Uso de lógica difusa como estrategia para evaluar la confianza y accesibilidad de los DataSet publicados en SPARQL Endpoints

Use of fuzzy logic as a strategy to evaluate trust and accessibility of DataSet published in SPARQL Endpoints

Jhon Francined Herrera-Cubides[1], Paulo Alonso Gaona-García[2], Carlos Enrique Montenegro-Marín[3], Esteban Arias-Caracas[4], Daniel Fernando Mendoza-López[5]

[1]*Universidad Distrital Francisco José de Caldas, Bogotá, Colombia,* jfherrerac@udistrital.edu.co

[2]*Universidad Distrital Francisco José de Caldas, Bogotá, Colombia,* pagaonag@udistrital.edu.co

[3]*Universidad Distrital Francisco José de Caldas, Bogotá, Colombia,* cemontenegrom@udistrital.edu.co

[4]*Universidad Distrital Francisco José de Caldas, Bogotá, Colombia,* eariasc@correo.udistrital.edu.co

[5]*Universidad Distrital Francisco José de Caldas, Bogotá, Colombia,* daltemen@hotmail.com

Fecha de recepción: 03/09/2018 Fecha de aceptación: 04/29/2018

DOI: https://doi.org/10.18041/1794-4953/avances.1.4737

Como citar: Herrera - Cubides, J. F., Gaona-García, P. A., Montenegro-Marín, C. E., Arias-Caracas, E. & Mendoza-López, D. F. (2018). Uso de lógica difusa como estrategia para evaluar la confianza y accesibilidad de los DataSet publicados en SPARQL Endpoints. *AVANCES: INVESTIGACIÓN EN INGENIERÍA,* 15 (1), 231-255. DOI: https://doi.org/10.18041/1794-4953/avances.1.4737

Resumen

La web semántica ha proporcionado herramientas como Linked Data, la cual ha permitido adelantar procesos de vinculación de datos abiertos. De cara a la creciente publicación de datos abiertos, se encuentran los retos de calidad de los datos vinculados, requerimientos vitales para el beneficio de los consumidores que deseen utilizar los datos publicados. Para abordar estos retos, se han generado modelos y herramientas que permiten evaluar la calidad de los datos. Una gran parte de estas herramientas basan su quehacer en la medición de variables haciendo uso de modelos matemáticos tradicionales, restringiendo la valoración misma de la calidad. Este artículo se orienta

en proponer un método de evaluación de datos abiertos bajo especificaciones Linked Open Data, publicados en SPARQL Endpoints, mediante la implementación de un modelo basado en lógica difusa. Este modelo permitirá comparar los modelos tradicionales de evaluación sin la necesidad de restringir los aspectos de calidad con los que se puede medir. Finalmente se presentan los resultados obtenidos y trabajos futuros.

Palabras clave: datos abiertos vinculados, lógica difusa, medición de la calidad, SPARQL.

Abstract

The Semantic web has provided tools such as Linked Data, which has allowed for the advancement of Linked Open Data processes. With the growing publication of open data, different challenges have been generated in terms of the linked data quality, vital requirements for the benefit of consumers who wish to use the published data. To address these challenges, models and tools have been generated to evaluate data quality. A large part of these tools base their work on the measurement of variables using traditional mathematical models, restricting the quality assessment itself. This paper aims to propose a method for evaluating open data under Linked Open Data specifications, published in SPARQL Endpoints, by implementing a model based on fuzzy logic. This model will allow comparing traditional evaluation models, without the need to restrict the quality aspects with which it can be measured. Finally, the results obtained and future work are presented.

Keywords: Linked Open Data, fuzzy logic, quality measurement, SPARQL.

1. Introducción

Tecnologías de la web semántica, tales como Linked Data, soportada sobre la filosofía de Open Data, han venido incursionando cada vez con más fuerza en los procesos de publicación, distribución y consumo de datos en la web [1]. Haciendo uso de dichas tecnologías, el cargue y actualización de datos en la web se ha vuelto una labor un poco más fácil. Estas actividades pueden ser realizadas por cualquier tipo proveedor, sea este individuos, pequeños grupos de personas, organizaciones educativas, sitios de redes sociales e incluso organismos gubernamentales [2]. Dada la gran variedad de proveedores de datos, la vinculación de datos abiertos ha crecido de manera exponencial, pasando de 12 DataSet publicados en 2007, a cerca de 300 en 2011, y 9,960 DataSet en 2016. Datos acumulados de tres de las principales colecciones de DataSet disponibles al público: data.gov, publicdata.eu y datahub.io [3].

Con este crecimiento de Linked Open Data (LOD), surge la necesidad de establecer estrategias o herramientas

que permitan evaluar y gestionar la calidad de los datos publicados en la web. Lo anterior dado que en estudios como [4] y [5] la información proporcionada por la mayoría de los recursos publicados en la web no posee una estructura adecuada para el proceso de vinculación, además de presentar desafíos en cuanto a información relevante para ser consultada sobre los recursos vinculados.

Con base en estos desafíos, diferentes investigaciones han identificado variables a evaluar en los procesos de vinculación de datos, tales como seguridad, estabilidad, rapidez y precisión [6]. Teniendo en cuenta que la calidad es un concepto muy subjetivo, no se puede definir de una sola manera por un simple juicio [2, 6]. Bajo este contexto, Tim Berners-Lee estableció un esquema técnico de publicación LOD en 2011 basado en 4 principios [7]:

- Usar URI para nombrar cosas u objetos conceptuales.
- Usar URI HTTP que sean interpretables por humanos y máquinas.
- Proveer información útil acerca de cada URI en algún estándar de la web (p. ej. RDF).
- Crear links entre URI.

Dichos principios se configuran como la base para el esquema de vinculación de datos, que consta de los siguientes niveles [7]:

- Nivel 1: Publicar la data en la web, descrita en algún tipo de formato, bajo una licencia abierta.
- Nivel 2: Los datos están publicados como datos estructurados.
- Nivel 3: Para la publicación se usan formatos no propietarios.
- Nivel 4: Usar URI para identificar cosas y propiedade, de manera que se pueda apuntar a los datos.
- Nivel 5: Los datos se encuentran enlazados a otros datos.

Este esquema ha definido un conjunto de requisitos mínimos para llevar a cabo un adecuado proceso de vinculación. Con el fin de evaluar, entre otros criterios, el cumplimiento de dichos requisitos mínimos, se han elaborado diferentes métodos para extraer y agrupar datos, con el objeto de evaluar su calidad. Algunas de las estrategias diseñadas para abordar dicho proceso contemplan la creación de software para evaluar la calidad de los datos, haciendo uso de procedimientos matemáticos.

Dentro de las estrategias de software de medición de calidad, se encuentra Luzzu [6], que corresponde a un framework genérico basado en DataSet Quality Ontology, que permite a los usuarios definir sus propias métricas de calidad. Esta es una plataforma integrada que:

- Evalúa la calidad de los datos vinculados utilizando una biblioteca de métricas de calidad específicas de dominio genéricas y provistas por el usuario de una manera escalable.
- Proporciona metadatos de calidad consultables en los conjuntos de datos evaluados.
- Ensambla informes detallados de calidad sobre los conjuntos de datos evaluados.

Aunque existen este tipo de herramientas, gran parte de los resultados obtenidos de su uso son extraídos de modelos matemáticos clásicos o convencionales. Estos modelos principalmente abarcan soluciones lineales o continuas, de acuerdo con los resultados. Es decir, que estos modelos matemáticos toman, como ejemplo, valores binarios como Sí o No para la evaluación de las distintas variables, generando que la calidad calculada posea cierto grado de incertidumbre.

A partir de este contexto, este artículo pretende responder a la problemática de: ¿Cómo calcular la calidad de los datos publicados como LOD, de una manera más precisa, con la capacidad de no perder tal precisión con mayores cantidades de variables de calidad?

Para dar respuesta a este interrogante, se planteará un modelado basado en lógica difusa para proporcionar la medición de calidad de datos vinculados, teniendo en cuenta la tolerancia de valores intermedios entre evaluadores convencionales [8].

Dentro de la literatura explorada sobre calidad de datos vinculados, se han identificado diferentes modelos que contienen dimensiones y métricas de calidad. Para esta investigación se tomó como base el modelo expuesto en [9]. Teniendo como referente este modelo de calidad, la presente investigación cubrirá las dimensiones de confianza y accesibilidad. Al interior de estas dimensiones, se representarán cuatro aspectos principales: tiempo de respuesta (*response time*), escalabilidad (*scalability*), confiabilidad (*trustworthiness*) y puntualidad (*timeliness*).

