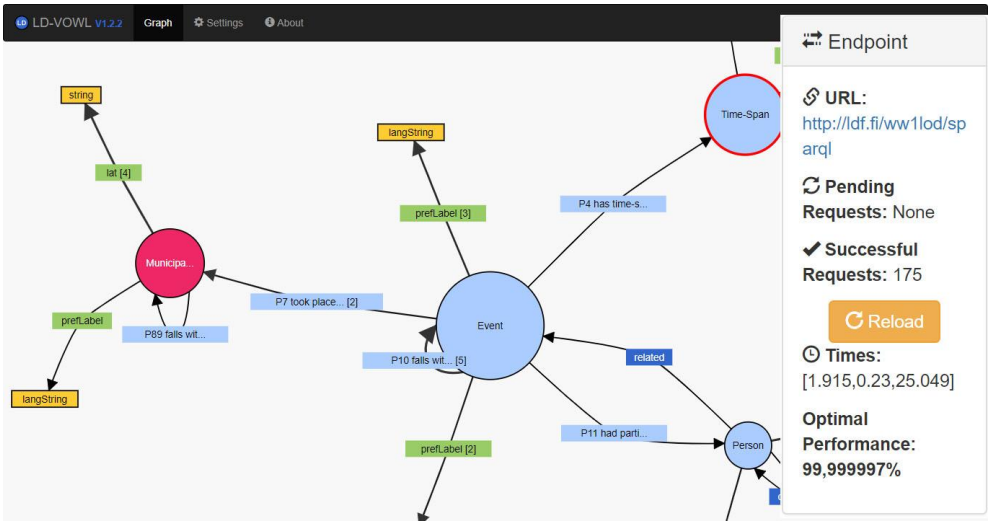


Figure 6 VOWL display with fuzzy logic component

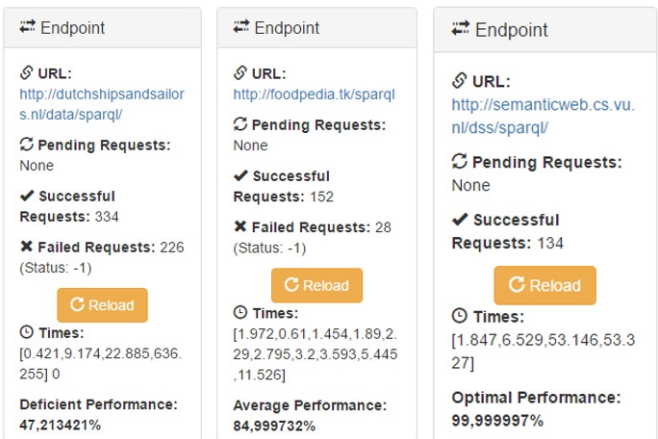


Figure 7 Endpoint consumption results

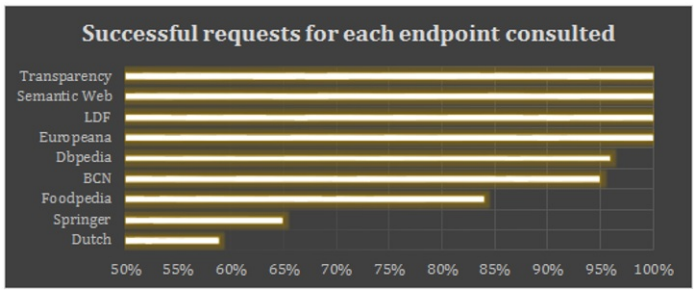


Figure 8 Percentage of successful requests

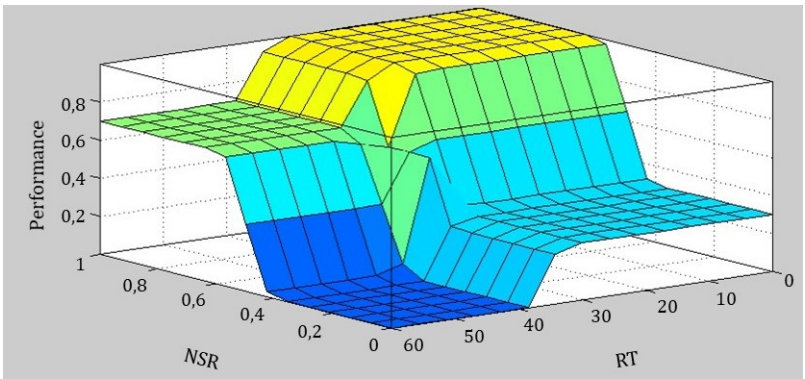


Figure 9 Surface Graph Interaction Arguments

Figure 10 Resultant Variables

of requests according to how the functions of belonging were conceived for the established fuzzy logic system. In this way, a favorable index for the search proposal of levels of trust was obtained, and it is presented in this research work. The above arguments can be evidenced in the research exposed in [34], where it is detailed that:

