UNIVERSIDAD TÉCNICA PARTICULAR DE LOJA

La Universidad Católica de Loja

ÁREA TÉCNICA

INGENIERO EN SISTEMAS INFORMÁTICOS Y COMPUTACIÓN

TRABAJO DE TITULACIÓN

Método de enriquecimiento de Grafos de Conocimiento de indicadores de ODS usando un enfoque de inferencia de entidades semánticas equivalentes desde fuentes de datos abiertos

Autor: Pinto Orellana, Diego Fernando

Director: Piedra Pullaguari, Nelson Oswaldo

LOJA - ECUADOR

2021

2021

Aprobación del director del Trabajo de Titulación

Loja, 01, de febrero, 2021

Magíster
Fernanda Maricela Soto Guerrero
Coordinadora de carrera de Sistemas Informáticos y Computación
Ciudad. -

De mi consideración:

El presente trabajo de titulación denominado: Método de enriquecimiento de Grafos de Conocimiento de indicadores de ODS usando un enfoque de inferencia de entidades semánticas equivalentes desde fuentes de datos abiertos, realizado por Diego Fernando Pinto Orellana, ha sido orientado y revisado durante su ejecución, por cuanto se aprueba la presentación de este. Así mismo, doy fe que dicho trabajo de titulación ha sido revisado por la herramienta antiplagio institucional.

Particular que comunico para los fines pertinentes.

Atentamente,

Firma:.....
Nelson Oswaldo Piedra Pullaguari.
C.I: 1102809462

Declaración de autoría y cesión de derechos

"Yo, Diego Fernando Pinto Orellana, declaro y acepto en forma expresa lo siguiente:

- Ser autor(a) del Trabajo de Titulación denominado: Nombre del trabajo, de la Titulación Método de enriquecimiento de Grafos de Conocimiento de indicadores de ODS usando un enfoque de inferencia de entidades semánticas equivalentes desde fuentes de datos abiertos de la Titulación Sistemas Informáticos y Computación, específicamente de los contenidos comprendidos en: Capítulo 1. Estado del Arte, Capítulo 2. Propuesta, Capítulo 3. Desarrollo de la propuesta, Capitulo 4. Pruebas y Resultado, siendo Nelson Oswaldo Piedra Pullaguari; y, en tal virtud, eximo expresamente a la Universidad Técnica Particular de Loja y a sus representantes legales de posibles reclamos o acciones judiciales o administrativas, en relación a la propiedad intelectual. Además, ratifico que las ideas, conceptos, procedimientos y resultados vertidos en el presente trabajo investigativo son de mi exclusiva responsabilidad.
- Que mi obra, producto de mis actividades académicas y de investigación, forma parte del patrimonio de la Universidad Técnica Particular de Loja, de conformidad con el artículo 20, literal j), de la Ley Orgánica de Educación Superior; y, artículo 91 del Estatuto Orgánico de la UTPL, que establece: "Forman parte del patrimonio de la Universidad la propiedad intelectual de investigaciones, trabajos científicos o técnicos y tesis de grado que se realicen a través, o con el apoyo financiero, académico o institucional (operativo) de la Universidad".
- Autorizo a la Universidad Técnica Particular de Loja para que pueda hacer uso de mi obra con fines netamente académicos, ya sea de forma impresa, digital y/o electrónica o por cualquier medio conocido o por conocerse, sirviendo el presente instrumento como la fe de mi completo consentimiento; y, para que sea ingresada al Sistema Nacional de Información de la Educación Superior del Ecuador para su difusión pública, en cumplimiento del artículo 144 de la Ley Orgánica de Educación Superior.

Firma: .....

Autor: Diego Fernando Pinto Orellana

C.I.: 1105642332

Dedicatoria

Este trabajo lo dedico:

A Dios por haberme dado la fortaleza en los momentos difíciles y haberme permitido llegar hasta este momento tan importante de mi formación profesional.

A mis padres, quienes me han acompañado durante todo este transcurso de mi vida, por sus consejos y apoyo incondicional.

A mi hermana, quien siempre me ha apoyado y motivado para seguir adelante para logar este gran sueño.

A mis abuelitas, tíos, primos y a toda mi familia, por haberme brindado su apoyo incondicional y por compartir conmigo buenos y malos momentos.

Diego Fernando Pinto Orellana

Agradecimiento

A Dios por guiar mi camino y haberme dado fuerzas para saber llevar cada obstáculo y las dificultades a lo largo de mi formación.

A Nelson Piedra, por brindarme las oportunidades que he tenido en varias etapas de mi formación, por la confianza en mi capacidad para desarrollar cada una de las ideas que tiene en mente, por dedicar su tiempo para revisar y corregir mis errores y enseñarme como tener una mejor versión de mi trabajo.

Agradezco la confianza y el apoyo de mis padres para alcanzar mis metas. A mi hermana, abuelitas, tíos y a toda mi familia por haberme brindado muchos momentos de alegría.

A todos mis docentes quienes fueron parte de mi formación académica, por compartir su conocimiento y experiencias para mi formación profesional.

A todos mis amigos/as, que siempre me animaron a ser mejor cada día y se alegran de cada uno de mis logros.

Índice de Contenido

Índice de Tablas

Índice de Figuras

Resumen

El objetivo de este trabajo es crear un método para enriquecer un grafo de conocimiento relacionado con los Objetivos de Desarrollo Sostenible (ODS). Los ODS buscan garantizar que las personas gocen de paz y prosperidad. Este trabajo describe una propuesta para el enriquecimiento de grafos de conocimiento relacionados con datos ODS. El contexto de fuentes de información y ODS se aborda a través de la gestión de datos heterogéneos y distribuidos a través del modelo de datos orientados a grafos. Los grafos aportan mayor nivel de abstracción que facilita la descripción semántica de información, en términos de nodos y relaciones, tanto del modelo de datos, como de las interacciones que se producen entre componentes de conjuntos de datos locales como distribuidos.

El propósito central del trabajo consiste en incorporar anotaciones a un grafo de conocimiento semántico con información relacionada con ODS, de manera que enriquezca la descripción semántica de los nodos con componentes de grafos semánticos externos. Finalmente, la propuesta desarrollada para este trabajo implementa una aplicación web piloto que permite interactuar y acceder a los recursos enriquecidos.

Palabras claves: Objetivos de Desarrollo Sostenible, Enriquecimiento de grafos, Web Semántica.

Abstract

The objective of this work is to create a method to enrich a knowledge graph related to the Sustainable Development Goals (SDGs). The SDGs aim to ensure that people enjoy peace and prosperity. This paper describes a proposal for the enrichment of knowledge networks related to SDGs data. The context of information sources and SDGs is addressed through the management of heterogeneous and distributed data via the graph-oriented data model. Networks provide a higher level of abstraction that facilitates the semantic description of information, in terms of nodes and relationships, both of the data model and of the interactions that occur between components of local and distributed datasets.

The central purpose of the work consists of incorporating annotations to a semantic knowledge graph with information related to SDGs, in a way that enriches the semantic description of the nodes with components of external semantic graphs. Finally, the proposal developed for this work implements a pilot web application that allows interacting and accessing the enriched resources.

Keywords: Sustainable Development Goals, Graphics Enrichment, Semantic Web.

Introducción

Los portales de datos de acceso público tienen como misión recopilar y compartir datos de múltiples fuentes y de una variedad de dominios de información. Los portales de datos abiertos son claves en acciones que contribuyan a la transparencia, el mejoramiento de la gestión participativa, la innovación, la promoción de la colaboración ciudadana y académica, entre otras características importantes. Existen avances importantes relacionados con la apertura y el reuso: licenciamiento abierto, uso de estándares abiertos, mejoras en las capacidades de integración e interoperabilidad de datos, buenas prácticas para mejorar la calidad de datos, la accesibilidad, mejorar la colaboración, el uso y reutilización de información abierta.

Los ODS comprenden 17 objetivos y 169 metas, y un conjunto preliminar de 330 indicadores, propuestos en marzo de 2015 y con un plazo para que se cumpla la resolución hasta el año 2030. Los portales de datos relacionados con los Objetivos de Desarrollo Sostenible (ODS) plantea retos en torno al acceso e interoperabilidad de datos a través de los diferentes mecanismos e interfaces de acceso a la información.

Actualmente no existe un proceso único que permita conectar conjuntos de datos abiertos relacionados con ODS con otras fuentes externas. Este trabajo busca crear un método de enriquecimiento de grafos de tal manera poder observar y determinar la calidad del grafo de conocimiento construido a partir de conjuntos de datos abiertos relacionado con los ODS, y que se publican a través de la plataforma CKAN. El trabajo se enfoca en desarrollar un método para enriquecer un grafo de conocimiento semántico, e implementa una interfaz de navegación Web sobre el grafo que forma parte del alcance de este trabajo. El sistema de enriquecimiento de grafos de conocimiento proporciona una forma de generar candidatos para un nodo y que puede ser desplegado en el grafo existente, y también una nueva forma de conectar nodos a través de equivalencia semántica inferida a través del enfoque propuesto.

El resultado del enriquecimiento de datos de este trabajo se mostrará a través de una aplicación Web. Se ha validado que la aplicación sea de fácil uso y acceso.

Este trabajo inició con una revisión sistemática relacionada con datos, grafos de conocimiento semántico, interoperabilidad y enriquecimiento. En la revisión sistemática de literatura se obtuvo información que permitió construir una forma para convertir recursos abiertos de múltiples fuentes en información interoperable y esta pueda ser enriquecida con más información. Entonces mediante un correcto análisis y procesamiento de datos se puede mejorar la accesibilidad, uso y reutilización.

Capítulo uno

Estado del Arte

1.1 Revisión Sistemática

El crecimiento intensivo de la información heterogénea, que se encuentra en la Web, plantea la necesidad de contar con maneras efectivas para mejorar la capacidad de descubrimiento, re-uso, analítica, personalización, y en general de obtención de valor.

A través de la Web, se acceden a diferentes plataformas de información abierta. Algunas de estas plataformas están destinadas a la publicación de datos abiertos (en inglés, Open Data - OD). Estas plataformas están relacionadas con la recolección y la distribución de información de diferente tipo, recolectada por diferentes instituciones. Las plataformas de datos abiertos publican información en forma de conjuntos de datos (denominados en inglés como *datasets*). Los *datasets* pueden ser de mucha ayuda en tareas de re-uso de información. RDF es un método para describir semánticamente piezas de conocimiento, a partir de fuentes de información estructurada o no estructurada. RDF se ha convertido en el estándar de la descripción semántica de recursos de la Web Semántica. Con RDF se puede tener un conjunto de *datasets* y estos conformarían repositorios semánticos como un gran conocimiento.

La gestión de la información contenida actualmente en la Web avanza hacia un paradigma de Web Semántica (WS). La WS busca que cada dato se pueda extraer, analizar, publicar y enlazar entre varios datos. De hecho, el interés en esta área sigue siendo alta debido a la abundancia de aplicaciones prácticas que ayudan a los usuarios a lidiar con diferentes tipos de información.

A lo largo de este tiempo en la web ha ido generando grandes volúmenes de información y a su vez esta siendo tratada con el fin de obtener conocimiento de ellos de diferente índole. Lo que se busca es con este trabajo de titulación es establecer un método para enriquecer un grafo de conocimiento con diferentes fuentes externas.

1.1.1 ***Metodología***

La revisión sistemática es realizó con las siguientes pautas proporcionadas por (Okoli & Schabram, 2010, p. 7) y (Kitchenham & Charters, 2007) con la finalidad de buscar documentos relevantes de investigación. Los resultados de las búsquedas permitirán identificar desafíos o direcciones futuras en la investigación, así proporcionando un método verificable para resumir la búsqueda de información en la orientación de métodos para enriquecer grafos de conocimiento. La **Figura 1** presenta la metodología a seguir para la revisión sistemática.

Las siguientes subsecciones describen los pasos que se llevaron a cabo en la revisión sistemática del trabajo de titulación.

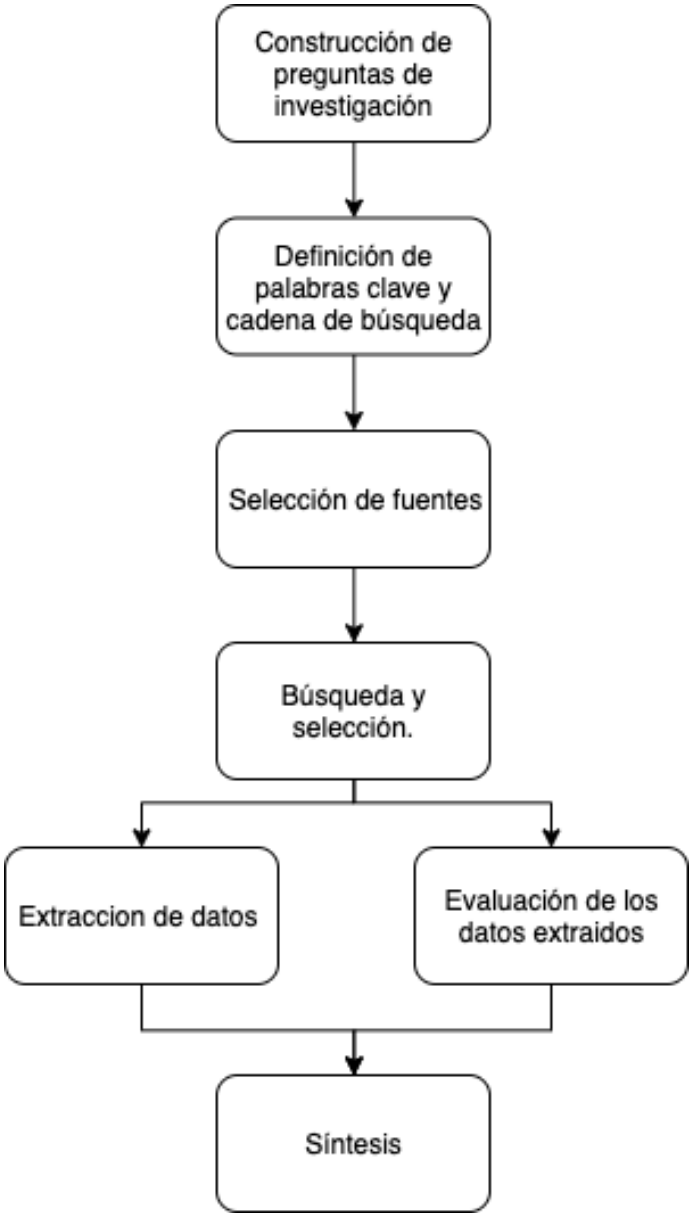
1.1.2 ***Definiendo las preguntas de investigación***

El objetivo de la revisión sistemática es comprender como un conjunto de datos o *datasets* pueden ser almacenados y esto a su vez con un método de enriquecimiento que pueden ser explotados para crear un conocimiento más grande. Por tanto, hemos definido las siguientes preguntas de investigación:

a) ¿Qué repositorios semánticos han desarrollado un método para enriquecer su grafo de conocimiento?
b) ¿Qué desafíos y problemas se han enfrentado para enriquecer los datos?
c) ¿Qué criterios y técnicas se utilizan para la evaluación el conjunto de datos?
d) ¿Qué direcciones son las más prometedoras para futuras investigaciones?