Por otra parte, el uso de un modelo de lógica difusa permitirá evaluar las dimensiones propuestas gracias a la compatibilidad que dichas dimensiones poseen con los esquemas de lógica difusa. Además, mostrar unos resultados con menor incertidumbre, dado que el cálculo matemático basado en el modelado centroide es más preciso comparado con los valores convencionales de respuesta binaria.

Para desarrollar esta investigación, en la sección 2 del presente artículo, se lleva a cabo la contextualización del marco teórico y trabajos relacionados en el tema. La sección 3 expone la metodología seguida del diseño metodológico usado para desarrollar esta investigación. La sección 4 presenta el modelo de lógica difusa propuesto. La sección 5 expone el análisis de los resultados obtenidos. La sección 6 presenta la discusión propuesta. Y finalmente la sección 7 las conclusiones y trabajo futuro.

2. Background

A continuación se exponen los referentes principales que sustentan esta investigación.

2.1 Linked Open Data (LOD)

En primera instancia, la apertura de datos denota la habilidad de diversos sistemas y organizaciones para trabajar juntos, en

otras palabras, la habilidad para interoperar o integrar diferentes DataSet.

Por ende, LOD trabaja sobre datos abiertos en RDF, permitiéndoles a los usuarios enlazar datos provenientes de diversas fuentes, instituciones u organizaciones, explorar y combinar estos datos de manera libre y sin restricciones de copyright para nuevos desarrollos web [10, 11].

Adicionalmente, los grafos de conocimiento (Figura 1), generados en los procesos de vinculación de los recursos expuestos en la web, ofrecen la posibilidad de clasificar los datos publicados acorde con sus posibles usos y aplicaciones, en dominios de conocimiento.

LOD ha venido experimentando un crecimiento en diferentes dominios, sobre todo en los ambientes públicos y de gobierno. Dada la incursión de diferentes organizaciones, y el incremento de la apertura de datos, se hace cada vez más necesario abordar estrategias acerca de

Figura 1. Grafo de conocimiento [7].

cómo poder definir la calidad de estos DataSet.

Para evaluar dicha calidad, existen formas o variables a tener en cuenta a la hora de medir la calidad de los datos. Sin embargo, investigadores como [10, 13] concuerdan en la necesidad de definir dimensiones o métricas para poder visualizar diferentes campos y así clasificar estas variables de manera más generalizada.

2.2 Dimensiones de calidad de datos

Como se plantea en [9], la calidad de los datos, entendida como la "aptitud para el uso", puede depender de varias dimensiones como precisión, puntualidad, integridad, relevancia, objetividad, credibilidad, comprensibilidad, consistencia, concisión, disponibilidad y verificabilidad. Estas dimensiones se pueden definir como características de un DataSet, y ser agrupadas de acuerdo con sus propiedades, de tal forma que se puedan clasificar para calcular la calidad, bien sea de manera subjetiva u objetiva. Lo anterior considerando el hecho de que darle un valor numérico a una variable de calidad puede ser subjetivo o abstracto.

Dentro de los modelos explorados se identifican diferentes dimensiones, tales como contextuales, de confianza, intrínsecas, accesibles, representativas, dinámica de DataSet, entre otros. Para esta investigación, y basados en el modelo de calidad propuesto por [9] (Figura 2), se seleccionaron las dimensiones de confianza y accesibilidad.

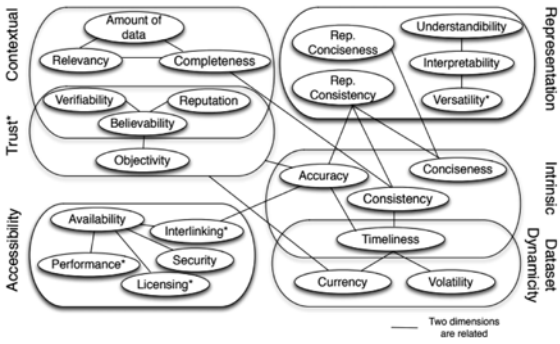

Figura 2. Calidad en LOD [9].

En dicho modelo, los autores definen que:

- La dimensión de confianza analiza el nivel de fiabilidad que un consumidor de datos considera acerca de los datos a su disposición.
- La dimensión de accesibilidad permite el análisis del rendimiento del conjunto de datos, lo cual representa la capacidad de que el DataSet responda a las solicitudes del usuario y retorne información de consulta de manera efectiva y rápida.

A su vez, dentro de las dimensiones de accesibilidad y confianza, se agrupan diferentes variables de calidad a tomar en cuenta. En cuanto a la dimensión de accesibilidad, esta posee seguridad, disponibilidad, desempeño, tiempo de respuesta, entre otros, mientras que, por su lado, la dimensión de confianza posee variables como origen, credibilidad, reputación, licenciamiento, entre otras. Bajo estos modelos de calidad, varias de las dimensiones de calidad no poseen una forma cuantificable de poder medirlas, generando así discrepancias a la hora de intentar obtener un cálculo matemático. Sin embargo, gracias a los aportes de distintos autores, y basados en el modelo presentado por [9], se logra una aproximación matemática para aplicarse al modelo de lógica difusa.

Para el planteamiento del presente modelo, se eligieron cuatro variables: tiempo de respuesta y escalabilidad (dimensión de accesibilidad), confiabilidad y puntualidad (dimensión de confianza). La selección de estas cuatro variables responde a la compatibilidad y capacidad de implementación dentro del modelo matemático de la lógica difusa.

2.3 Lógica difusa

Como lo plantea [8], la lógica difusa permite una lógica multivaluada, la cual acepta valores intermedios entre evaluaciones convencionales como verdaderas o falsas, sí o no, alto o bajo, entre otros. Este tipo de sistemas permiten un cálculo más preciso de acuerdo con las entradas matemáticas que requieran, en comparación a los modelos matemáticos binarios.

Tal como lo describe [14], cualquier problema del mundo puede resolverse dado un conjunto de variables de entrada (espacio de entrada), obtener un valor adecuado de variables de salida (espacio de salida). La lógica difusa permite establecer este mapeo de una forma adecuada, atendiendo a criterios de significado (y no de precisión).

Para el trabajo en esta investigación, y con base en propuestas como [15],

donde se plantea el modelamiento e implementación de un sistema difuso de tipo-I basado en reglas lógicas, orientado a la toma de decisiones sobre la incertidumbre que presenta la definición de niveles de confianza en el consumo de DataSet LOD; se escogió un modelo tradicional de lógica difusa. El modelo de centroide permite resultados dentro de rangos porcentuales, lo cual facilita el establecimiento de un enfoque más aproximado a una medición de calidad de LOD.

2.4 Trabajos relacionados

Partiendo del auge de LOD durante los últimos años, junto con la inclusión de la lógica difusa como metodología para calcular procesos matemáticos, como resultado de la revisión bibliográfica realizada, acerca del uso de lógica difusa como estrategia para evaluar la calidad en LOD, se encontraron los siguientes referentes importantes:

Uno de los trabajos más destacados es el realizado por Lewis y Martin [16], en cuya investigación se utiliza la lógica difusa para el análisis de las ontologías y los vocabularios dentro de LOD. Haciendo uso del modelo de centroide, los autores demostraron su implementación haciendo uso de la aproximación X-μ., con un enfoque orientado hacia las ontologías LOD.

Por otro lado, existen frameworks desarrollados para evaluar la calidad de los datos vinculados. Uno de estos frameworks se conoce como Luzzu, planteado por [6] (Figura 3). Este trabajo aborda la medición de calidad de los datos a través del uso de mecanismos de puntuación de usuario. Adicionalmente, plantea el manejo de 22 aspectos de calidad o métricas, relacionadas a través de nueve dimensiones de calidad [6]. Luzzu utiliza cálculos matemáticos de aproximaciones logarítmicas con redondeo, a la vez de cálculos binarios para variables de calidad como seguridad.

Otro de los frameworks identificados corresponde al propuesto en [15]. Este modelo se compone de un framework de visualización (VOWL), y de un conjunto de reglas, basadas en Sistema Fuzzy Tipo-I, para la declaración de niveles de inferencia con respecto a las dimensiones de consumo de datos.