- The higher the accessibility metric is, it is better (it is recommended > 1 req. p/sec)
- The lower the response time metric is, it is better (<0.5 seconds per request)

Metrics that when contrasted allow obtaining the categorization of the "Performance" proposed in this research (Optimal, Acceptable, Poor and Critical).

Finally, in order to compare this approach with other, different investigations like [21, 22] have used VOWL to expose a uniform visual notation for OWL ontologies based on aspects such as clarity and distinguishability of elements, ease of use with interactive highlighting and design features, as well as aesthetics.

On the other hand, the implementation of fuzzy logic techniques has made it possible to use the information provided by the Endpoint queries, and to be able to establish classification criteria that allows evaluating the level of trust offered by each Endpoint. In proposals such as [25, 26, 35], the visualization of the information resulting from the Endpoint Consumption is observed. Also, the suggestion for future works regarding the inclusion of new quality criteria that will allow evaluating the trust offered by such Endpoints.

In conjunction with the above, it becomes apparent that the application of fuzzy sets for the evaluation of the metrics of availability and response time, in addition to aligning with the proposed theoretical references, offers

a development based on specific analyses of datasets. These fuzzy sets allow determining appropriate and practical criteria to provide a guidelines of reliability in the consumption of data, and therefore in data quality.

7. Conclusions

The process of adding trust to published resources starts from an adequate, clear and public abstraction of a data model. Subsequently, monitoring the recommendations described to carry out the exposure and linkage process contributes to the fact that the published data offers both greater usability and a higher value of trust to the user.

The above can be seen in aspects such as: the data have metadata, its resources have an identifier and can be referenced, the data are published in suitable formats, they can be reused, and the platforms where they are stored offer services for query, among other factors. All of this contributes to the construction of trust by those who aspire to consume those resources.

In addition, evaluating different quality dimensions, offering consumers relevant mechanisms and indicators of these dimensions resulting from the application of these mechanisms contributes to the building of trust in the resources exposed. Such is the case of indicators like availability and response time, indicators that allow evaluating if a resource is available and how long it takes to resolve its referencing. These factors allow the consumers to build trust by giving them information about whether the resource they want to access is available, referenceable, and resolved in a relatively acceptable time.

As a result, the Type-I Fuzzy Systems were selected as mechanisms to evaluate the proposed quality dimensions with a design of fuzzy logic rules based on two variables. These variables were obtained from the query component to the Endpoint: the number of successful requests on the total number of requests in a query, and the average response time taken by each query round in resolving data consumption.

This mechanism, coupled with a visualization tool such as VOWL, allowed evidencing and analyze the "Time" and "Performance" indexes associated to data consumption and included in this research, which provide information about the response time and performance level obtained, categorizing in this way the level of trust (Optimal, Acceptable, Poor, Critical) obtained in the consumption of such resources. Therefore, through the proposed approach, the fuzzy aspects immersed in a process of data consumption were conceived in order to take these feedbacks and contemplate a more horizontal scenario in this area.

In general, the design and implementation of a Type-I fussy system was projected, based on logical rules for the definition of trust levels, as a preliminary model that allows a more robust control of the data recovery in the consumption processes of the Endpoint.

On the other hand, the use of such sets opens the possibility of constructing more proposals or models, more rigorous ones by adding different classifications of each of the schemas and datasets, optimizing and greatly improving the levels of trust, which will increasingly favor the quality of use and consumption of Linked Data. However, it is necessary to consider other work scenarios in other repositories and SPARQL connection points to define rules and methods that encompass and collect more schemas and resources in the Linked Data framework.

As further work, it is proposed to include more elements that feed the fuzzy system to achieve a much more detailed analysis of each of the schemas available in the linked data framework. In this context, it is intended to gather more dimensions in the framework of computational learning, which allows defining levels of trust from more rigorous processes. Moreover, it is intended to reduce the uncertainty about the response time of Endpoints with a greater coverage of variables. On the other hand, future solutions might consider parameters that could be inferred in the queries while being able to increase the feasibility and trust, for example, in broken links.