Figura 1

Revisión sistemática de manera general.



Nota. Elaboración propia

1.1.3 ***Búsqueda y selección***

La obtención de conjuntos de documentos relevantes para esta revisión sistemática requiere de búsquedas en las bases de datos científicas: Scopus[1], IEEE Xplore[2] y Google Scholar[3]. Estas bases de datos fueron elegidas en base a las exigencias que requiere la

investigación y porque estas son mundialmente reconocidas por sus investigaciones y publicaciones. La **Tabla 1,** se define un conjunto de cadenas de búsquedas para determinar los tipos de documentos serán óptimos.

Tabla 1

Cadenas de consultas

Base de datos científicas	**Cadena de búsquedas**
- SCOPUS[1] -IEEE Xplore[2] -Google Scholar[3]	(data AND enrichment AND in AND semantic AND web)
	(enrichment AND of AND open AND data AND graphs)
	(adding AND value AND to AND linked AND open AND data)
	(data AND enrichment AND in AND semantic AND web) AND PUBYEAR > 2015
	(data AND enrichment AND in AND semantic AND web) AND DOCTYPE (ar OR re) AND PUBYEAR > 2015

Nota: Representación de las cadenas de consultas en las bases de datos científicas.

El objetivo es encontrar artículos y documentos de calidad con la finalidad de obtener un refinamiento en los resultados de búsquedas y a su vez los resultados sean relacionados al tema de investigación. Sin embargo, debido al gran número de publicaciones de todo tipo se optimizarán las búsquedas, especificando los años y los tipos de documentos requeridos. Adicionalmente, que la etapa de búsqueda y selección de artículos fue de gran ayuda debido a que con esta forma se pudo definir de mejor manera que información será de gran relevancia y de que bases de datos científicas serán parte para el TT.

1.1.4 Extracción de artículos y documentos

Una vez determinado e identificado cada uno de los documentos científicos los cuales serán la base necesario para el desarrollo de la investigación. Y después de finalizar la búsqueda, se procede a extraer los metadatos de cada uno de los artículos consultados. Al extraer esta información de los diferentes sitios científicos estos contienen

[1] https://www.scopus.com/
[2] https://ieeexplore.ieee.org/
[3] https://scholar.google.com/

diferentes metadatos. La **Tabla 2** describe las características que se toma en cuenta para la extracción por cada búsqueda, los cuales son:

Tabla 2

Elementos de los resultados de la búsqueda

Elementos de datos	**Descripción**
Titulo del documento	Título del trabajo
Autor(es)	Los nombres del autor
Año	Año de publicación
Fuente y tipo de documento	El tipo de la investigación o área de interés
DOI	identificar un objeto digital
Tipo de acceso	Si el documento pude se accedido libremente por cualquier persona sin pago.

Nota: Representación de resultados de las búsquedas en las bases de datos científicas.

Una vez extraída la información de cada una de las búsquedas en las diferentes bases de datos científicas mencionadas. Se procede a excluir ciertos resultados con el fin de evitar la duplicidad de datos, limpieza de datos de los resultados obtenidos en las tres plataformas que fueron de ayuda. Por último, la estrategia de análisis fue leer el abstract, introducción y las conclusiones de los diferentes artículos o documentos para cerciorarse de que contiene todos los datos necesarios para este TT. Estos artículos o documentos se encuentran en el idioma inglés.

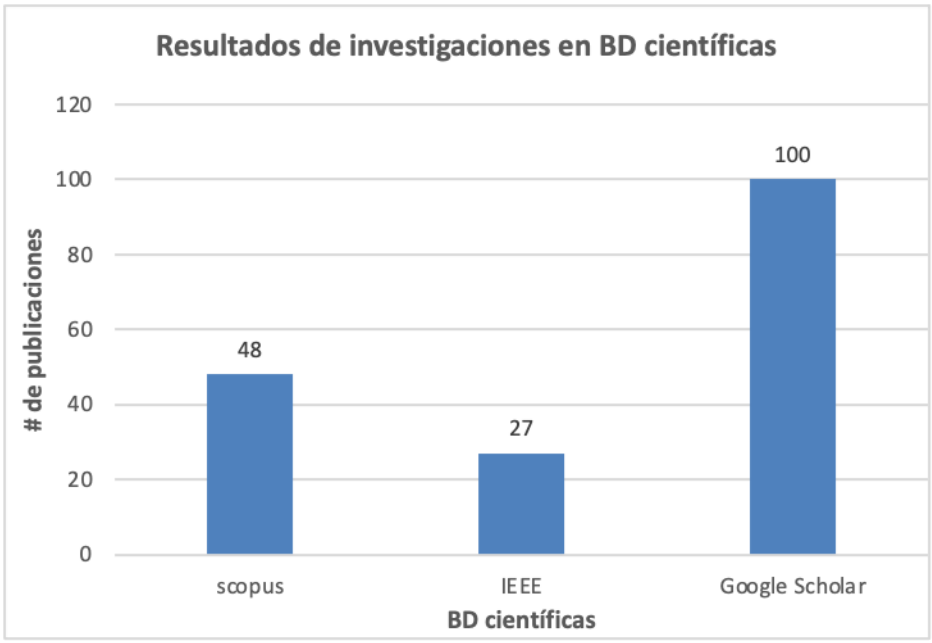
*1.1.5 **Evaluación de los datos extraídos***

Después de extraer y limpiar los datos de los artículos y documentos se procedió al análisis los datos artículos extraídos con el fin de tener un criterio más claro y determinar que información es más relevante. Además para el análisis se utilizan criterios que surgieron de las preguntas planteadas al inicio de la investigación. Los criterios usados para el análisis se los puede definir como temáticas o palabras claves de tal manera que se pueda obtener un resultado conciso y más acercado a lo que se pretende revisar. Las temáticas contienen las siguientes palabras claves: semantic web, linked data, knowledge graphs, enriching a knowledge graph.

Con base a los datos extraídos, las búsquedas y selección de los resultados se presentan en la **Figura 2**.

Figura 2

Resultado de búsquedas en BD científicas



Como resultado de las búsquedas se obtuvo un total de 175 artículos de los cuales se usarán 12 documentos que serán los base para la TT y la definición del método de enriquecimiento de grafos de conocimiento. Para la selección de estos 12 documentos se tomó en cuenta al año de publicación y el enfoque de la investigación que estos tienen en relación al trabajo de TT.

La **Figura 2** muestra el resultado final de las búsquedas obtenidas en cada una de las bases de datos científicas. Hay que tener en cuenta que la búsqueda en Google Scholar arroja un resultado total de búsqueda que inicialmente se obtuvo un valor aproximado de 1710 documentos, entonces se asumió que las diez primeras hojas del resultado de búsqueda son las óptimas, esto es los 100 primeros documentos.

Por otro lado, los criterios que fueron parte esencial para determinar los tipos de documentos que son claves para la investigación teniendo en cuenta los años de publicación, la participación de los estudios primarios como se muestra en la **Figura 3**.

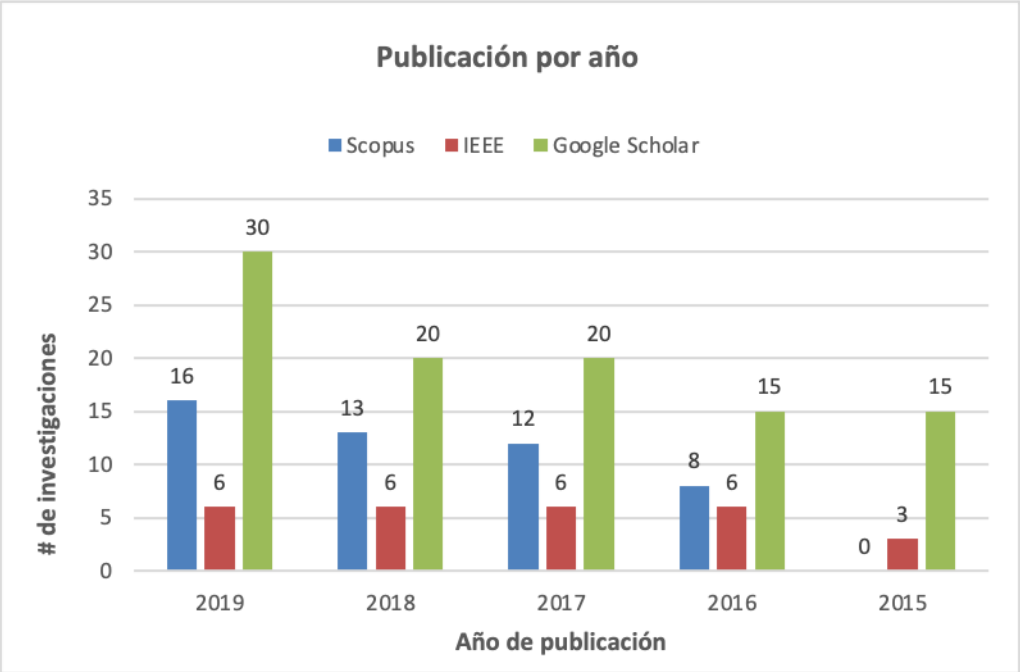
Cabe destacar que los documentos científicos seleccionados se publicaron desde el 2015 hasta la actualidad 2019 o 2020 en caso de que existan algunos.

1.1.6 Año de publicación y distribución geográfica

La **Figura 3** demuestra que esta área de investigación es moderna, novedosa y lleva varios años en la búsqueda de aportar conocimiento. En la **Figura 3** se puede observar el resultado de la distribución del año de publicación de todos los trabajos seleccionados tiene un incremento relativo con relación al año anterior donde se publicaron un total de 18 trabajos (10,2%) en el 2015, 29 trabajos (16,4%) en el 2016, 38 trabajos (21,5%) en 2017, 39 trabajos (22%) 2018 y 52 trabajos (29,4%) hasta noviembre del 2019. Por lo tanto, el número de publicaciones aumenta cada año, lo que refleja el aumento de interés en esta área de investigación.

Figura 3

Año de publicación de los documentos primarios seleccionados.



La **Figura 4** muestra la distribución geográfica de los países que están en la búsqueda sobre este tema. Existen más 20 países que tienen trabajos relacionados, pero se toma en cuenta los 10 primeros países que tienen gran número de trabajos para la

representación gráfica. En esta **Figura 4** se muestra la distribución geográfica donde la atención de la investigación tiene más impacto. El mayor porcentaje de trabajos o documentos es de 21,74% en Estados Unidos, y fueron publicados por universidades o empresas en función del proyecto de investigación. Después de esto, tenemos España Alemania y Grecia que comparten un porcentaje del 10,87%. En el resto de los países comparten valores similares entre ellos.

Figura 4

Distribución geográfica de los países en relación con el tema.