3. Metodología planteada

Para esta investigación se decidió utilizar un estudio de tipo cuasiexperimental, el cual permite estimar el impacto de los diferentes aspectos de calidad para obtener una calidad global, sin el uso

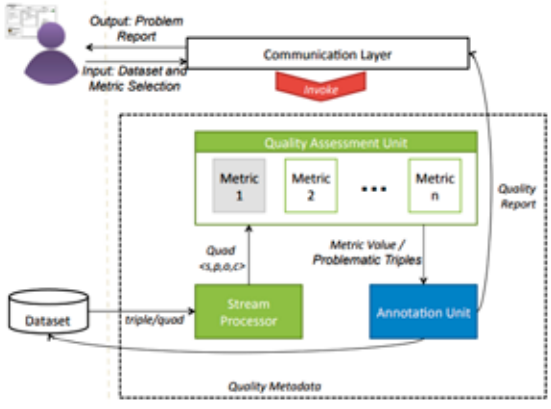

Figura 3. Framework Luzzu [6].

de variables aleatorias en el proceso. Para llevar a cabo una mejor representación, a continuación en la Figura 4 se presenta el diseño metodológico utilizado en el estudio.

Para llevar a cabo el diseño metodológico, en una primera fase se realiza la obtención de datos internos de cada DataSet, conocidos como metadata. Para tal fin, se consultan los SPARQL Endpoints (SPARQL Protocol and RDF Query Language), a través de un flujo de salida que agrupa esta metadata. Los Endpoints permiten ejecutar consultas de datos tipo grafos RDF, a través de sentencias tipo query. La metadata obtenida es usada para obtener los valores de entrada del modelo, y se clasifican de acuerdo con las dimensiones. Es necesario recalcar que el enfoque utilizado requiere tanto la metadata obtenida de la consulta SPARQL, para medir los valores de confiabilidad y puntualidad, como datos externos dentro del framework, para medir los valores de tiempo de respuesta y escalabilidad. De esta forma, es posible conocer o determinar una mejor aproximación de calidad general de LOD bajo las dimensiones de calidad seleccionadas.

Como población para esta investigación, se toma en cuenta el DataSet de opendata.cz, conocido por su interés en "construir una infraestructura de datos abierta que permita acceder a los datos públicos en la República Checa" [17]. La razón principal de la elección de este Endpoint reside en su accesibilidad y capacidad de consulta. Uno de los mayores problemas de los Endpoints es su capacidad para consultar gran cantidad de DataSet relacionados, llevando a resultados incompletos o no funcionales. Esta problemática puede ser controlada, gracias al uso de la plataforma Virtuoso, la cual permite realizar consultas a los Endpoint; el uso de los estándares de la web semántica de RDF

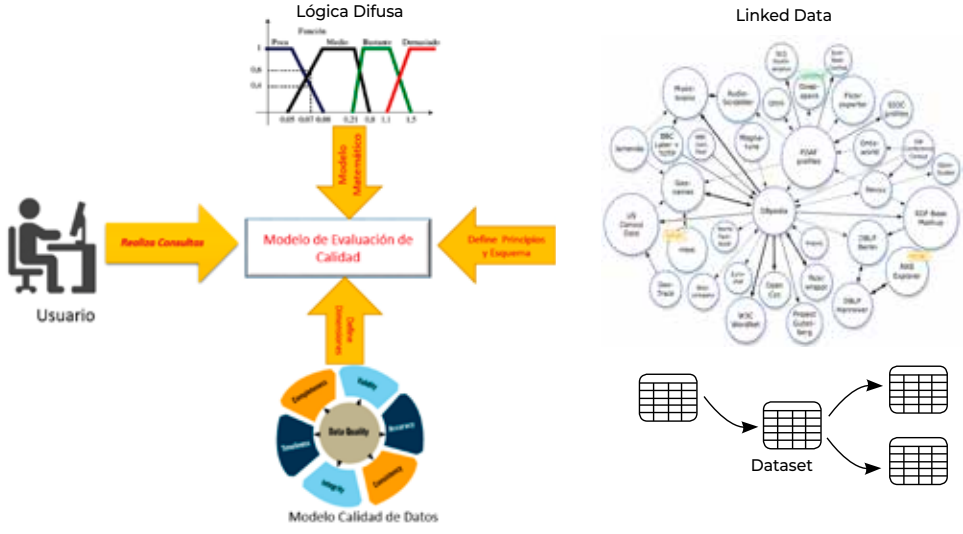

Figura 4. Diseño metodológico.
Fuente: Elaboración propia.

y el uso de los Vocabularios DCTERMS (Dublin Core Terms) y FOAF (Friend of a Friend) dentro del Endpoint utilizado.

Para llevar a cabo este trabajo, se tomaron 100 DataSet asociados a Opendata.cz, cuyo dominio principal se orienta a temas relacionados con sitios educativos u organizaciones públicas de República Checa. La consulta SPARQL utilizada en el modelo de prueba obtiene como algunos resultados de metadata: título, descripción, creador (importante para el análisis de la confiabilidad), contribuyentes (importante para el análisis de la confiabilidad), fechas de creación y modificación (importantes para el análisis de la puntualidad). Los resultados obtenidos se agrupan en un archivo JSON (JavaScript Object Notation), para su posterior adición al modelo. El SPARQL Endpoint utilizado posee una mayor cantidad de vocabularios RDF que DCTERMS y FOAF (Friend of a Friend), creando así una discrepancia en cuanto a elegir una mayor cantidad de conjuntos de datos elegidos para el estudio, considerando que las consultas serían imprecisas en cuanto a su información.

Como variables a analizar, se tomaron los cuatro aspectos previamente mencionados: tiempo de respuesta, escalabilidad, confiabilidad y puntualidad, asociados a las dimensiones de accesibilidad y confianza. La razón por la cual estas dimensiones fueron elegidas reside en la importancia que cada dimensión genera en los factores vinculados a la opinión del cliente y a la funcionalidad del DataSet. La combinación de estos factores permite realizar una aproximación más cercana a la medición de la calidad de los datos, así como poder construir un modelo más comprensible para el usuario.

Como instrumentos para la recolección de la información se utilizó el SPARQL Endpoint correspondiente al DataSet principal de estudio.

Dentro del plan de análisis del modelo, se decidió comparar valores de calidad entre dos casos, para la visualización del efecto de más o menos variables incluidas dentro del modelo de lógica difusa. Con relación a ello, para el primer caso se toman los cuatro aspectos de calidad y se mide la calidad con ellos. En el segundo caso, se toman 2 de los aspectos y se redefinen en sus peores valores posibles, con el fin de evaluar la calidad de los datos de manera similar al primer caso. A su vez, se decidió comparar tiempos de respuesta del modelo para poder evaluar su capacidad de rendimiento frente a las variables. Los resultados obtenidos se presentan a través del uso de gráficas comparativas entre ambos casos, para los valores resultantes de la calidad evaluada, además de usar tablas para las pruebas estadísticas tales como la desviación y el error estándar.

4. Modelo propuesto de lógica difusa

El uso de la lógica difusa en este modelo se basa en la importancia del cambio del paradigma con relación a los resultados matemáticos, expuesto en estudios durante los últimos años. Los resultados

de la medición en los modelos tradicionales pueden presentar ambigüedades, dado que dicho resultado puede ser un valor parcialmente bueno o parcialmente malo, parcialmente cierto o parcialmente falso [18].

Por su parte, la lógica difusa, considerando que utiliza reglas que definen la calidad en segmentos, y resultados que pueden ser representados en más de dos maneras o variables, puede definir adecuadamente una mejor aproximación de la evaluación de calidad de datos.

Como razón adicional, que sustenta el uso de la lógica difusa dentro del modelo, es la existencia de una gran compatibilidad de las variables analizadas frente a los modelos matemáticos usados, unido a una facilidad en su implementación, sin la necesidad de grandes requerimientos tanto para su diseño como su portabilidad, permitiendo así agregar más variables de calidad y más dimensiones a trabajar, para su posterior cálculo de calidad de manera más precisa.

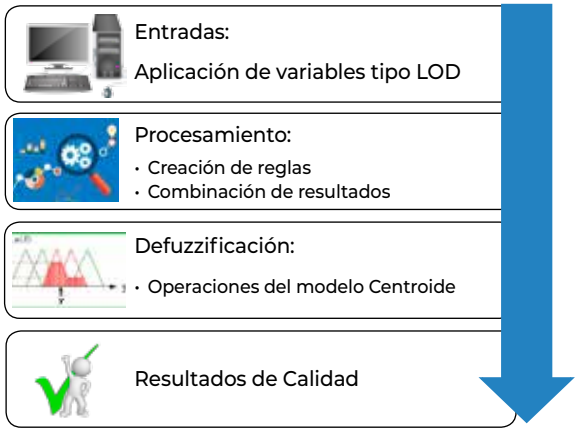

Figura 5. Gráfica de estructura de la lógica difusa.