8. Acknowledgement

This research has been developed within the framework of the doctoral research project on Linked Data, at the District University Francisco José de Caldas. In the same way, the subject matter is working as a line of Research Group GIIRA.

9. References

[1] T. Berners, J. Hendler, and O. Lassila. "The Semantic Web," *Scientific American*, vol. 284, no. 5, pp. 29-37, 2001.

[2] A. A Rodríguez "Las nuevas pautas para el acceso a la información," *Investigación Bibliotecológica: Archivonomía, Bibliotecología e Información*, vol. 30, no. 69, pp. 117-135, 2016.

[3] W3C, Data on the Web Best Practices, 2017. [Online]. Available: `https://w3c.github.io/dwbp/bp.html`. Accessed on: July 25, 2017.

[4] A. Zaveri, *et al.*, "Quality Assessment Methodologies for Linked Open Data. A Systematic Literature Review and Conceptual Framework," *IO Press Journal*, no. 1, pp. 1-5, 2012.

[5] L. Page, S. Brin, R. Motwani, and T. Winograd, "The PageRank citation ranking: Bringing order to the web," Stanford University, Stanford, CA, Tech. Rep., Jan. 1998.

[6] J. Pattanaphanchai, "DC Proposal: Evaluating Trustworthiness of Web Content Using Semantic Web Technologies," in *The Semantic Web ISWC: 10th International Semantic Web Conference*, Berlin, Heidelberg, 2011, pp. 325-332.

[7] E. Rajabi, S. Sanchez, and M. A. Sicilia, "Analyzing broken links on the web of data: An experiment with DBpedia," *Journal of the Association for Information Science and Technology*, vol. 65, no. 8, pp. 1721-1727, 2014.

[8] M. E. Vidal, *et al.*, "Retos para el Procesamiento Semántico de Datos Enlazados en la Nube de los Datos Abiertos Enlazados," *Revista Venezolana de Computación*, vol. 1, no. 1, pp. 34-42, 2014.

[9] LinkedData.Center, *What is a Dataset?* 2017. [Online]. Available: http://sites.linkeddata.center/help/business/starter-kit/dataset. Accessed on: July 20, 2017.

[10] A. Ahmad, R. Troncy, and A. Senart, "What is up LOD Cloud? Observing the State of Linked Open Data Cloud Metadata," in *International Semantic Web Conference*, Portorož, Slovenia, 2015, pp. 247-254.

[11] T. Berners, *Linked Data*, 2009. [Online]. Available: https://www.w3.org/DesignIssues/LinkedData.html, Accessed on: May 20, 2017.

[12] Semantic Web,*SPARQL Endpoint*, 2011. [Online]. Available: http://semanticweb.org/wiki/SPARQL_Endpoint.html. Accessed on: June 20, 2017.

[13] Europeana pro, *Europeana SPARQL Endpoint*, 2016. [Online]. Available: https://pro.europeana.eu/post/europeana-sparql-endpoint. Accessed on: July 12, 2017.

[14] Y. Kabutoya, R. Sumi, T. Iwata, T. Uchiyama, and T. Uchiyama, "A Topic Model for Recommending Movies via Linked Open Data," in *IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology*, Macau, China, 2012, pp. 625-630.

[15] Y. Gil and D. Artz, "Towards content trust of web resources," *Web Semantics: Science, Services and Agents on the World Wide Web*, vol. 5, no. 4, pp. 227-239, 2007.

[16] P. Vandenbussche, A. Hogan, J. Umbrich, and C. Buil, "SPARQL: A Gateway to Open Data on the Web?" *ERCIM NEWS Linked Open Data*, vol. 1, no. 96, pp. 31-31, 2014.

[17] VisualDataWeb, *LD-VOWL: Visualizing Linked Data Endpoints*, 2016. [Online]. Available: http://vowl.visualdataweb.org/ldvowl.html. Accessed on: May 16, 2017.

[18] S. Lohmann, F. Haag, and S. Negru, *Towards a Visual Notation for OWL: A Short Summary of VOWL*, 2013. [Online]. Available: https://pdfs.semanticscholar.org/2c26/fdeac23c325f2832ed786a0beba656605522.pdf. Accessed on: June 28, 2017.