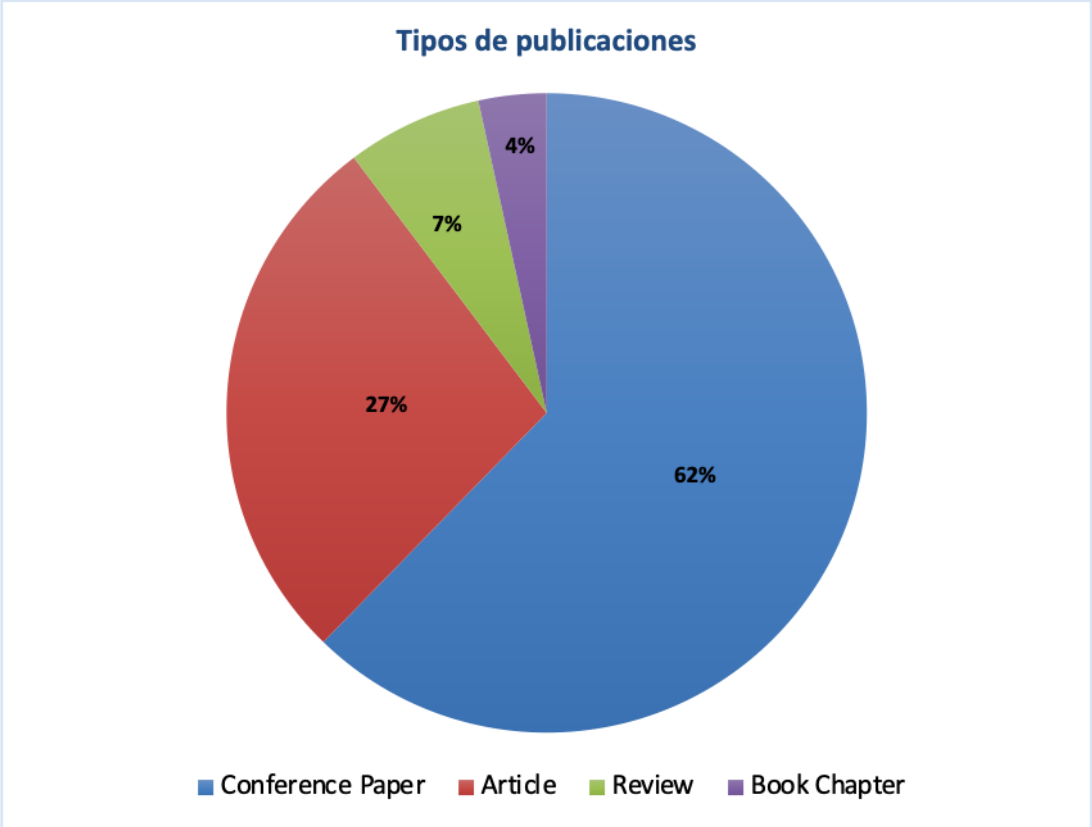
*1.1.7 **Tipos de publicaciones***

La **Figura 5** muestra el total de publicaciones de las cuales fueron seleccionadas para formar parte de esta investigación. Con un 62,29% del total de todos los documentos seleccionados fueron publicados en artículos de conferencia, seguido del 27,43% en

artículos científicos y 6,86% se publicaron en artículos de revisión. Una de las bases de datos científica o bibliografía con más número de publicaciones es Scopus, en este se indexan gran cantidad de documentos de investigación de diferentes categorías de las cuales se basan para determinar que método de enriquecimiento será usado para el TT.

Figura 5

Tipos de publicaciones

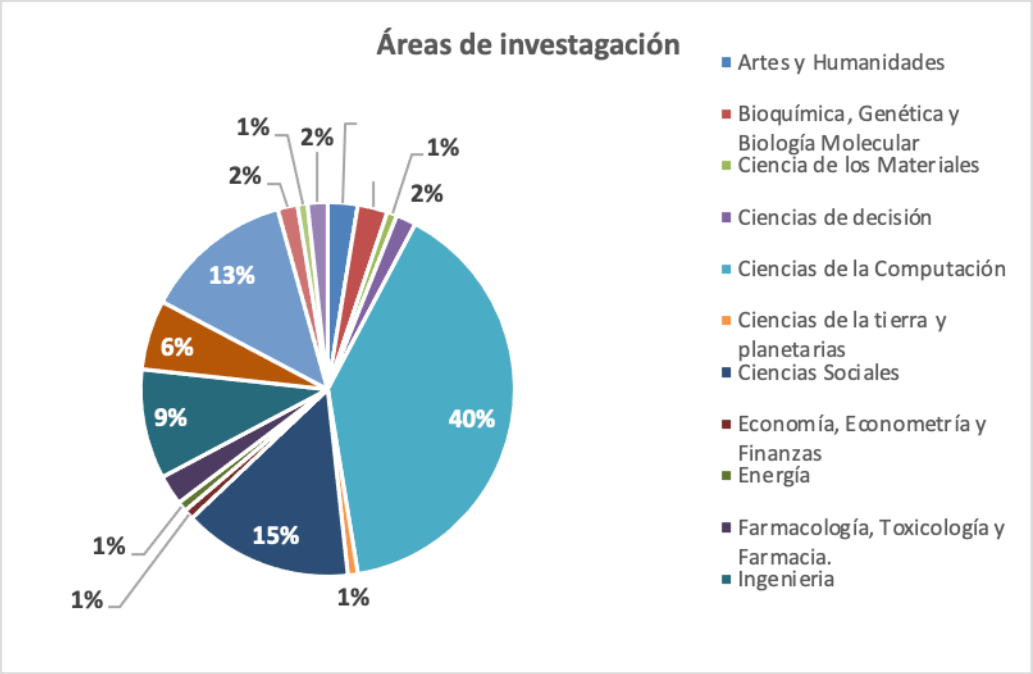


1.1.8 ***Áreas de aplicación o investigación.***

La **Figura 6** se aprecia mediante un diagrama de pastel como esta temática se ha ido aplicando en cada una de las áreas de investigación.

Figura 6

Áreas de investigación



Nota. Elaboración propia

La gran mayoría de los temas de investigación de las universidades se centran en las áreas técnicas y socio humanísticas, donde se puede apreciar gran cantidad de investigaciones relacionadas a este tema. Las áreas con mayor aporte es ciencias de computación que tiene un 43% y representa las áreas que están interesadas en resolver o buscar métodos para implementar y adaptar las diferentes tecnologías para este tema del TT, le sigue ciencias sociales con un 15% y luego el área medica que comparte un 13% de todas las investigaciones en este apartado.

1.1.9 ***Discusión y Resultados.***

Finalmente, después de todas las fases anteriores se llegó a la selección de los diferentes métodos mencionados en cada artículo analizado, basándose desde la fase de recolección de los datos, transformación y hasta el enriquecimiento de grafos de conocimiento, siendo esto la parte fundamental para el desarrollo del TT.

Son 12 artículos científicos que son base para este TT. Para la selección de esto se basa en el año de publicación y el enfoque del contenido que estos tienen. De estos 12, en

5 artículos mencionan diferentes procesos o métodos de enriquecimiento de grafos y los restantes son los que permiten entender sobre la web semántica, grafos de conocimiento y más temas relacionados al TT. A continuación de describen algunas formas para realizar el proceso de enriquecimiento.

Según (Quattrini et al., 2017) propone un modelo de enriquecimiento datos basándose en usar un etiquetado semántico, una organización semántica a todo el conjunto de datos y para relacionar o enlazar los *datasets* con otros usaron *Revit* el usar parámetros compartidos para enriquecer datos. Estos parámetros usados por *Revit* son de crear, agrupar por áreas temáticas, agrupar conjuntos de parámetros de una ontología de dominio y diferentes formas de agrupas como texto, números, referencias de enlaces externos, entre otros.

Otras formas de enriquecer grafos son planteadas por (Jabbar & Bulbul, 2019) y (Sánchez et al., 2018), los cuales proponen dos formas, la primera definida por *Jabbar* y *Bulbul* mencionan tener un almacén de datos semánticos donde tiene un motor de transformación a *linked data*, y el enriquecimiento se realiza usando el motor de enriquecimiento semántico de *Jena* y así tener un servicio de linked data para luego su explotación.

Por otro lado, (Sánchez et al., 2018) en su trabajo proponen una vinculación híbrida entre vinculación automática y manual. La vinculación automática se realiza para todos los recursos mediante el servicio de reconciliación ofrecido por RDFRefine, donde se establece Sparql Endpoint y la propiedad con la que se buscan los recursos relacionados y por otro lado la vinculación manual trata de enlazar datos, pero con los criterios que requiere para la vinculación de datos extendidos.

En (Escobar et al., 2020) se plantea un método de enriquecimiento de grafos, partiendo desde la extracción de datos hasta el enriquecimiento. Primero se debe especificar las fuentes de datos (utilizan un proceso ETL para normalización de datos obtenidos de fuente de datos heterogéneos), establecen un modelo RDF de datos (transformación de datos usando vocabulario Data Cube), generación datos (OpenRefine)

esta es la parte que se encarga de enriquecer los datos y finalmente la publicación y explotación de los datos.

Otro método propuesto por (Song & Park, 2018) menciona en usar un modelo basado en traducción, esto consiste en integrar entidades y relaciones en un solo conjunto de datos de la misma área o del mismo vector. Con la traducción de entidades busca mapear vectores de entidad en diferentes vectores en espacios de relación según la entidad y la relación.

Todos estos métodos planteados por los investigadores mencionados serán de gran ayuda para definir un método que se adapte a la necesidad para resolver en problema de este trabajo de titulación.

Por otro lado, al concluir este análisis e interpretación de los métodos usados en las investigaciones mencionadas se tiene como resultado las respuestas a la incógnitas planteadas para esta revisión sistemática. La primera pregunta planteada menciona lo siguiente, ¿Qué repositorios semánticos han desarrollado un método para enriquecer su grafo de conocimiento? En la web existen múltiples repositorios semánticos que se usan para tareas de enlazado y enriquecimiento, por ejemplo en (Escobar et al., 2020) se presenta el proceso o método para la recolección, tratamiento, formulación de estructura semántica y el enriquecimiento con otras fuentes como GeoNames, RDF Data Cube con la finalidad de ampliar el repositorio Barcelona Open Data (BOD)..

Después de haber revisado todos los posibles desafíos y problemas que se puede encontrar al momento de enriquecer los datos se resumen en lo siguiente. El problema es que no existe un proceso estándar que permita que todos los datos existen en la web puedan ser estructurados semánticamente, llevando esto a un nuevo desafío de poder generar un método que permita estructurar y esto a su vez permita añadir más información a los datos existentes. De la misma manera los criterios y técnicas se utilizan para la evaluación el conjunto de datos se basa en los criterios que los investigadores planteaban en sus artículos.

Finalmente las direcciones más prometedoras para plantear las investigaciones acorde al tema es que se pueda crear modelos estándares que permitan que cualquier tipo de información semántica se pueda enriquecer. Por consiguiente el objetivo de esta revisión sistemática fue identificar investigaciones de temas predominantes dentro del campo de las tecnologías de web semántica y posibles aplicabilidades de métodos que permitan enriquecer un grafo de conocimiento semántico.

En resumen se menciona lo siguiente, que mediante una fase de 5 pasos se puede resolver un poco el problema de los datos no enlazados o enriquecidos, estas 5 fases son las siguientes:

1. Extractor: permite la extracción rutinaria y automatizada de nuevos datos de fuentes especificas.
2. NER (Reconocimiento de Entidades Nombradas): permite la extracción de información que se busca localizar y clasificar en categorías predefinidas, tales como: personas, organizaciones, lugares y las entidades nombradas encontradas en un texto.
3. NLP (Procesamiento del Lenguaje Natural): reconocimiento de entidades en el cual se pueden idear búsquedas lingüísticas específicas y detectar similitudes entre entidades.
4. Flujo de trabajo: los usuarios podrán hacer uso del API de los datos.
5. Publicación: la publicación de los datos en el piloto y el resultado de visualización serán recursos fáciles de procesar por parte del usuario final.

Es importante tener en cuenta que varios de estos los temas son parte fundamental para el desarrollo de la propuesta de solución del TT.

1.2 **Conceptos**

En esta sección se detallan conceptos claves para el desarrollo del Trabajo de Titulación (TT) que permitieron la mejor compresión del contexto del TT y la presentación una propuesta relevante que permita resolver o satisfaga a los objetivos del tema. En las secciones 1.2.1 y 1.2.2 se presentan partes fundamentales para el entendimiento del desarrollo del TT. La sección 1.2.3 analiza la importancia de tener claro los conceptos que sean útiles en el campo de la web semántica con finalidad de crear una propuesta de solución para el trabajo. La sección 1.2.4. muestra la técnica para la integración, extracción de información y herramientas que se usaran para resolver el problema. Finalmente, en la sección 1.2.5. realizará una discusión del contenido obtenido.

1.2.1 ***Datos abiertos***

Se debe tener en cuenta que los datos abiertos, también conocido como Open Data (OD). Según (Open Data Handbook, 2019) define que los datos abiertos son "*datos que pueden ser utilizados, reutilizados y redistribuidos libremente por cualquier persona*", nos referimos a datos que están disponibles de forma libre por la Web. Estos datos tienen que pasar por varios niveles para poder ser completamente libres y tienen que cumplir ciertas características como la disponibilidad y acceso, pues tienen que estar disponibles por completo y así estos pueden ser accedidos por cualquier interesado. Otra característica es la reutilización que se tiene que hacer bajo unos permisos o términos que permitan su uso y explotación de forma libre y no pongan ninguna restricción a la hora de trabajar con otros tipos de datos. Todos los interesados que deben poder acceder a los datos y trabajarlos como ellos lo requieran, pero sin ninguna restricción alguna.

1.2.1.1 Importancia

La importancia de OD radica en la posibilidad de acceder a los datos de forma libre y así mejorar el conocimiento[4]. También los datos deben contribuir a mejorar eficientemente los servicios de cualquier entidad o grupo que usen estos datos (Pfenninger et al., 2017). Principalmente permite el acceso, la responsabilidad y la transparencia con la finalidad de que estos recursos abiertos permitan beneficiar a varios individuos o también permite ayudar con búsquedas de soluciones.

Hay muchas áreas donde podemos esperar que los datos abiertos sean valiosos, por lo tanto, (Duus & Cooray, 2016) mencionan en su portal que la disponibilidad de datos abiertos crean oportunidades para todo tipo de organizaciones, gobiernos y organizaciones con o sin fines de lucro para encontrar nuevas formas de abordar los problemas, como esto se promueve que exista transparencia en las acciones que estos grupos realicen.

Los datos abiertos juegan un papel fundamental para poder de ampliar la visión de cada persona en el uso y manejo de los datos. Según (Duus & Cooray, 2016) menciona que los datos describen los patrones detrás de cómo vivimos puede ayudarnos a resolver problemas de formas que no hubiéramos previsto. El autor concluye que la apertura de los datos abiertos es de gran ayuda con la finalidad de tener una perspectiva diferente a lo que cada uno tiene actualmente de la información porque los OD buscan resolver problemas.

1.2.1.2 Calidad de los datos abiertos

Dentro de la calidad de OD vienen también los retos que estos pueden tener como poder realizar la captura de datos y de estos tener obtener una información bien estructura.