Fuente: Elaboración propia.

4.1 Modelo propuesto

La Figura 5 representa el funcionamiento del modelo de lógica difusa propuesto frente a las entradas dadas.

El modelo de lógica difusa requiere la evaluación a través de una serie de reglas, de cada una de las cuatro vías de medición de calidad, conocidos como los antecedentes o las entradas del modelo. Posterior a ello, se evalúa las salidas obtenidas, con el fin de determinar si la calidad de los datos es alta, media o baja.

Después de esto, el intervalo porcentual se calcula dentro de los resultados agregados de cada regla de lógica difusa, operando cada antecedente o entrada, y verificando aspectos tales como si todos los antecedentes están presentes.

Es necesario tener en cuenta que la calidad medida entre las reglas clasificadas por dimensión representa la calidad según esa dimensión, mas no la calidad global del conjunto de datos.

Seguidamente, en el proceso de defuzzificación, se calcula el consecuente o la salida del modelo a través del método centroide, también conocido como "centro de gravedad". Esta selección se debe a su avanzada compatibilidad de intensidad en comparación con otro tipo de métodos como el método bisector, así como ser considerado la técnica más utilizada [19].

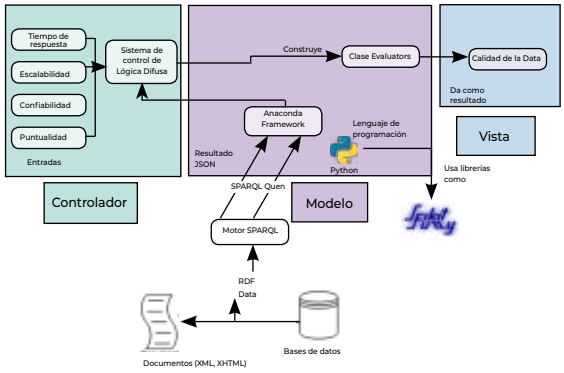

Figura 6. Infraestructura del modelo implementado.

Fuente: Elaboración propia.

Como paso final, el framework logra dar como resultado salidas de cálculos de porcentajes por cada conjunto de datos ingresado.

La Figura 6 presenta el framework utilizado para la construcción y el procedimiento del modelo de lógica difusa.

4.2 Ambiente tecnológico

El análisis de los datos se realiza a través de Anaconda Framework, impulsado por el lenguaje de programación Python, que permite evaluar los datos de una notación de objetos de JavaScript (JSON).

La consulta SPARQL que obtiene información muestra los DataSet asociados, definiendo su relación, punto de origen y datos internos como por ejemplo el autor de DataSet o su fecha de creación [20]. Estos datos se toman como las entradas del framework para la ejecución del modelo. Después de obtener los datos principales, el modelo procede a utilizar controladores y evaluadores definidos para cada antecedente, como parte de la estructura en su programación, cada uno con su función específica.

Los controladores permiten extraer y dividir los datos JSON en segmentos, definidos por las cuatro dimensiones de la calidad de los datos, y los evaluadores ejecutan las operaciones en los datos para prepararlos para su procesamiento en el modelo.

Dentro de las librerías utilizadas para desarrollar los componentes que permitirán evaluar la calidad de los datos, se encuentran:

- Para la simulación de cuatro usuarios y la concurrencia de los mismos, se utilizó una función de multiprocesamiento dentro de la librería Python, denominada *multiprocessing*. Con su uso, se simuló un escenario real en el que cuatro personas con su propio PC intentan acceder a los DataSet, ejecutando la misma consulta al mismo tiempo, con el fin de calcular los tiempos promedio. Este proceso se ejecutó 50 veces más, obteniendo como resultado tiempos de respuesta para uno y cuatro usuarios por separado, para un total de 250 veces.
- Con el uso de la librería Scikit-Fuzzy, derivado de SciPy, se implementó el modelo de lógica difusa dentro del lenguaje de programación Python. A la vez, con la implementación de la librería Requests, se permitió calcular los tiempos de respuesta del SPARQL Endpoint.

4.3 Entradas del modelo

La primera entrada a evaluar es el tiempo de respuesta. Entrada que se puede calcular usando la disponibilidad del DataSet analizado, calculando el tiempo que se demora en responder a las solicitudes. La métrica para definir el tiempo de respuesta se basa en el modelo expuesto en [21]. Este modelo permite analizar la disponibilidad del DataSet con un query genérico, recibiendo o no una respuesta. De acuerdo con esta solicitud, se toma el tiempo que tarda el framework en recibir la respuesta de la URL. Independientemente de la respuesta, si esta existe, el DataSet se considera disponible [21]. Según lo expuesto, se plantea como resultado: el valor de 1 si se recibe una respuesta a la hora de solicitar datos internos de cada DataSet, y el valor de 0 en caso contrario.

La segunda entrada a medir en el modelo es la escalabilidad. La escalabilidad mide la capacidad para responder a múltiples usuarios o peticiones al mismo tiempo [9]. La evaluación de este concepto se toma con base en la concurrencia que soporta el servicio que expone el Endpoint. Como desarrollo metodológico, se plantea la concurrencia de uno y cuatro usuarios hacia cada uno de los DataSet, realizando una simulación en la cual cada uno de estos usuarios realiza una consulta hacia el Endpoint al mismo tiempo, simulando los cuatro usuarios con multihilos.

De manera similar, se realiza el proceso del tiempo de respuesta. Este modelo también se basa en el trabajo expuesto en [21]. Modelo en el cual se maneja la escalabilidad a través del cálculo de tiempos de respuesta de 10, 50 y 100 clientes paralelos para unas 1000 solicitudes aproximadamente. Después de obtener los resultados se aplica la Ecuación 1.

$$Escalabilidad = \frac{Tiempo\ promedio\ 1\ Usuario}{Tiempo\ Promedio\ 4\ usuarios} \quad (1).$$

La relación entre el promedio de tiempo y la generación de la fórmula se basó en la fórmula del modelo de evaluación propuesto en [22]. Trabajo que, dada su similitud y método de evaluación utilizado, permitió verificar matemáticamente en un intervalo, adaptable, este modelo de lógica difusa. Debido a la dificultad en la cual una cantidad numerosa de usuarios intente acceder a los datos al mismo tiempo en un escenario de la vida real, se hizo necesario el uso de simulaciones dentro de un mismo ordenador.

Como se aprecia en la Figura 7, se establece un rango de 0 a 8 de acuerdo con la capacidad de escalabilidad que el DataSet posee, de modo que 0 es el peor caso, y 8 el mejor posible.

La tercera entrada para analizar es la confiabilidad. La confiabilidad permite evaluar si la información y los datos se clasifican como verdaderos y correctos, dentro de las necesidades de cada usuario [10].

La información de procedencia de cada DataSet se configura como una métrica con la que se pueden medir los valores

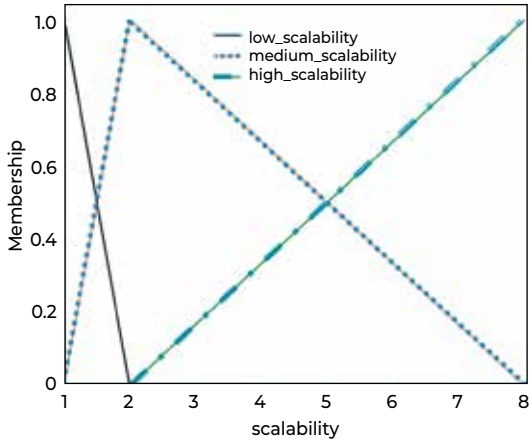

Figura 7. Modelo de lógica difusa para el segundo caso: escalabilidad.

Fuente: Elaboración propia.

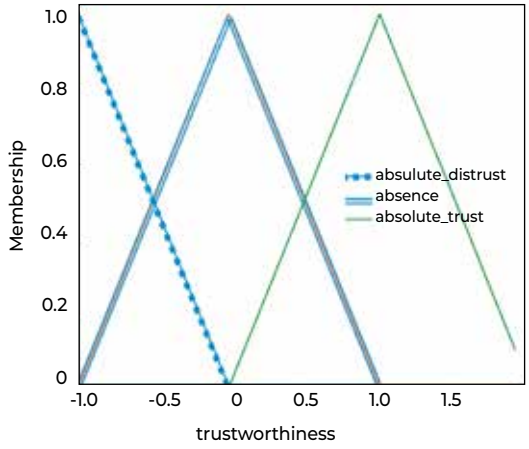

Figura 8. Modelo de lógica difusa para el tercer caso: confiabilidad.