[19] AIMS Agricultural Information Management Standards, *VOWL: Visual Notation for OWL Ontologies*, 2014. [Online]. Available: http://aims.fao.org/community/general-information/blogs/vowl-visual-notation-owl-ontologies. Accessed on: July 8, 2017.

[20] T. Ramdane and S. Bachir, *OWL Visual Notation and Editor*, 2014. [Online]. Available: http://www.univ-tebessa.dz/fichiers/ouargla/icaiit2014_submission_217.pdf. Accessed on: July 9, 2017.

[21] M. Weise, S. Lohmann, and F. Haag, "LD-VOWL: Extracting and Visualizing Schema Information for Linked Data," in *2nd International Workshop on Visualization and Interaction for Ontologies and Linked Data*, Kobe, Japón, 2016, pp. 120-127.

[22] S. Lohmann, S. Negru, F. Haag, and T. Ertl, *Visualizing Ontologies with VOWL*. [Online]. Available: http://www.semantic-web-journal.net/system/files/swj1114.pdf. Accesed on: July 23, 2017.

[23] D. Bold, S. Lohmann, and S. Negru, *Protégé VOWL: VOWL Plugin for Protégé*, 2014. [Online]. Available: http://vowl.visualdataweb.org/protegevowl.html. Accessed on: July 23, 2017.

[24] M. D. Rojas, E. Zuluaga, and J. Ochoa, "Propuesta de medición de la confianza en la información utilizando un sistema difuso," *Revista Ingenierías Universidad de Medellín*, vol. 10, no. 19, pp. 113-123, 2011.

[25] S. Javanmardi, M. Shojafar, S. Shariatmadari, and S. S. Ahrabi, "FR TRUST: A Fuzzy Reputation Based Model for Trust Management in Semantic P2P Grids," *International Journal of Grid and Utility Computing*, vol. 6, no. 1, pp. 57-66, 2014.

[26] I. Jacobi, L. Kagal, and, A. Khandelwal, "Rule-Based Trust Assessment on the Semantic Web," in *International Workshop on Rules and Rule Markup Languages for the Semantic Web*, Barcelona, Spain, 2011, pp. 227-241.

[27] S. Kamal, S. F. Nimmy, and L. Chowdhury, "Fuzzy Logic over Ontological Annotation and Classification for Spatial Feature Extraction," *International Journal of Computer Applications*, vol. 41, no. 6, pp. 18-22, 2012.

[28] R. Aringhieri, E. Damiani, S. De Capitani, S. Paraboschi, and P. Samarati, "Fuzzy Techniques for Trust and Reputation Management in Anonymous Peer-to-Peer Systems," *Journal of the American Society for Information Science and Technology*, vol. 57, no. 4, pp. 528-573, 2006.

[29] S. Lohmann, F. Haag, and S. Negru, "Towards a Visual Notation for OWL: A Brief Summary of VOWL," in *International Experiences and Directions Workshop on OWL*, Bologna, Italy, 2016, pp. 143-153.

[30] W3C, *OWL Web Ontology Language Overview*, 2004. [Online]. Available: https://www.w3.org/TR/owl-features/. Accessed on: July 25, 2017.

[31] 31. GitHub, Inc, *LD-VOWL*, 2016. [Online] Available: https://github.com/VisualDataWeb/LD-VOWL. Accessed on: April 25, 2017.

[32] J. M. Mendel, "Fuzzy logic systems for engineering: a tutorial," *Proceedings of the IEEE*, vol. 83, no.3, pp. 345-377, 1995.

[33] A. Peregrín, "Integración de Operadores de Implicación y Métodos de Defuzzificacion en sistemas Basados en Reglas Difusas. Implementación, análisis y Caracterización," M.S. thesis, Universidad de Granada, Granada, España, 2000.

[34] C. Bizer, Z. Miklos, J. Calbimonte, A. Moraru, and G. Flouris, *D2.1 Conceptual model and best practices for high-quality metadata publishing*, 2012. [Online]. Available: http://planet-datawiki.sti2.at/uploads/archive/d/d7/20120217160352%21D2.1.pdf. Accessed on: May l5, 2017.

[35] VOILA 2016 Visualization and Interaction for Ontologies and Linked Data, *Visualization for Ontology Evolution, 2016*. [Online] Available: http://voila2016.visualdataweb.org/VOILA2016_proceedings.pdf. Accessed on: May 30, 2017.