Las herramientas de análisis de datos cada vez más sofisticadas, así permitiendo analizar los datos de nuevas maneras para descubrir tendencias y hallazgos que han dado forma a decisiones cruciales. Estas herramientas permiten ver soluciones entre datos que nunca han sido explorados. Para esto la información debe estar abierta y disponible.

[4] Tecnologías de la información y la comunicación

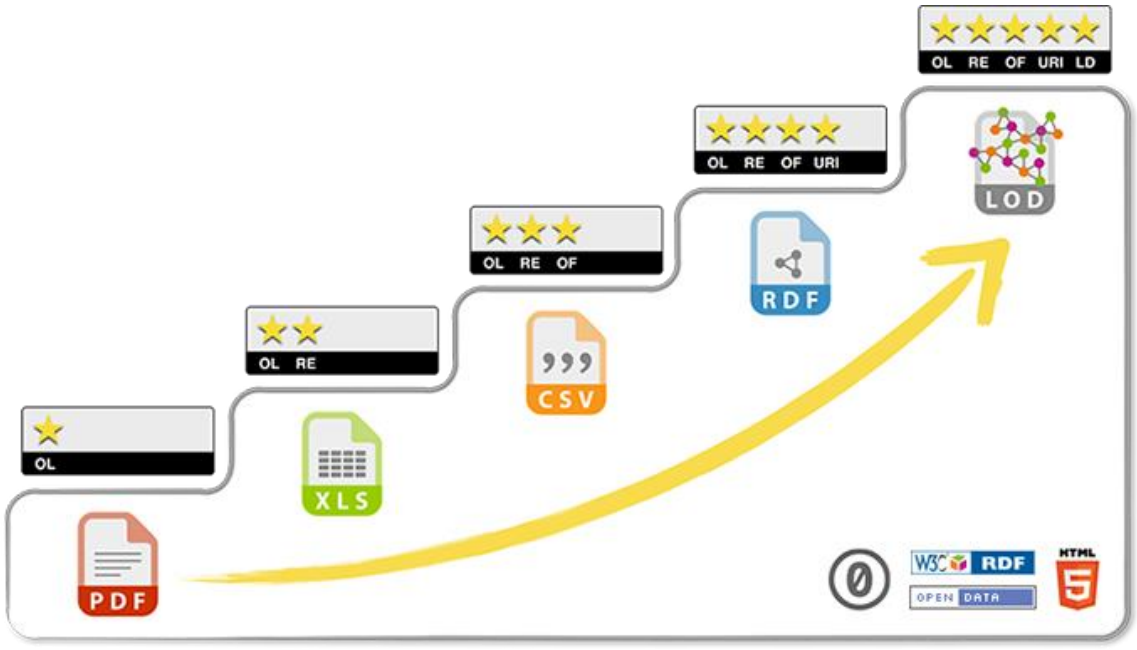
La disponibilidad de información y el acceso público a los datos cada vez mayor y el crecimiento exponencial de la información día a día es muy grande desde el nacimiento de la World Wide Web (WWW) que fue planteado por Tim Berners-Lee. Uno de los mayores retos que menciona (Stauffacher et al., 2012) es que el acceso a datos o información se traduce en el empoderamiento con el fin de poder tomar decisiones informadas, resolver problemas y para mejorar el nivel de vida de las personas.

Para que cada uno de estos datos puedan ser usado debe de pasar por diferentes procesos de análisis con la finalidad de determinar a que estrella o nivel se encuentra y de esta manera saber si esta cumpliendo con cada una de las características que tiene la Open Data. Tim Berners-Lee, el creador de la WWW, sugirió un esquema de desarrollo de 5 estrellas de OD[5] el cual describe en niveles los estados de los datos en los que se en cuenta desde el primer nivel hasta llegar a la ultimo nivel en la cual representa que los datos son abiertos, de acceso libre y con una estructura fácil de manejar y mantener. La **Figura 7** muestra de manera resumida los niveles de estrellas en los que se encuentra los datos.

[5] https://5stardata.info/en/

Figura 7

Calidad de OD en 5 niveles



Nota. Adaptado de 5 stard data [Fotografia], por (Berners-Lee, 2015)

Como se aprecia en la **Figura 7** cada nivel está representado por una estrella a través de la que se evalúa que tan accesible es la información. Según (Berners-Lee, 2015) describe los niveles de la siguiente manera en la ***Tabla 3***:

Tabla 3

Niveles de las 5 estrellas de la Open Data

★☆☆☆☆	Los datos deben estar disponibles en la Web (cualquier formato) bajo una licencia abierta
★★☆☆☆	Deben ser datos estructurados, por ejemplo, un archivo usado en Excel.
★★★☆☆	Deben estar disponible en un formato abierto no propietario, por ejemplo, CSV en lugar de Excel.
★★★★☆	El uso de URI para denotar cosas, de modo que las personas puedan señalar sus cosas.
★★★★★	Los datos deben estar vinculados entre datos para proporcionar contexto.

Nota: Presenta los 5 niveles de los datos abiertos planteados por (Berners-Lee, 2015)

En este trabajo se ha considerado una versión ampliada del modelo de cinco estrellas de calidad de los datos abiertos, reportada en (Alesso & Smith, 2006).

1.2.1.3 Plataformas de datos abiertos

Existen diversas plataformas de datos abiertos o Open Data Platform (ODP), las cuales permiten obtener datos o a su vez publicar todo tipo de información relevante de varias temáticas, son una parte fundamental en el proceso de apertura de datos según (Techopedia, n.d.).

Las ODP sirven para almacenar, compartir, conectar y visualizar todo tipo de datos en una base de datos que estas posean. También son las puertas de entrada a la búsqueda de estrategias de una organización o grupos con el beneficio de abrir sus datos y así los ciudadanos pueden ver esta información (Techopedia, n.d.).

Dentro de las plataformas no existe un único modelo para los datos, sino los portales poseen varias características con algunas secciones concretas, pues todo esto no permite tener una misma estructura de la información de una plataforma con otra.

Hay información y herramientas que permiten trabajar con datos que son abiertos o públicos. Algunas aplicaciones de varias funciones en la sociedad se han desarrollado gracias a estos datos, ha permitido también que esta información ayude a predecir eventos que puedan suceder. A continuación, enumeraremos las partes fundamentales que tiene un portal de datos abiertos (Danneels et al., 2017):

a. **Catálogo de datos:**

   Los catálogos de datos son la parte esencial de las plataformas de datos abiertos, estos poseen una lista en una tabla los cuales describen el posible contenido que estas puedan ofrecer, el nombre de los responsables de la información, el formato en la que se encuentran (pdf, xlsx, csv, json entro otros), la frecuencia que estos son actualizados, número de visitas entre más características que cambian según los portales.

b. Colección de datos:

La colección de datos es un listado de información. La colección puede tener un enfoque con temas relacionados (salud, economía, clima, educación) o puede tener alguna organización que esta a cargo de estos datos.

c. Visualización de datos:

Es el sitio donde se da a conocer de otra manera los datos, es decir, por medio de gráficas o vista de datos que fueron desarrollados por los usuarios o administrador del sitio, todo esto depende si la plataforma nos permite.

d. Historias de datos:

Las historias basadas en datos son consideradas una prioridad para muchas organizaciones. Estas historias de datos se conforman por versiones previas que se han llevado durante un tiempo. Estas historias de datos pueden ayudar a tomar decisiones o también a determinar como ha ido cambiando la organización a lo largo de este tiempo ya que nos permite dar una retroalimentación de como estas han ido avanzando a lo largo del tiempo y ver si las cosas que han hecho han sido la mejor opción o no.

*1.2.2 **Objetivos de desarrollo sostenibles.***

En el 2015 se reunieron en las Naciones Unidas[6] (UN) todos los lideres de las diferentes naciones para adoptar un conjunto de objetivos con una visión de mejorar la vida en un futuro para todos, estas metas están planteadas en la Agenda 2030 de la siguiente manera, 17 Objetivos de Desarrollo sostenibles[7] (ODS) con 169 objetivos y 330 indicadores (Griggs et al., 2013). Las Naciones Unidas ha estimado un incremento de la

[6] https://www.un.org/es/
[7] https://www.un.org/sustainabledevelopment/sustainable-development-goals/

población de 66% para el 2050 lo que esto implica más retos y tratar de solucionar de manera rápida y oportuna a estos problemas.

1.2.2.1 Indicadores de los ODS alineados con los datos abiertos

A través de los ODS, los países cuentan con indicadores específicos, observables y medible que permiten mostrar el cambio y progreso en las actividades relacionadas a los ODS (*Indicadores de ODS*, 2010).

En la búsqueda de ayudar a mejorar la calidad de vida de todos con los ODS, varios grupos o personas han desarrollado herramientas que permite tener de cierta manera características de cada uno de los ODS y así estos puedan ser accedidos y poder relacionar grupos de datos abiertos a cada criterio con los indicadores y así saber de manera mas rápida el crecimiento que tiene el objetivo de desarrollo sostenible. Las Naciones Unidas con el objetivo de ayudar a determinar esto desarrollaron una API[8] que permite saber tantos detalles de los objetivos como de los indicadores. Estos recursos pueden ayudar de gran manera a integrar varios recursos abiertos a cada uno de los ODS e indicadores.

*1.2.3 **Web Semántica***

A lo largo de estos años la web ha ido variando desde su primera versión la 1.0 basándose en que las personas puedan intercambiar información en la web. Después pasó a la web 2.0 que se describe como la segunda generación de World Wide Web (WWW)[9] donde las páginas estáticas son reemplazadas por páginas interactivas para una experiencia dinámica para los usuarios. La Web 2.0 es creada con el objetivo de permitir a las personas compartir su información en la red a través de las plataformas de redes sociales, blogs, chats y poder contribuir en la web (Cheaney, 2012).

La Web 3.0 es el tercer nivel de desarrollo de la web a lo que se denomina la Web Semántica (WS) la que demuestra la naturaleza semántica de los contenidos. Tim

[8] https://unstats.un.org/SDGAPI/swagger/
[9] Es un término que es utilizado para referirse a la World Wide Web, también conocida como Internet.

Berbers-Lee fue el mentor del término de web semántica lo que nos dice que los datos que se encuentran en la Web es información que se puede tratar y buscar relaciones entre ellas con situaciones reales (Sharma, 2015).

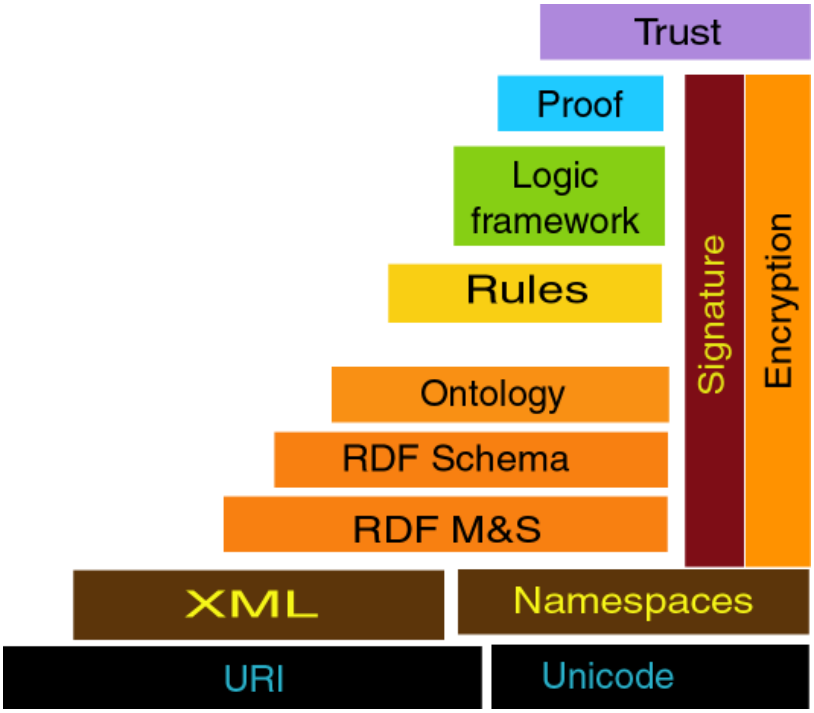
Tim Berners-Lee, el padre la web, señala que "*La Web Semántica es una extensión de la web actual en la que la información tiene un significado bien definido, permitiendo que las computadoras y las personas trabajen mejor en cooperación*", (Berners-Lee et al, 2000) esta visión, evoluciona la Web actual en una red de datos y significados global, con un enorme pontencial para apoyar a la resolución de problemas que requieren gestionar conocimiento global, de manera rápida y oportuna.

1.2.3.1 Tecnologías de la Web Semántica

La Web Semántica desde sus inicios se ha planteado varios criterios que se deben considerar el desarrollo de la WS y sus posibles aplicaciones. La **Figura 8** presenta la organización por capas de las tecnologías de la WS cual será el criterio clave para el desarrollo del TT.

Figura 8

Las capas de las tecnologías web semánticas



Nota. Adaptado de w3: capas de las tecnologías [Fotografía], por (Hazaël-Massieux & Berners-Lee, 2003)

Las capas fundamentales para el desarrollo del TT se las describe a continuación:

a. URI

Uniform *Resource Identifier* (URI), su significado en español es Identificador de Recursos Uniforme, su función es nombrar recursos en la web, lo que un navegador recibe es información o le proporciona un método/ubicación de acceso al recurso (Miessler, 2020).