Fuente: Elaboración propia.

de confianza, basada en el trabajo de [22]. En este trabajo, el autor implementa un modelo de medición de tipo cuantificable, para poder medir los niveles de confiabilidad en los datos tipo RDF en un intervalo [-1.1]. Como resultado se considera el valor de -1 como desconfianza absoluta y 1 como confianza absoluta [23, 24].

A su vez, define el concepto de ausencia, que representa el nivel de falta de desconfianza/confianza a la hora de evaluar este concepto. La adaptación de este modelo al modelo de lógica difusa se puede visualizar en la Figura 8.

La cuarta entrada del modelo corresponde a la puntualidad, la cual define el momento en que los DataSet se actualizaron o modificaron por última vez. Para tal efecto, se tiene en cuenta qué tan recientes son los datos y cuán pequeña es la brecha entre la creación y modificación de los mismos.

Para medir la puntualidad, es necesario tener en cuenta conceptos tales como vigencia y volatilidad. En cuanto a la adaptación dentro del modelo, se ha definido la vigencia como la diferencia entre la fecha actual (2017) y la fecha de creación del DataSet [25].

Con relación a la volatilidad, esta se maneja como la diferencia entre la fecha actual (2017) y la fecha de la última modificación del DataSet [22]. Estos dos valores buscan evaluar el concepto conocido como *data must be timely*, expuesto dentro de los principios de LOD [7], concepto que intenta explicar la necesidad de que la data debe mantenerse tan actualizada y reciente como sea posible [10, 26]the Linked Data (LD, y depende de su estructura para saber si la información es realmente relevante y actualizada.

Considerando que la mayoría de los DataSet fueron creados en un tiempo ya relativamente largo (desde 2007 y

en crecimiento) [3], se decidió utilizar el valor de los años como medida de diferencia. A su vez, la muestra de DataSet tomada abarca DataSet relacionados con información no atada a eventos cercanos, y por lo tanto, su prioridad de sensibilidad temporal es baja.

Con los valores de vigencia y volatilidad definidos, la fórmula planteada para la medición de la puntualidad está establecida en la Ecuación 2. El resultado se representa en un intervalo [0.1] en el cual el valor de 0 representa incertidumbre y el valor de 1 certeza [23].

$$Puntualidad = 1 - \left(\frac{Fecha\ actual - Fecha\ creación}{Fecha\ actual - Fecha\ modificación}\right) \quad (2)$$

Dada la similitud de respuestas en cuanto a las figuras, se han agrupado dos de las cuatro variables de calidad: tiempo de respuesta y puntualidad, dentro de un intervalo entre 0 y 1, visualizado en la Figura 9, de acuerdo con los parámetros dados por cada variable.

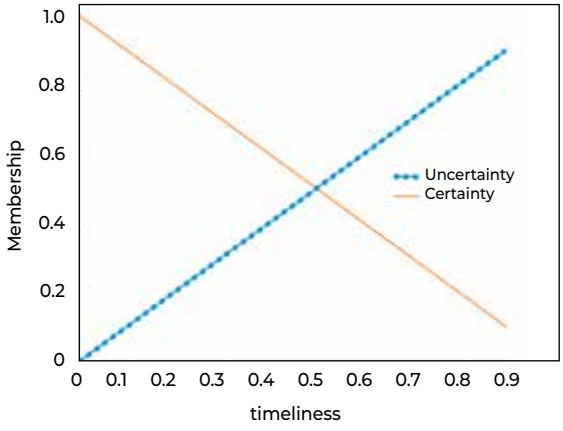

Figura 9. Modelo de lógica difusa para el primer y cuarto caso: tiempo de respuesta y puntualidad.

Fuente: Elaboración propia.

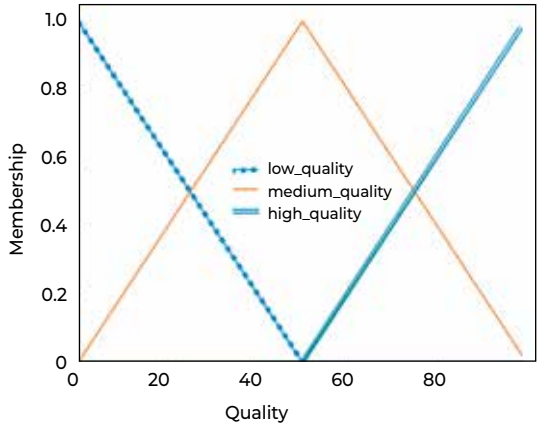

Figura 10. Modelo de lógica difusa para la salida: calidad.

Fuente: Elaboración propia.

4.4 Salidas del modelo

La salida del modelo se representa en un intervalo del 0% y del 100%. En este intervalo, se definen tres rangos que definen la calidad como baja, media o alta, representando los valores como mala, regular y de buena calidad. La Figura 10 representa esta valoración.

4.5 Reglas del modelo

Cada caso del modelo de lógica difusa tiene tres reglas específicas, en que cada una define un caso posible de calidad dentro de cada dimensión. Es necesario aclarar que el valor de salida llamado calidad dentro de cada regla para cada DataSet es según la dimensión en la cual está clasificada la regla, mas no corresponde a la calidad global.

De acuerdo con las Figuras 7, 9 y 10, las reglas para la dimensión de accesibilidad son:

- Si (tiempo de respuesta es Baja) y (Escalabilidad es Alta) Entonces (Calidad es Alta)

- Si (tiempo de respuesta es Baja) y (Escalabilidad es Media) Entonces (Calidad es Media)
- Si (tiempo de respuesta es Alta) y (Escalabilidad es Baja) Entonces (Calidad es Baja)

Es necesario recalcar que la relación que posee el tiempo de respuesta y la escalabilidad es principalmente el rendimiento que ambas variables especifican a la hora de evaluar el DataSet. La escalabilidad mide de manera sucesiva el tiempo de respuesta de cada DataSet, por lo que el resultado del tiempo de respuesta va a afectar de manera directa el resultado de la escalabilidad.

De acuerdo con las Figuras 8, 9 y 10, las reglas para la dimensión de confianza son:

- Si (Confiabilidad es Confianza) y (Puntualidad es Confianza) Entonces (Calidad es Alta)
- Si (Confiabilidad es Confianza) y (Puntualidad es Incertidumbre) Entonces (Calidad es Media)
- Si (Confiabilidad es Desconfianza) y (Puntualidad es Incertidumbre) Entonces (Calidad es Baja)

Al interior del modelo de lógica difusa, las entradas se actualizan, procediendo a comparar los resultados con las 6 reglas previamente definidas, y permiten obtener un resultado de salida por DataSet.

La relación entre la confiabilidad y puntualidad responde al análisis interno de cada DataSet, es decir, tomar en cuenta la información de consulta del SPARQL Endpoint y su procesamiento, para transformarlas en antecedentes para el modelo de lógica difusa. Tanto la confiabilidad como la puntualidad toman datos internos del DataSet como fecha de creación, autores, DataSet de origen, entre otros, los cuales definen de manera interna la dimensión de confianza para la obtención de la calidad de los datos.

Es necesario tener en cuenta que cada una de las reglas se enmarca en los rangos de cada dimensión, es decir, que el modelo toma cada dimensión y las evalúa según los antecedentes, sin necesidad de ignorar criterios. Como ejemplo, si la regla 1 se cumple, quiere decir que según la dimensión de accesibilidad la calidad de un DataSet es alta. Sin embargo, puede existir la posibilidad de que, según la dimensión de confianza, la calidad es baja, por lo que su puntaje tendrá en cuenta todos los criterios de cada dimensión, tanto los incluidos como los que se pueden agregar como trabajo futuro.

5. Resultados obtenidos

En primera instancia, se llevó a cabo el consumo de los 100 DataSet a analizar, descargados en un archivo tipo JSON.

Posteriormente se obtuvieron los datos necesarios para operar el modelo de lógica difusa. Para ejemplificar los resultados, solo se muestran en esta sección cuatro DataSet de ejemplo obtenidos del SPARQL Endpoint del DataSet

Tabla 1. Ejemplos de resultados de tiempos de respuesta

Nombre de DataSet	Tiempo de respuesta calculado	Valor de tiempo de respuesta
Czech municipalities	48.1128183467	1
Job applicants in regions of Czech Republic	47.8042141867	1
Institutional research plans	49.1565653333	1
R&D Programmes	47.9116424533	1

Fuente: Elaboración propia.

opendata.cz, siendo estos Czech municipalities, Job applicants in regions of Czech Republic, Institutional research plans y R&D Programmes.

Para el caso de la métrica de tiempo de respuesta, el promedio obtenido fue de 43.73 m, calculado a través de las peticiones SPARQL hacia el DataSet. Este procedimiento se ejecutó aproximadamente 50 veces más y luego se obtuvo el promedio de esos tiempos, en el que dio el resultado de 48.32 m. Con estos resultados se pudo comprobar que cada uno de los DataSet analizados dieron resultados dentro del JSON según la petición dada, es decir, no se obtuvieron valores nulos o no respuesta por parte de los DataSet, por lo que el valor de tiempo de respuesta para cada uno de estos DataSet corresponde al valor de 1.

Para el caso de escalabilidad, el procedimiento seguido consistió en realizar una solicitud SPARQL para extraer datos internos de cada DataSet, usados dentro de los casos de confiabilidad y puntualidad.

Como paso siguiente, se añadieron cada uno de los cuatro tiempos de usuario, obteniendo datos de un único grupo de tiempo de respuesta, que luego se compararon con los datos de un usuario, utilizando el tiempo medio de ambos grupos. Con esto se aplica la fórmula definida, cuyo resultado de escalabilidad promedio es igual a 0.900363599953. Como referencia, este valor está entre un intervalo de [0.1]. Los resultados de escalabilidad obtenidos se visualizan en la Tabla 2.

Para el caso de la confiabilidad, los datos internos de cada DataSet son consultados. Para tales efectos, solo se tuvieron en cuenta los autores y contribuyentes de cada DataSet. La forma natural de

Tabla 2. Ejemplos de resultados de escalabilidad

Nombre de DataSet	Resultado de escalabilidad calculado
Czech municipalities	0.917890563599
Job applicants in regions of Czech Republic	0.891315467557
Institutional research plans	0.937077953964
R&D Programmes	0.921803070866

Fuente: Elaboración propia.

medir este caso es tomar los valores de autor y colaboradores de cada DataSet, y evaluar si estos datos están presentes (sin valor nulo). Sin embargo, debido al estado empírico del estudio para los 100 DataSet de ejemplo, la confiabilidad se evaluó estáticamente como 1, considerando que todos los DataSet contienen autores y contribuyentes verificados por Datahub, plataforma diseñada por la Open Knowledge Foundation. Dado que algunos DataSet no siempre pueden tener una licencia explícita y estar libres desde un punto de vista de la web [20], la confiabilidad puede tomar valores diferentes a esta simulación, razón por la cual se define el valor 0 como ausencia, es decir, no se puede definir correctamente de confianza o no.

Para el caso de puntualidad, se tomó en cuenta la fecha de cada DataSet, en cuanto a su creación y última modificación. Con el fin de especificar y generalizar el cálculo de este factor, solo se tuvo en cuenta el año actual de esta investigación (2017), para comparar ambas fechas. Es importante anotar que la fórmula propuesta puede trabajar con días y meses añadidos, incluso horas [10].

Tabla 3. Ejemplos de resultados de puntualidad

Nombre de DataSet	Resultado de puntualidad
Czech municipalities	0.8
Job applicants in regions of Czech Republic	1.0
Institutional research plans	1.0
R&D Programmes	1.0

Fuente: Elaboración propia.

Como resultado, se obtiene un intervalo entre 0 y 1 por DataSet, resultados visualizados en la Tabla 3.

Como modelo de salida, cada una de las entradas fue introducida en el modelo de lógica difusa, posteriormente procesada con las reglas de lógica difusa, y sus respectivos resultados visualizados como un porcentaje de salida. Este porcentaje representa el valor aproximado de la calidad de los datos de los DataSet. Este proceso se realizó en dos momentos para representar cada caso. El primer caso toma en cuenta todas las 5 entradas para calcular la calidad de los datos, y para comparar los resultados el segundo caso toma 2 de estas entradas, tiempo de respuesta y confiabilidad, y aplica el peor valor posible a cada uno de ellos. El tiempo de respuesta toma como valor 0, y la confiabilidad asume como su peor puntuación posible el valor de -1. Con estos valores predeterminados, y teniendo en cuenta que el resto de entradas son iguales al caso 1, el modelo se ejecuta de nuevo, arrojando los valores de salida correspondientes al segundo caso. Estos resultados pueden ser visualizados en las Tablas 4 y 5.

Tabla 4. Calidad del DataSet: primer caso

Nombre de DataSet	Porcentaje de calidad de la data
Czech municipalities	57.40730834
Job applicants in regions of Czech Republic	57.22276088
Institutional research plans	57.22276088
R&D Programmes	57.22276088

Fuente: Elaboración propia.

Tabla 5. Calidad del DataSet: segundo caso

Nombre de DataSet	Porcentaje de calidad de la data
Czech municipalities	17.22222222
Job applicants in regions of Czech Republic	16.81818182
Institutional research plans	16.81818182
R&D Programmes	16.81818182

Fuente: Elaboración propia.

Tabla 6. Desviación estándar y error estándar de entradas/salidas

	Desviación estándar	Error estándar
Tiempo de respuesta	0.435425506	0.061578466
Escalabilidad	0.05859876	0.008287116
Calidad caso 1	0.399888771	0.056552812
Calidad caso 2	0.63081278	0.089210399

Fuente: Elaboración propia.

5.1 Análisis de resultados

Teniendo en cuenta el plan de análisis propuesto, se tomaron dos perspectivas principales, las cuales abarcan tanto:

El desempeño de la aplicación, el cual permite inferir la capacidad del framework para obtener métricas de calidad de acuerdo con un número considerable de variables de calidad.

El análisis de los datos obtenidos a través de la inferencia estadística, soportado en referentes como la desviación estándar y el error calculado dentro de las operaciones, el cual permite diferenciar la capacidad de la lógica difusa en el análisis, de su desempeño dentro de cada variable de calidad por separado.

Como primer paso, se estudió la desviación estándar y el error de los resultados del modelo, tanto para los casos de calidad como para el tiempo de respuesta y escalabilidad, debido al uso de sus valores medios para ejecutar el modelo.

La desviación estándar de ambos casos de calidad es bastante pequeña (Tabla 6), lo que permite definir la alta precisión de los datos calculados, debido al bajo rango de dispersión entre la media y los valores obtenidos [27].

Es pertinente resaltar que el caso 2 tiene una desviación estándar más grande. Lo anterior se debe a que toma solo la puntualidad y el tiempo de respuesta como factores que fluctúan el resultado, pero todas las demás variables continúan con los mismos valores que en el caso 1, así que una correlación entre ambos existe y tiene resultados de desviación estándar similares, por lo tanto, definen una alta precisión en los resultados de salida del modelo de la lógica difusa.

Como método de dispersión, con la ayuda de la desviación estándar se puede comprobar la capacidad del modelo para obtener resultados de manera precisa, ya que, por ser resultados tan bajos, estos concluyen que las variables obtienen pocos valores extremos. Esto ocurre debido a que todos los DataSet son obtenidos dentro de un mismo SPARQL Endpoint, por lo cual el modelo puede poseer un ligero

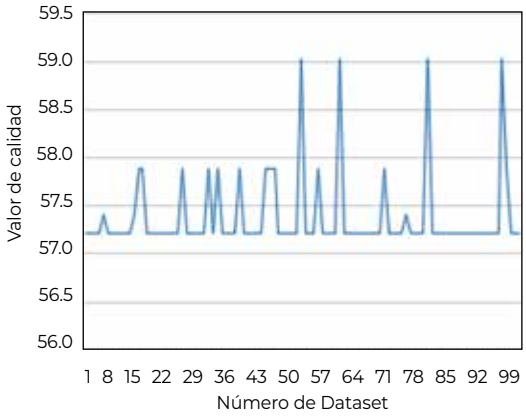

Figura 11. Calidad de los resultados: primer caso.
Fuente: Elaboración propia.