Esté termino surge de la unión de Uniform Resource Locator (URL) y Uniform Resource Name (URN).

b. XML

XML es conocido como eXtensible Markup Language, traducido como Lenguaje de Marcado Extensible o Lenguaje de Marcas Extensible, el cual es un lenguaje de marcado similar a HTML y es una especificación de W3C, es decir que XML no está predefinido, por lo cual debe ser definido las propias etiquetas (Gavin, 2018).

c. RDF

Resource Description Framework (RDF) es un estándar que permite la interoperabilidad entre datos en la Web. Según (Lapuente & Lamarca, 2018) “RDF es un lenguaje para representar información sobre recursos, … particularmente sobre metadatos de recursos web, tales como el título, autor, modificaciones de los datos de la página web, copyright y otras licencias de información sobre documentos web, así como la disponibilidad para algunos recursos compartidos”. Todo esto se trata de un modelo de datos para objetos o recursos.

1.2.3.2 Ontologías y vocabularios

Las ontologías tienen componentes que sirven para representar el conocimiento de algún domino. Son descritas como una estructura para la comunicación entre humano y máquinas. Estas permiten mejor las búsquedas, la comunicación entre conocimientos y entre otros aspectos.

Según (W3C, 2015) los vocabularios definen conceptos y relaciones que son usados para describir y representar cosas. Estos vocabularios pueden ser muy complejos a la hora de usar una gran cantidad de datos.

La W3C dice que no existe la división exacta entre vocabulario y ontologías, ya que la tendencia es usar la palabra *ontología* para una colección de términos más complejos. Mientras que *vocabulario* es usado para la construcción de bloques con técnicas básica a seguir.

1.2.3.3 Linked data

Linked data es un término que hace referencia a un método de publicación de datos, pero estructurados. Esta tecnología tiene dos ideas, las cuales engloban en ser más grandes y poderosas.

Este método permite interrelacionar conjuntos de datos, por lo que nos permite que sea mas fácil y útil para la búsqueda de información. Según (Bizer, 2009) menciona que la linked data se refiere “*al conjunto datos que buscar mejorar la información para publicar y conectar datos estructurados en la web*”, siendo este el siguiente paso de la open data con

la finalidad de que estos datos tengo una estructura pero que estos se puedan relacionar con más conjuntos de datos abiertos.

Linked data (LD) tiene algunas capacidades o características que según (Bizer, 2009) pueden ser la siguientes:

- Cualquiera puede publicar datos en la Web de datos vinculados.
- Los enlaces conectan entidades creando un gafo que abarca fuentes de datos y permite el descubrimiento de nuevas fuentes de datos.
- Los datos se auto describen. Si una aplicación encuentra datos representados usando un vocabulario desconocido, la aplicación puede resolver los URI que identifican los términos del vocabulario para encontrar sus definiciones.
- La Web de datos vinculados está abierta, lo que significa que las aplicaciones pueden descubrir nuevas fuentes de datos en tiempo de ejecución siguiendo los enlaces.

El enfoque de datos enlazados puede ser aplicado en cualquier tipo de información. A través de LD la información se estructura como piezas de conocimiento, que se pueden enlazar con otros datos enlazados, y así expandir el conocimiento (Piedra, N. & Suárez, J.P, 2017)..

1.2.3.4 Aplicaciones de Web Semántica

A la medida que pasa el tiempo, la Web Semántica ha ido teniendo mayor auge y se ha visto que está siendo aplicada en diferentes áreas de investigación. Al igual que las necesidades han incrementado, el volumen de los datos también, en los cuales su estructura se ha vuelto más heterogénea y difícil de procesar, pues gracias a esto sea visto la necesidad de tratar y dar un valor agregado a esta información con la finalidad de encontrar nuevas oportunidades y tratar de crear nuevas aplicaciones con esta información como: Buscadores semánticos, servicios de web semántica, e-Learning, entre otras.

Finalmente, las aplicaciones web semánticas se las pueden definir como aplicaciones basadas en la web que aprovechan el contenido semántico y estructurado

que estas poseen incluyendo no solo información, sino también metadatos que describen cada uno de los recursos obtenidos según (Alesso & Smith, 2006).

1.2.4 ***Procesamiento de lenguaje natural***

El procesamiento de lenguaje natural (NLP[10]) es la capacidad de interpretar o analizar el lenguaje humano con la finalidad de saber de lo que están hablando en el conjunto de datos que son analizados.

Otra definición que nos entrega (Wonderflow, 2018) menciona que es la "interacción entre el lenguaje humano con las computadoras". Una de las características es que el NLP es superior a las actividades de un ser humano en la capacidad de interpretar gran cantidad lenguajes y datos. El NLP es una rama de la inteligencia artificial (AI[11]).

El desarrollo del NLP es desafiante debido a los algoritmos que son usados para realizar los procesos que necesitan para una compresión más fuerte para poder interpretar la información por un computador. Para que esto se realice debe ser mediante comandos de voz o texto, los cuales son interpretados por los computadores. Para el desarrollo del NLP existen varias librerías o framework en los diferentes lenguajes los cuales permiten realizar el análisis de la información que se quiere interpretar.

En la web existen varios métodos que son fáciles de aplicar, por ejemplo, tenemos a **fasttext**[12] desarrollado para Python y el cual es una librería de aprendizaje de incrustaciones de palabras y de clasificación de texto, este modelo permite un aprendizaje supervisado y no supervisado dando un resultado de una representación de vectores para las palabras y soporta 294 idiomas.

Con esta librería se busca entender el contexto de cada palabra y así formar la oración que serán usadas para la búsqueda, de tal manera refinar la búsqueda y obtener un resultado más acorde a lo que se quiere.

---

[10] Natural Language Processing
[11] Artificial Intelligence
[12] https://fasttext.cc/

Finalmente, el Procesamiento de Lenguaje Natural permite realizar varias actividades de manera más rápida y oportuna. El ser humano no puede realizar análisis o interpretar grandes volúmenes de datos en pocos segundos.

1.2.5 ***Base de datos semántica con GraphDB***

GraphDB es una base de datos grafos que sirve a las organizaciones para almacenar, organizar y gestionar contenido en forma de datos inteligentes semánticamente enriquecidos y es altamente eficiente, escalable y robusta con soporte para RDF y SPARQL. GraphDB incluye integración de datos e interconexión, cumplimiento de estándares W3C, modelo de datos expresivo, rico y flexible, espacio único de información interconectado formado por datos estructurados y documentos de texto, razonamiento, compatibilidad con datos abiertos vinculados y procedencia de datos. Según (Ontotext, 2020), empresa que creo GraphDB menciona que tiene tres tipos de licencias, una gratis, la estándar y versión empresarial; en la que se trabajó es con la licencia gratis.

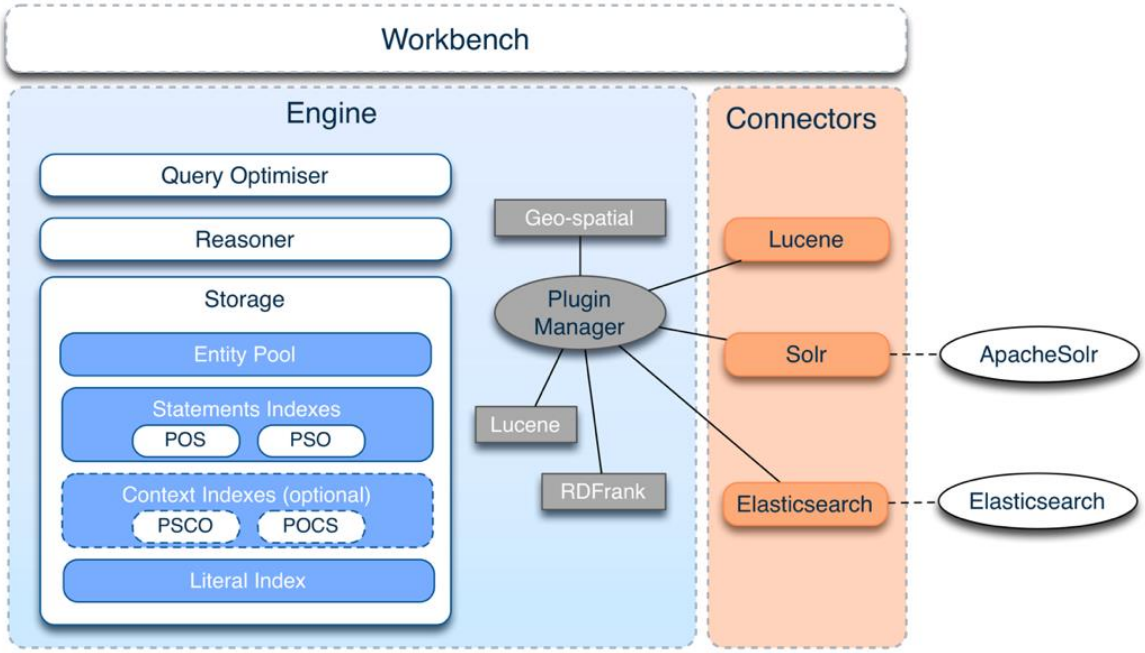
1.2.5.1 Arquitectura y componentes

GraphDB se estructura con una capa de almacenamiento e inferencia (SAIL[13]) para esto usa las propiedades del framework RDF4J y hace un amplio uso de las características y la infraestructura de RDF4J, especialmente el RDF modelo, analizadores RDF y motores de consulta.

[13] Storage And Inference Layer

Figura 9

Arquitectura de alto nivel GraphBD



Nota. Tomado de Ontotext architecture GraphDB [fotografía], por (Ontotext, 2020)

1.2.5.1.1 RDF4J

RDF4J es un framework para almacenar, consultar y razonar con datos RDF. Es compatible con el lenguaje de consulta W3C SPARQL. También es compatible con los formatos de archivo RDF más populares y los formatos de resultados de consultas incrustado en una aplicación como una biblioteca Java. Para comunicarse con RDF4J se puede utilizar una API con las funciones de JDBC y de RESTful HTTP.

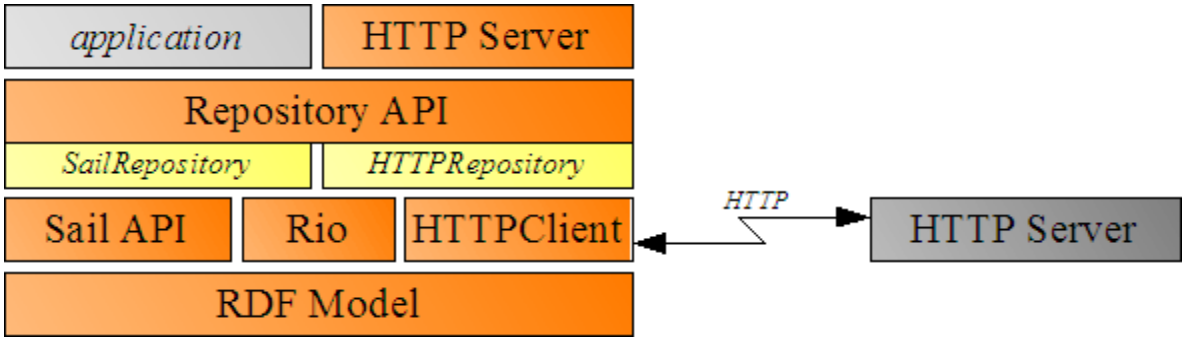
La **Figura 10** nos muestra una representación esquemática de la arquitectura de RDF4J y una breve descripción de los componentes principales.

Las aplicaciones normalmente se comunicarán con RDF4J a través de la API del repositorio y este proporciona un nivel de abstracción lo suficientemente alto como para que los detalles de los componentes subyacentes particulares permanezcan ocultos, es decir, se pueden intercambiar diferentes componentes sin necesidad de modificar la aplicación.

Este TT está basado en Python y la comunicación que se usa para el desarrollo se el framework SPARQLWrapper. El formato de comunicación del framework sigue manteniendo la misma estructura arquitectónica que plantea RDF4J.

Figura 10

Arquitectura RDF4J



Nota. Tomado de Ontotext architecture *RDF4J* [Fotografía], por (Ontotext, 2020)

1.2.5.1.2 Workbench

Workbench es una herramienta de GraphDB que permite la administración de su *triplestore*, Según (Ontotext, 2020) proponen los siguientes puntos en lo que se puede usar:

- Gestionar repositorios GraphDB
- Carga y exportación de datos
- Ejecutar consultas y actualizaciones SPARQL
- Gestión de espacios de nombres
- Manejo de contextos
- Visualización / edición de recursos RDF
- Consultas de seguimiento
- Monitoreo de recursos
- Gestión de usuarios y permisos
- Conectores de gestión
- Proporciona API REST para automatizar varias tareas para administrar y administrar repositorios

1.2.6 ***Discusión***

Este trabajo propone como obtener las ventajas que entrega la Web Semántica con esto mejorar todas las formas de relacionar su contenido con más fuentes de datos semánticos. Se busca realizar el proceso de enriquecimiento de datos para esto se debe tener en cuenta que los datos deben de estar en un formato abierto y semántico con la finalidad de poder realizar consultas SPARQL sobre el contenido que se quiera enriquecer.

Para esto se debe tener en cuenta todos los criterios que se plantear en la sección 1.2, siendo estos la base fundamental para tener la idea principal del trabajo de titulación.

Al definir cada uno de los conceptos que se mencionan en esta sección (1.2), se busca identificar que aspectos son los necesarios para estar al tanto de que va el TT. Por otra parte, se define ciertas herramientas que serán usadas para el desarrollo de modelo de solución del TT.

Las herramientas mencionadas en esta sección cumplen el rol fundamental para el manejo y uso de los datos CKAN semánticos y así sea posible poner la solución de TT.

Capítulo dos

Propuesta

La propuesta presentada en este capítulo está fundamentada en la revisión de la literatura realizada en el capítulo anterior, y enfocada en alcanzar los objetivos planteados para el Trabajo de Titulación.

En la propuesta se define un método de enriquecimiento de un grafo de conocimiento relacionado con indicadores de ODS, se define los conjuntos de metadatos que serán usados para el análisis e integración con otros metadatos y así estos se puedan relacionar con los ODS, se crea una base de conocimiento, se aplica un método de enriquecimiento, y se desarrolla una herramienta Web para visualización de datos.