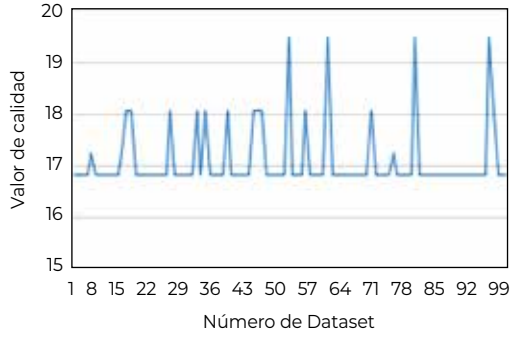

Figura 12. Calidad de los resultados: segundo caso.
Fuente: Elaboración propia.

aumento del resultado de desviación estándar a la hora de analizar varios Endpoints al mismo tiempo.

A la vez, la adición del error estándar permite confirmar el hecho de que la muestra de los datos tomados no posee una alta discrepancia. Específicamente con el resultado de la calidad en el caso 1, el cual contiene todas las cuatro variables de calidad, situación que permite identificar que esta discrepancia disminuye, y así permite reconocer que el modelo puede soportar una mayor cantidad de variables de calidad, sin que fluctúe mucho la discrepancia hallada.

En relación con los resultados de la calidad para ambos casos, se observa en las Figuras 11 y 12 que el primer caso tiene un porcentaje más alto (promedio de 57.37966662 %) que el segundo caso (promedio de 17.08383838 %). Lo anterior debido a tener una mejor puntuación en dos de los valores de dimensión; sin embargo, se observa que en ambos casos se tienen picos que significan mayor calidad. La razón principal de este fenómeno es que la puntualidad y el tiempo de respuesta fluctúan los resultados.

En un escenario real, los picos están representados por valores que no poseen una alta puntualidad, probada por no tener el mejor valor posible en el intervalo 0.1. Esto hace que los valores cambien, y a su vez los valores de las reglas de la lógica difusa. Lo anterior considerando que la estructura del modelo consta de tres reglas construidas por cada dimensión, lo que afecta el resultado de salida, debido al método centroide usado dentro del modelo de la lógica difusa.

Con el fin de verificar el rendimiento del modelo de lógica difusa, junto con la calidad de salida, se obtuvo un tiempo de respuesta entre el inicio del procesamiento del modelo y el final de este, para cada DataSet.

En la Tabla 7, los tiempos de respuesta para el primer caso (promedio de 0.016356885 s) son muy similares a

Tabla 7. Ejemplos de tiempo de procesamiento de modelos de lógica difusa

Nombre de DataSet	Tiempos de calidad caso 1	Tiempos de calidad caso 2
Czech municipalities	0.0160699390508s	0.0158510595177s
Job applicants in regions of Czech Republic	0.0166211912081s	0.0138444504652s
Institutional research plans	0.0166536178056s	0.0154051938023s
R&D Programmes	0.0165887646106s	0.0158510595177s

Fuente: Elaboración propia.

los del segundo caso (promedio de 0.015976563 s). Este fenómeno puede definir un alto rendimiento del modelo de lógica difusa. Esto permite establecer un buen comportamiento del modelo, con n variables, independiente de sus valores o resultados.

Por otro lado, se observa que el desempeño de este modelo es bastante bueno, considerando que los tiempos de ejecución de las operaciones del modelo fueron pequeños, permitiendo mayores pruebas en las que tanto los números de DataSet como los factores de calidad pueden ser expandidos. La lógica difusa permite la precisa evaluación de un número mayor de factores de calidad, en términos de recursos e interoperabilidad.

Con base en este enfoque, los resultados porcentuales obtenidos permiten identificar un resultado de calidad media, como se aprecia en los cuatro DataSet de ejemplo. Principalmente con estos DataSet, su calidad a través de las dimensiones de accesibilidad y de confianza son buenas en el caso 1, en comparación con el caso 2, el cual obtuvo un porcentaje de puntuación mucho menor.

Estas comparaciones permiten apreciar la funcionalidad de las reglas de lógica difusa, el modelo centroide usado para la defuzzificación y aspectos como las fluctuaciones entre puntajes. De igual forma, permite confirmar uno de los objetivos principales del modelo desarrollado, consistente en la capacidad de analizar gran cantidad de variables de calidad con valores de rendimiento y procesamiento aceptables.

6. Discusiones

Mientras que Luzzu utiliza una normalización de resultados de tipo logarítmica para obtener resultados más legibles, el modelo propuesto decidió no normalizar los resultados, dado que, a pesar de que son más confusos, de acuerdo con sus decimales relativamente extensos, el modelo centroide para la lógica difusa puede ser afectado con la falta de estos datos matemáticos, perdiendo parte de su objetivo, el cual corresponde a obtener resultados de calidad más precisos que los modelos tradicionales.

Similarmente al modelo propuesto, el framework Luzzu también logra

apreciar las fallas que los SPARQL Endpoints poseen a la hora de ejecutar gran cantidad de DataSet, permitiendo un tipo de nueva problemática en cuanto a cómo lograr visualizar o procesar gran cantidad de DataSet relacionados dentro de nubes de tipo LOD. A pesar de que los modelados de ambos framework son distintos, se puede apreciar que su crecimiento en cuanto a procesamiento y velocidad, basados en el tiempo de ejecución de los modelos, son de tipo lineal. Sin embargo, el modelado del framework Luzzu se expande en mayor manera al modelado de lógica difusa cuando existe una gran cantidad de variables dentro del procesador de SPARQL Endpoint, uno de los tres procesadores usados y el más inestable dentro de ellos.

Sin embargo, a pesar de utilizar fórmulas matemáticas similares para el cálculo de los datos [6, 10], Luzzu utiliza un modelamiento clásico para el procesamiento de los datos de una manera porcentual, en los que solo puede definir de manera dual la calidad, como por ejemplo verdadero o falso, sin las ventajas que ofrece la lógica difusa, en la generación de resultados más precisos, como el establecimiento de reglas de acuerdo con las dimensiones, lo cual se configura como valor agregado de esta investigación.

Es prudente aclarar que el modelo de lógica difusa propuesto en esta investigación aún requiere trabajo en cuanto a su operatividad y resultados, en comparación con Luzzu, para permitir visualizar al cliente de manera más puntual el valor de la calidad del DataSet, a la vez de permitir resultados de dimensiones y variables como seguridad, relevancia o consistencia, previamente estudiadas para su uso dentro de este modelo de lógica difusa, pero descartados por su complejidad a la hora de su implementación.

Sin embargo, en comparación con Luzzu, el modelo de lógica difusa posee una contribución principal, orientada a mejorar la capacidad de tiempo computacional y procesamiento dinámico, de acuerdo con una gran cantidad de variables, lo cual afirma la capacidad del modelo para poder procesar una mayor cantidad de variables y obtener un resultado de calidad global más preciso.

7. Conclusiones

La calidad de los datos vinculados cada vez más se convierte en un factor crítico, dada la proliferación de proveedores de datos, quienes bien por falta de experiencia, bien por desconocimiento, no siguen las buenas prácticas sugeridas por el esquema de publicación de datos.

Con la disposición de repositorios de datos tales como Datahub, que han facilitado el proceso de publicación de recursos, LOD ha venido experimentando un crecimiento exponencial en los últimos años. Este tipo de crecimiento se puede visualizar en diferentes esfuerzos, como por ejemplo Europeana - biblioteca digital, sobre la cual se han realizado estudios acerca de su uso para el desarrollo de objetos de aprendizaje en áreas de conocimiento específicas, a través de la reutilización de recursos

digitales de acceso abierto [28]. Pero en la otra cara de la moneda, y como lo muestran algunos de los estudios consultados, muchos proveedores de datos, sobre todo aquellos del dominio público o gubernamental, publican sus datos de forma abierta pero no avanzan hacia un proceso de vinculación de forma adecuada.

Para abordar los problemas de calidad suscitados por las deficiencias en los procesos de publicación y vinculación de datos, se han generado estrategias que configuran modelos de calidad de datos, y en algunos de los casos, soportados por herramientas de software.

Muchas de las aproximaciones hacia la evaluación de la calidad de los datos vinculados basan su valoración en modelos matemáticos tradicionales, lo cual puede llegar a restringir la apreciación de la calidad de los datos.

Una aproximación a la valoración de la calidad de los datos basada en lógica difusa es una solución que tiene muchas ventajas, representadas en factores como la capacidad de definir matemáticamente la calidad y enfocarse en ella profundamente, más que los enfoques matemáticos clásicos. Sin embargo, en este caso se evaluó el mismo Endpoint, lo que significa que los aspectos de calidad fueron similares, aparte de hechos como la necesidad de añadir más dimensiones de calidad y, por lo tanto, el uso de más reglas y operaciones.