En la sección 2.1 trata del contexto del trabajo, la sección 2.2 se menciona cuál es el problema general del Trabajo de Titulación, la sección 2.3 se trabajo en la propuesta del método de solución y en la sección 2.4 se presenta el modelo de la arquitectura que se emplea para el método de enriquecimiento de grafos.

2.1 *Contexto*

Los Objetivos de Desarrollo Sostenible son indicadores que permiten medir el progreso de la información en la web en relación de los objetivos, metas e indicadores que estos presentan. Desde el punto de vista de los datos se busca saber si la información semántica está relacionada con los indicadores y determinar si está surgiendo un cambio o no, debemos tener en cuenta que todos los datos relacionados están en formatos abiertos en la web. Sin embargo, existen más fuentes de datos abiertos que podrían ser de gran ayuda para aplicar la visión de los datos semánticos en la web.

2.2 *Problema*

La web ha ido creciendo rápidamente y la información en ella también, y eso ha generado datos heterogéneos; por consiguiente, un enorme volumen y variedad de datos inadecuados para los ordenadores tradicionales, lo cual a dado origen a la Big Data. Los datos abiertos son aquellos que pueden ser accedidos y utilizados cuando están disponibles en un formato común legible por máquina, lo que permite que una

computadora los lea y procese automáticamente. Tim Berners-Lee propuso un nuevo modelo llamado Linked Data para publicar información legible por máquina como datos estructurados, basado en RDF. Los datos abiertos vinculados (LOD) son datos que se publican bajo una licencia abierta.

El problema que al existir grandes volúmenes de datos y de diferentes fuentes es que estos no pueden incrementar el conocimiento que estas poseen (Eguiguren Palacios, 2019). Por otro lado, se tiene a los ODS que permiten analizar que datasets pueden estar vinculados a sus objetivos, metas e indicadores. Por consiguiente, lo que se busca es este Trabajo de Titulación es enriquecer un dataset con otros datos vinculados, para lo cual se crea un método que permite enriquecer los datos y se enlacen con otros datasets y a su vez con los Objetivos de Desarrollo Sostenibles.

2.3 *Propuesta de solución*

Se requiere desarrollar un método de enriquecimiento de Grafos de Conocimiento de indicadores de ODS usando un enfoque de inferencia de entidades semánticas equivalentes desde fuentes de datos abiertos. Para logar esto se realizó previamente una revisión sistemática en bases de datos científicas para encontrar investigaciones relacionadas que describe diferentes métodos usados para el enriquecimiento de grafos de conocimiento ya que estas investigaciones son la base para plantear la solución.

Para enriquecer los datos del datasets CKAN semántico, se usará el dataset de la Dbpedia debido a que este posee una gran información de varios temas o recursos semánticos que tienen relación al dataset de estudio que es el CKAN semántico.

Este trabajo se basa en los conjuntos de datos reportados en (Eguiguren Palacios, 2019), y en (Eguiguren Palacios, 2019). Estos conjuntos de datos están descritos en tripletas y en diferentes idiomas según el conjunto de datos. Sin embargo, con el fin de ampliar el alcance de búsqueda a otros lenguajes, se incorporó un componente de traducción al inglés como parte de la solución propuesta, también una aproximación basada en NLP (procesamiento de lenguaje natural) y en NER (Reconocimiento de entidades nombradas).

Una vez hecho los pasos anteriores se procede a determinar que valores o entidades tienen mayor relación y con las que se puedan crear una relación entre un dataset A y un dataset B que describen o se relación con el contexto. De esta manera se puede tener la inferencia semántica entre entidades y las nuevas relaciones entre diferentes conjuntos de datos y de esta forma cumplir con los objetivos del trabajo de titulación.

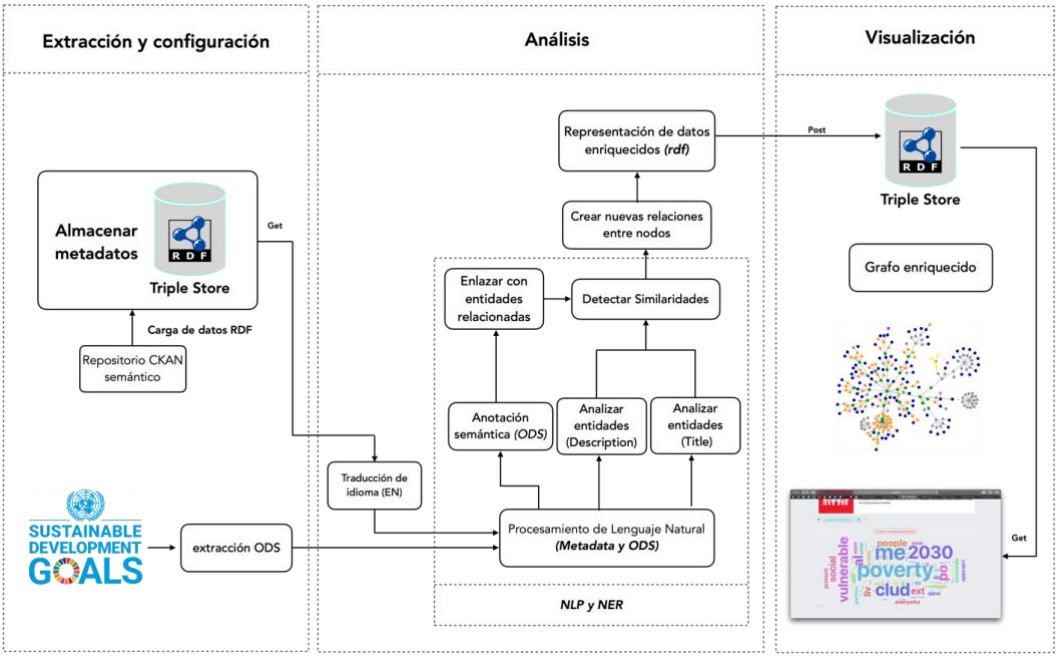
2.4 *Componentes arquitectónicos*

Para cumplir con los objetivos planteados para este Trabajo de Titulación es necesario crear una arquitectura que cumpla con cada uno de los elementos o características mencionadas en la propuesta de solución. Con esta arquitectura se busca establecer el proceso para el enriquecimiento de grafos semánticos a partir de otros conjuntos de datos abiertos y ver sus posibles relaciones con los Objetivos de Desarrollo Sostenible.

La **Figura 11** representa cada uno de los elementos esenciales para determinar el proceso del método de enriquecimiento de grafos y la asociación con los ODS. Esta representación arquitectónica muestra de manera general el funcionamiento del prototipo para el método de enriquecimiento de grafos con indicadores de ODS e inferencias semánticas con fuentes de datos abiertos.

Figura 11

Arquitectura de prototipo de método de enriquecimiento de grafos con ODS



Primero se realiza la representación de los ODS, lo que se busca es la extracción de cada unos de los objetivos, metas e indicadores. Una vez obtenida esta información se somete a un análisis de Procesamiento de Lenguaje Natural (NLP) con la finalidad de identificar palabras o partes de oraciones que permitan relacionar cada ODS a un recurso o entidad de otro conjunto de datos semánticos y a su vez crear un enlace entre ambos y estos sean almacenados en un triplestore.

De la misma manera se emplea el NLP para el conjunto de repositorios que es la base de estudio y así buscar que entidades semánticas que se relacionan entre si y de esta forma crear una nueva relación semántica.

Finalmente, lo que se busca es tener una nueva base de conocimiento que contenga no solo un conocimiento, sino varios y esta permita tener una mejor precisión en la toma de decisiones o mejor conocimiento en sus relaciones entre datos. También permite identificar las relaciones que cada dataset o cada recurso tiene con los Objetivos de Desarrollo Sostenible y de esta manera se pueda medir que el crecimiento la información enriquecida.

Capítulo tres

Desarrollo de la propuesta

Este capítulo especifica la implementación de la propuesta realizada en el capítulo anterior. Tomando en consideración cada una de las fases desarrolladas para el TT en base a la arquitectura planteada en la **Figura 11**. La sección 3.1 describe la configuración que se tiene que realizar con la herramienta de almacenamiento orientado a grafos GraphDB y la extracción de los ODS. En la sección 3.2 se describe el proceso de creación de enlaces entre los ODS y el dataset semántico. Finalmente, la sección 3.3 describe la visualización de los datos enriquecidos.

3.1 *Configuración y extracción*

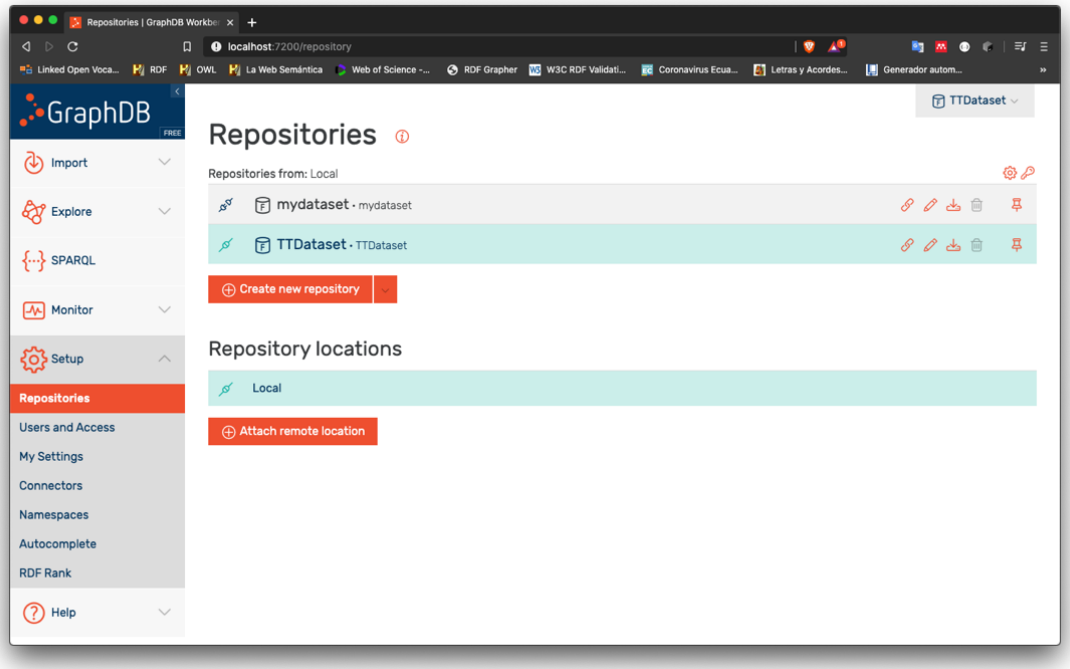
Previamente al almacenamiento de la información en GraphDB se tiene que crear el repositorio en el que se importará la información para obtener el tripleStore y de esta manera trabajar para el consumo de los datos en tiempo real como se plantea en la arquitectura de la **Figura 11** y la extracción de la información acerca de los objetivos, metas e indicadores de los ODS.

3.1.1 Creación de repositorios y carga de información

Dentro del entorno de GraphDB está Workbench, este permite de manera fácil y rápida la carga de la información. La **Figura 12** presenta el administrador para la creación del repositorio, y para la creación de este se da clic en la opción “create new repository” y luego se le asigna un ID o identificador único para el repositorio y continuar con los demás ítems que se necesita para la creación.

Figura 12

Administrador para la creación del repositorio



Una vez creado el repositorio se procede a la carga de la información. Esta sección se puede hacer mediante la carga directa desde la aplicación dando clic en “Upload RDF files” como se muestra en la **Figura 13**, o la otra forma mediante el uso del Endpoint de cada repositorio al que se quiera añadir la información, primero se debe obtener la URL del repositorio como se aprecia en la ***Figura 13*** y la configuración para la carga es la siguiente:

url -X POST --header "Content-Type:multipart/form-data" -F "config=@./config.ttl" "http://localhost:7200/<ID>/repositories"

Con este comando se debe reemplazar el ID de la URL por el nombre del repositorio que usará y también se debe especificar el formato serializable RDF del archivo.

Figura 13

Administrador para la carga de la información en el repositorio

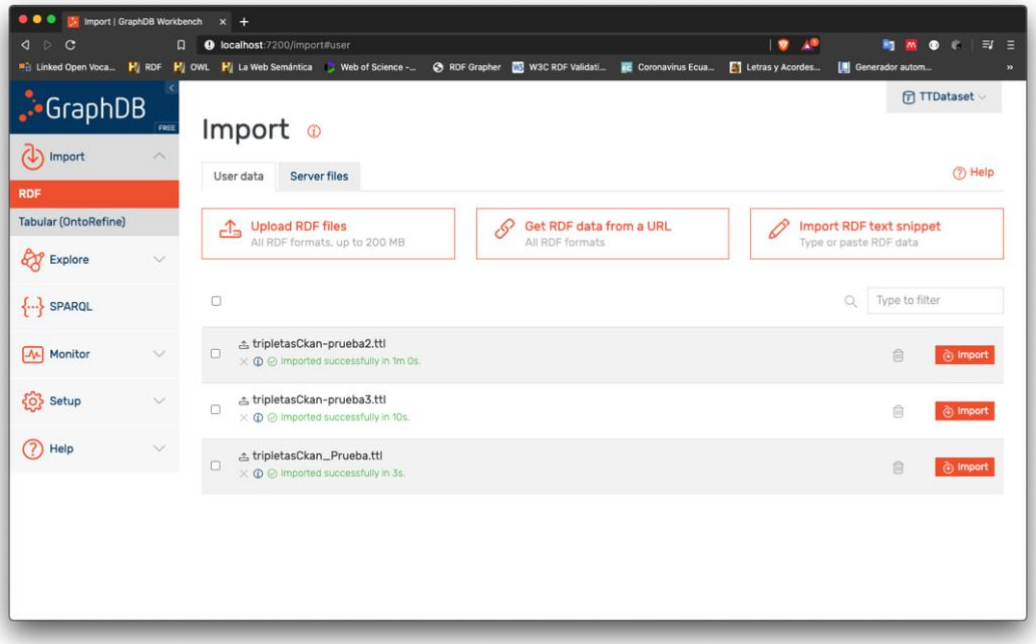
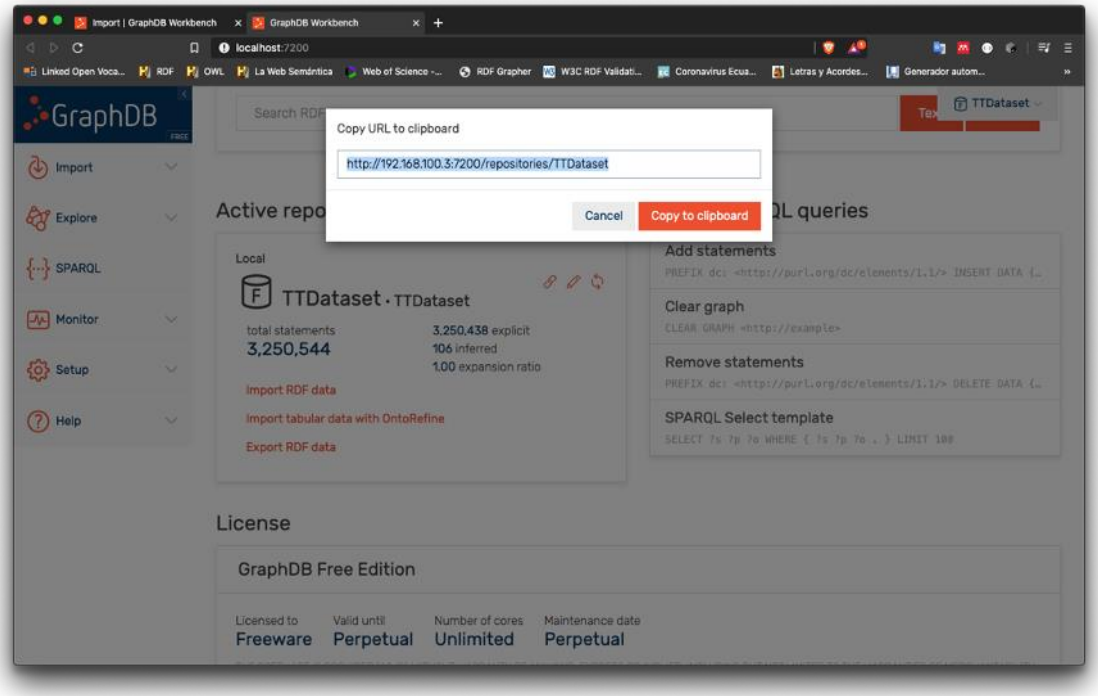


Figura 14

Administrador para obtener la URL del repositorio



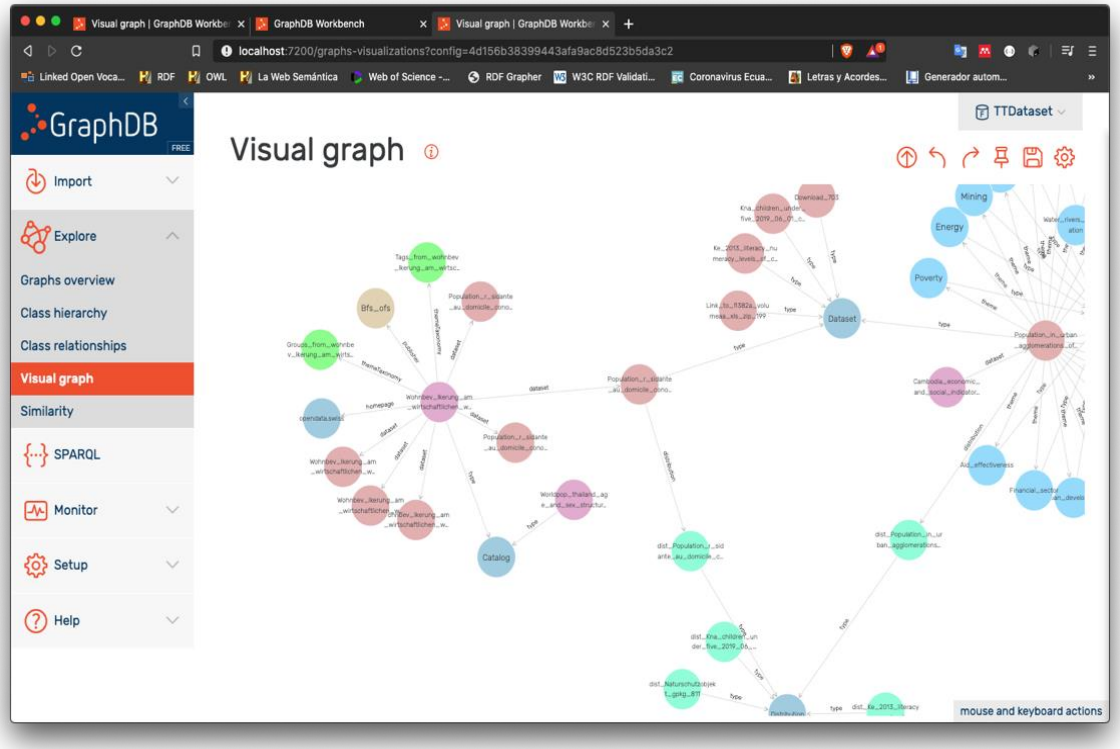
3.1.1.1 Exploración y visualización de datos

Esta plataforma GraphDB tiene muchas funcionalidades que son muy potentes, siendo así fácil de explorar los datos y sus herramientas que nos permiten visualizar instancias, clases, propiedades y sus relaciones a través de una jerarquía de clases del conjunto de datos que dan lugar al grafo de conocimiento.

Dentro de la aplicación Workbench también permite realizar la visualización del grafo con los datos existen del repositorio que se creo. Para la construcción del grafo se busca los recursos que están dentro del los datos almacenados o también se puede crear uno en base a una configuración previa. También se busca a través de una instancia poder explorar todas las relaciones y dando clic se pude ver más información de la instancia principal y de esta manera generar el grafo que se muestra en la **Figura 15**.

Figura 15

Visualizador de grafos



3.1.1.2 Jerarquía y relaciones de clases

Como se viene apreciando en las gráficas anteriores que GraphDB es una plataforma muy robusta que permite realizar múltiples cosas y en este caso se explica la jerarquía de las clases en base al número de instancias almacenadas como se lo presenta en la **Figura 16**. Dado este grafo se observa las clases hijas y padres.

La **Figura 17** presenta un diagrama que muestra las relaciones y la cantidad de enlaces que están asociadas a las clases. La **Figura 17** muestra el número total de las relaciones existentes a las clases como ejemplo en número de relaciones asociadas a dcat:Dataset.

Figura 16

Visualizador de grafos

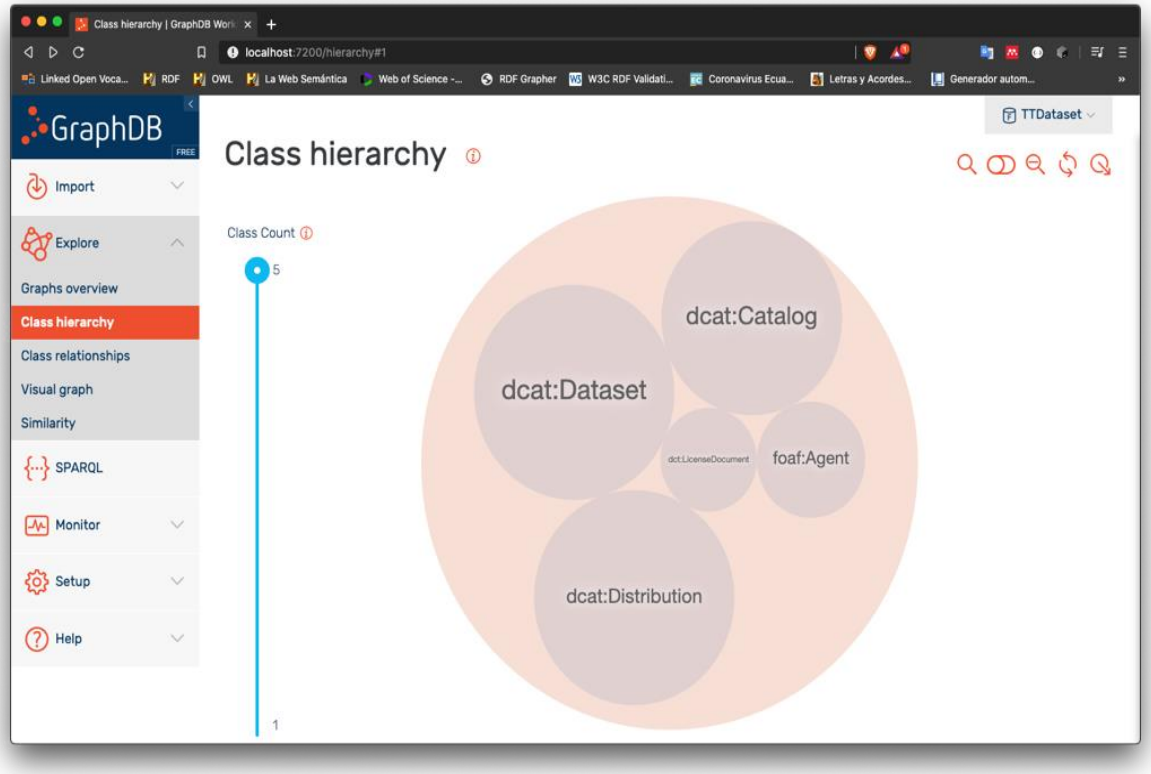
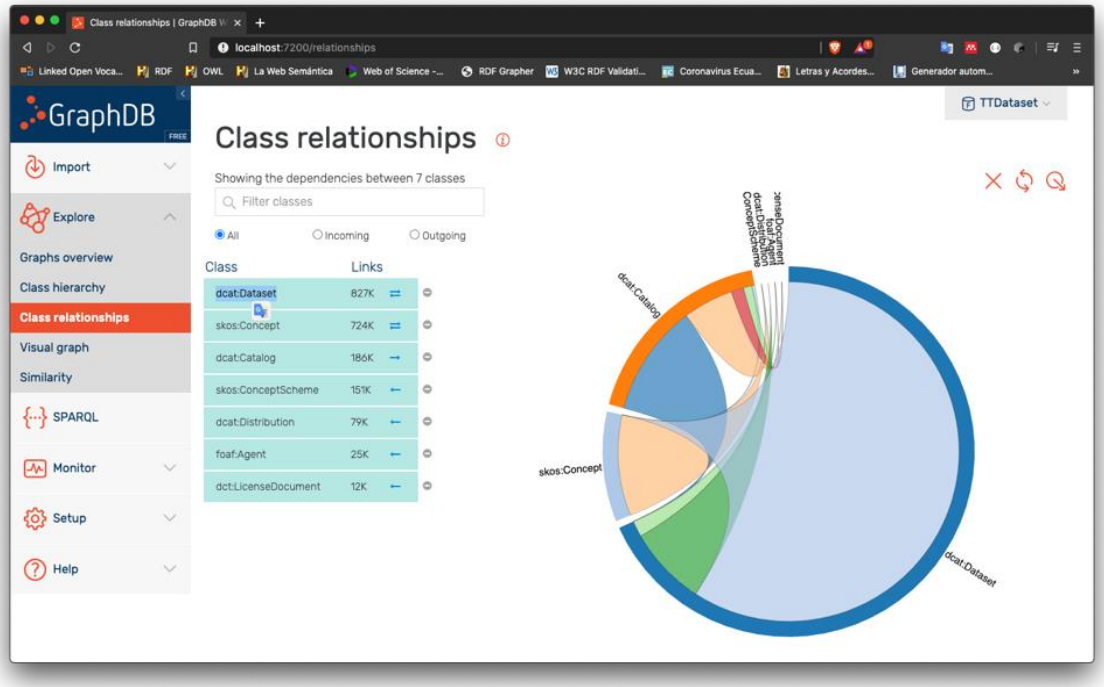


Figura 17

Diagrama de relaciones de clases

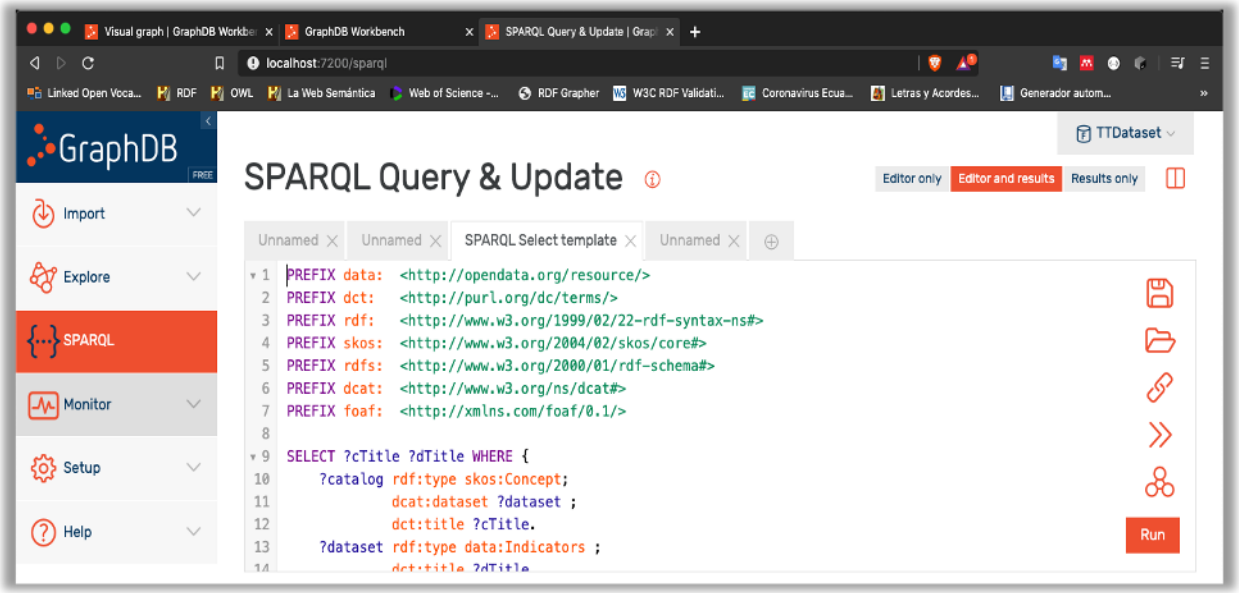


3.1.1.3 Consulta de datos

Finalmente se tiene las consultas de los datos que se encuentran en el repositorio de GraphDB, estos pueden ser usados para realizar las consultas mediante el editar SPARQL en Workbench. La **Figura 18** muestra un ejemplo de como se ven las consultas en el entorno. Las consultas SPARQL son súper sencillas y fácil de hacer con todas las funciones que da Workbench.