Con relación a la pregunta de investigación planteada, se ha logrado demostrar que el rendimiento de la aplicación frente a un número elevado de variables de calidad es estable y ágil, por lo que el framework planteado tendría la capacidad de aumentar de manera considerable estas variables para que el estudio de calidad sea aún más preciso.

A su vez, la contribución clave del modelo de lógica difusa es la capacidad del cálculo de modelados porcentuales, con rangos más detallados, los cuales llevan a una aproximación más detallada de la calidad de los datos vinculados, en comparación con otros modelos, como el modelado dual o el modelado logarítmico utilizado por el framework Luzzu. Esta contribución lleva a que el modelado de lógica difusa posea un gran potencial a la hora de calcular este tipo de resultados.

Estas dos pautas juntas logran afirmar que la calidad calculada puede llegar a ser más precisa y exacta, en cuanto a la opinión del usuario que necesite esta data, a la vez que lo provee de mayor información tanto interna como externa dentro de cada DataSet.

Como trabajo futuro se plantea avanzar en el trabajo de expandir este modelo mediante la adición de nuevos aspectos de calidad y nuevas dimensiones durante su clasificación. Por lo que, en resumen, se busca la adición de más variables de calidad para una evaluación más abierta y precisa, a la vez de la adición de nuevas interacciones con el usuario en cuanto a las opiniones que poseen frente a los DataSet a su disposición.

Agradecimientos

Esta investigación se lleva a cabo en el marco de la formación doctoral en ingeniería, en la Universidad Distrital Francisco José de Caldas. De igual forma, la temática planteada se configura como una línea de investigación de Grupo GIIRA.

Referencias

[1] Hu, B., Rodrigues, E., & Viel, E. (2014). Capri. In Proceedings of the 16th International Conference on Information Integration and web-based Applications & Services - iiWAS '14 (pp. 217–223). New York, New York, USA: ACM Press. https://doi.org/10.1145/2684200.2684336

[2] Bonatti, P. A., Hogan, A., Polleres, A., & Sauro, L. (2011). Robust and Scalable Linked Data Reasoning Incorporating Provenance and Trust Annotations. web Semantics: Science. *Services and Agents on the World Wide web,* 9(2), 165–201. En: http://aidanhogan.com/docs/saor_ann_final.pdf

[3] Ermilov, I., Lehmann, J., Martin, M., & Auer, S. (2016). LODStats: *The Data web Census DataSet*. International Semantic web Conference., 38–46. En: http://jens-lehmann.org/files/2016/iswc_lodstats.pdf

[4] Herrera-Cubides, J., Gona-García, P., & Gordillo-Orjuela, K. A. (2017). View of the web of Data. Case Study: Use of Services CKAN. *Ingeniería,* 22 (1), 111-124.

[5] Gordillo, K., Gómez Acosta, A., Gaona-García, P. & Montenegro Marín, C. (2015). Visualizing security principles to access resources based on Linked Open Data: Case study *DBpedia. Information Journal, Vol* 21(1) ,109-122.

[6] Debattista, J., Auer, Sö., & Lange, C. (2016). Luzzu—A Methodology and Framework for Linked Data Quality Assessment. Journal of Data and Information Quality, 8(1), 1–32. En: https://doi.org/10.1145/2992786

[7] Berners-Lee, T. (2016). 5 Stars principles of Linked Open Data. Consultado August 8, 2017, en: https://joinup.ec.europa.eu/sites/default/files/ckeditor_files/files/W3C04.pdf

[8] Hellmann, M. (2001). Fuzzy Logic Introduction. Universite de Rennes, 1–9. En: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.85.9757&rep=rep1&type=pdf

[9] Zaveri, A., Rula, A., Maurino, A., Pietrobon, R., Lehmann, J., & Auer, S. (2012). Quality Assessment Methodologies for Linked Open Data A Systematic Literature Review and Conceptual Framework. Semantic web, 7(1), 63-93. En: http://www.semantic-web-journal.net/system/files/swj414.pdf

[10] Zaveri, A. (2015). *Linked Data Quality Assessment and its Application to Societal Progress Measurement*. University of Leipzig. En: https://core.ac.uk/download/pdf/35206278.pdf

[11] Naumann, F. (2002). *Quality-driven query answering for integrated*

information systems. (G. Goos, J. Hartmanis, & J. van Leeuwen, Eds.) (1st ed.). Berlín: Springer. En: //dl.acm.org/citation.cfm?id=1791545

[12] Portal BCN. Datos Abiertos Enlazados. Biblioteca del Congreso Nacional de Chile. [Online]. Available: http://datos.bcn.cl/es/informacion/que-es

[13] Batini, C., & Scannapieca, M. (2006). *Data quality: concepts, methodologies and techniques*. Springer. En: https://dl.acm.org/citation.cfm?id=1177291

[14] González Morcillo, C. (2011). *Lógica difusa: Una Introducción Práctica.* En: http://www.esi.uclm.es/www/cglez/downloads/docencia/2011_Softcomputing/LogicaDifusa.pdf.

[15] Herrera-Cubides, J. et al. A Fuzzy Logic System to Evaluate Levels of Trust on Linked Open Data Resources. (2018). *Revista Facultad de Ingeniería*. 86, 40-53. Available at: <http://aprendeenlinea.udea.edu.co/revistas/index.php/ingenieria/article/view/328937>.

[16] Lewis, D. J., & Martin, T. P. (2015). Managing Vagueness with Fuzzy in Hierarchical Big Data1. *Procedia Computer Science,* 53, 19–28. https://doi.org/10.1016/j.procs.2015.07.275

[17] Nečaský, M., Klímek, J., Chlapek, D., Kučera, J., Mynarz, J., & Svátek, V. (2015). *OpenData.cz*. Consultado September 5, 2017, En: https://opendata.cz/

[18] Stella-vagaska Alena, H. (2012). Application of Fuzzy Principles in Evaluating Quality of Manufacturing Process. *WSEAS TRANSACTIONS ON POWER SYSTEMS,* 7(2), 1–10. En: http://www.tuke.sk/fvtpo

[19] Pradeep, A., Thomas, J., & Joseph, S. (2015). Performance Assessment for Students using Different Defuzzification Techniques. IJIRST – International Journal for Innovative Research in Science & Technology|, 2(6). En: www.ijirst.org

[20] Auer, S., Ermilov, I., Lehmann, J., & Martín, M. (2016). *LODStats - 9960 datasets*. Consultado August 14, 2017, En: http://stats.lod2.eu/

[21] Lehmann, J., Vandenbussche, P.-Y., Umbrich, J., Matteis, L., Hogan, A., & Buil-Aranda, C. (2017). SPARQLES: Monitoring Public SPARQL Endpoints. *Semantic web,* 1–17. En: http://www.semantic-web-journal.net/system/files/swj1381.pdf

[22] Hartig, O., & Zhao, J. (2009). *Using web Data Provenance for Quality Assessment*. Proceedings of the First International Conference on Semantic web in Provenance Management, 526, 29–34. En: http://ceur-ws.org/Vol-526/paper_1.pdf

[23] Hartig, O. (2008). Trustworthiness of Data on the web. N STI Berlin and CSW PhD Workshop, 1–5. En: https://pdfs.semanticscholar.org/dd04/829d811b2284f01be7fa421b4c2141d32256.pdf

[24] Feeney, K. C., O'Sullivan, D., Tai, W., & Brennan, R. (2014). Improving Curated web-Data Quality with Structured Harvesting and Assess-

ment. *International Journal on Semantic web and Information Systems,* 10(2), 35–62. https://doi.org/10.4018/ijswis.2014040103

[25] Fürber, C., & Hepp, M. (2011). SWIQA – A Semantic web information quality assessment framework SWIQA – a Semantic web information quality assessment framework. *Association for Information Systems AIS Electronic Library (AISeL),* (76), 1–13. En: http://aisel.aisnet.org/ecis2011

[26] Eslake, S. (2006). The importance of accurate, reliable and timely data. Discussion Paper prepared for a Group of "Eminent Australians." En: https://www.saul-eslake.com/wp-content/uploads/The-Importance-of-Data.pdf

[27] DataStar, I. (2013). *How to Interpret Standard Deviation and Standard Error in Survey Research*. Consultado September 7, 2017, En: www.surveystar.com

[28] Gaona-García, P., Sánchez-Alonso, A., & Fermoso García, A. (2017). Visual analytics of Europeana digital library for reuse in learning environments: A premier systematic study. *Online Information Review, Vol. 41 (6), 840-859,* https://doi.org/10.1108/OIR-04-2016-0114