Figura 18

Diagrama de consultas Sparql



Nota: (Ontotext, 2020)

3.1.2 Anotaciones semánticas de los objetivos de desarrollo sostenibles

Los datos son usables si están en un formato abierto y con una estructura. La estructura que poseen los recursos de los Objetivos de Desarrollo Sostenibles (ODS) no son lo suficientemente abiertos para su análisis y reutilización. De esta manera se busca que estos datos se los pueda procesar y luego crear un enlace entre ellos poder obtener las anotaciones semánticas de los ODS.

3.1.2.1 Extracción de información de ODS

Para obtener la información de los ODS acerca de los objetivos, metas e indicadores se realizó una extracción de datos desde el sitio oficial en inglés en el cual se tomó en cuenta todas las características mencionadas para su extracción o también existe otra forma de extracción que es mediante el API que posee Naciones Unidades[14] y para esto se usa los Endpoints[15] que ofrece la ONU.

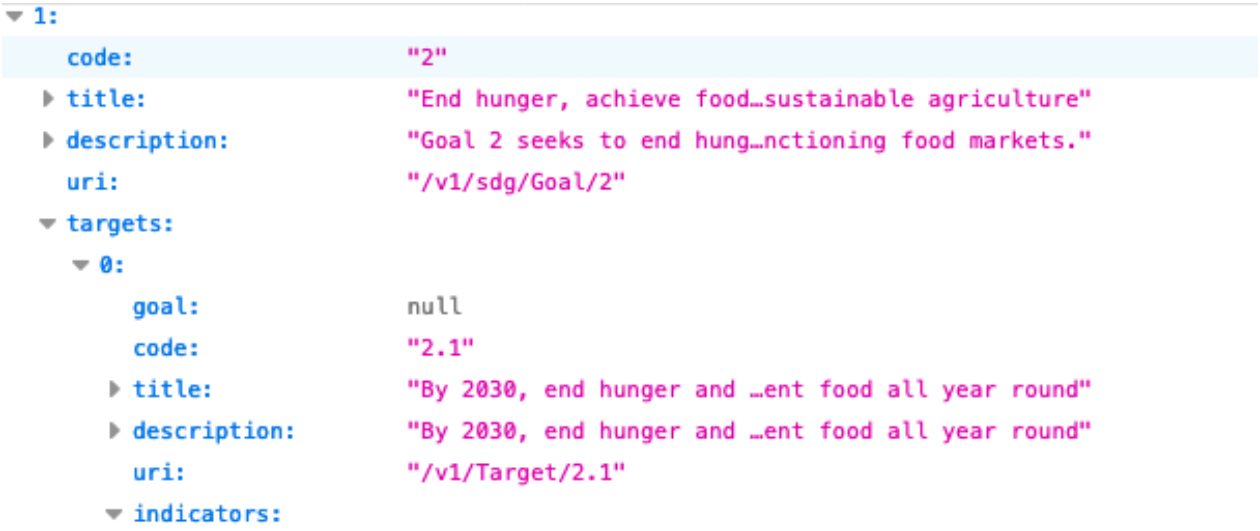
Con cualquiera de estas dos formas de extracción se obtuvo los 17 objetivos y cada uno de estos con sus características titulo, objetivos, metas e indicadores y así saber la

[14] https://unstats.un.org/SDGAPI/swagger/
[15] *https://unstats.un.org/SDGAPI/v1/sdg/Goal/List?includechildren=true*

finalidad de cada uno de estos objetivos. Si la extracción es mediante la API, el cual permite distribuir las solicitudes en diferentes parámetros para obtener una mejor información y características de los ODS y el resultado obtenido es cada objetivo con sus metas e indicadores como se aprecia en la **Figura *19*.**

Figura 19

Resultado de consulta mediante la API de la ONU



Nota. Tomado de ODS ONU [Fotografía], por (ONU, 2019)

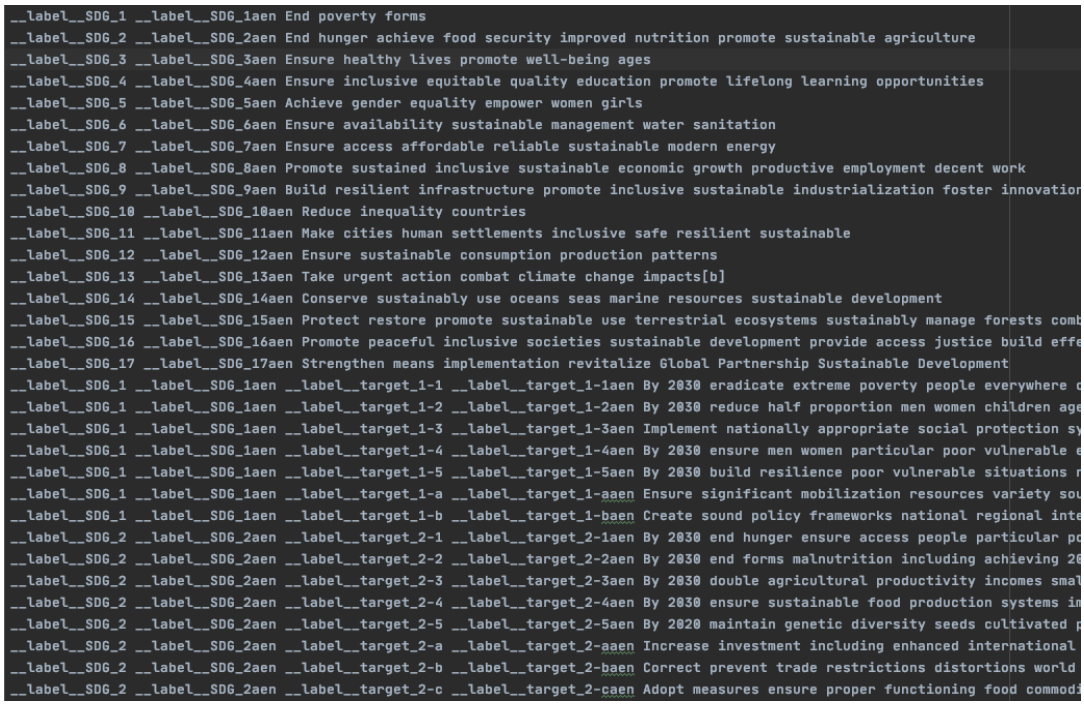
*3.1.3 **Transformación de información de ODS***

La transformación de estos datos se realiza para obtener una representación de conocimiento de cada unos de los objetivos, metas e indicadores. En esta fase se realizó dos procesos, uno para la creación de base de conocimiento esta se encuentra en inglés (**Figura 20**) para obtener esta información se realizó un proceso semiautomático el cual consiste en obtener la información manualmente y guardar un archivo; después mediante un algoritmo eliminar las Stop Word existentes en el texto obtenido. Después de esto se tiene correctamente la base de conocimiento. Se procede a realizar el procesamiento de lenguaje natural (NLP[16]) usando el motor de fasttext el cual permite interpretar esta información y así mejorar el conocimiento obtenido y buscar posibles relaciones con otra información.

[16] Natural Language Processing

Figura 20

Base de conocimiento de los ODS



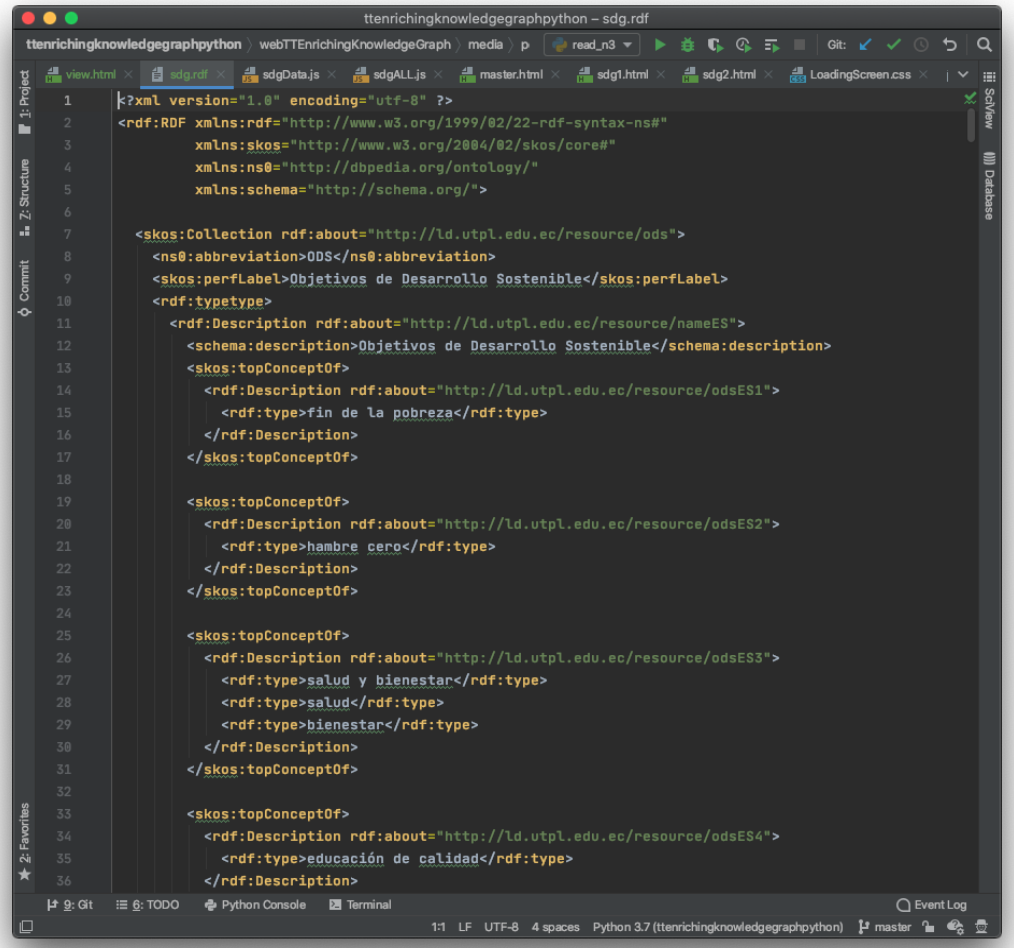
Lo que se busca el NLP es poder identificar del dataset principal y analizar que recursos tienen una probabilidad de ser semejantes y de esta manera se puedan relacionar los ODS a los demás recursos semánticos.

Y en la segunda parte una vez identificadas las entidades más importantes que se esta relacionando a cada uno de los 17 objetivos, 33 metas y 21 indicadores. Para esto se usó la librería rdfLib la cual permite analizar las entidades resultantes del NLP y buscar las posibles relaciones a cada ODS el cual contiene un parser/serializador a RDF/XML. Este resultado del uso de esta librería es un conjunto de tripletas serializados en RDF/XML, el resultado esta almacenado en un archivo denominado SDG.rdf (**Figura 21**).

Luego de obtener el análisis por cada ODS, se transformó y consolidó con la taxonomía de SKOS y para relacionar sus clases se usó DCAT y Schema.

Figura 21

Resultado de análisis de los ODS



Con estas tripletas en formato RDF/XML se puede identificar fácilmente los objetivos, metas e indicadores que han sido representados por el nombre y características que se agruparon con todos los conceptos. De tal manera se puede usar el análisis y buscar si existe alguna relación con otros textos. Finalmente, estos datos pueden ser usados para buscar si existe la probabilidad de tener relaciones entre los ODS y un recurso, y estos permitirán identificar como están incrementando el conocimiento de los ODS con relación a los dataset.

3.2 **Análisis - creación de enlaces entre ods y dataset semánticos**

La creación de enlaces entre el Dataset ckan Semántico (dataset base) con otros datasets externos y los ODS, de esta manera obtener una nueva representar semántica del conjunto de datos ckan.

Inicialmente para la creación de estos enlaces se busca identificar cuales son las entidades principales y cuales son literales, con estos se pueden interpretar que cada uno de los recursos que serán procesados en la parte del procesamiento de lenguaje natural (NLP). Con esto se busca interpretar el contenido de las entidades, de los literales y determinar la similitud entre los conjuntos de datos de tal manera que permitan obtener nuevos enlaces entre ellos. En los siguientes ítems se describe todos los pasos para la creación de los enlaces.

3.2.1 ***Traducción***

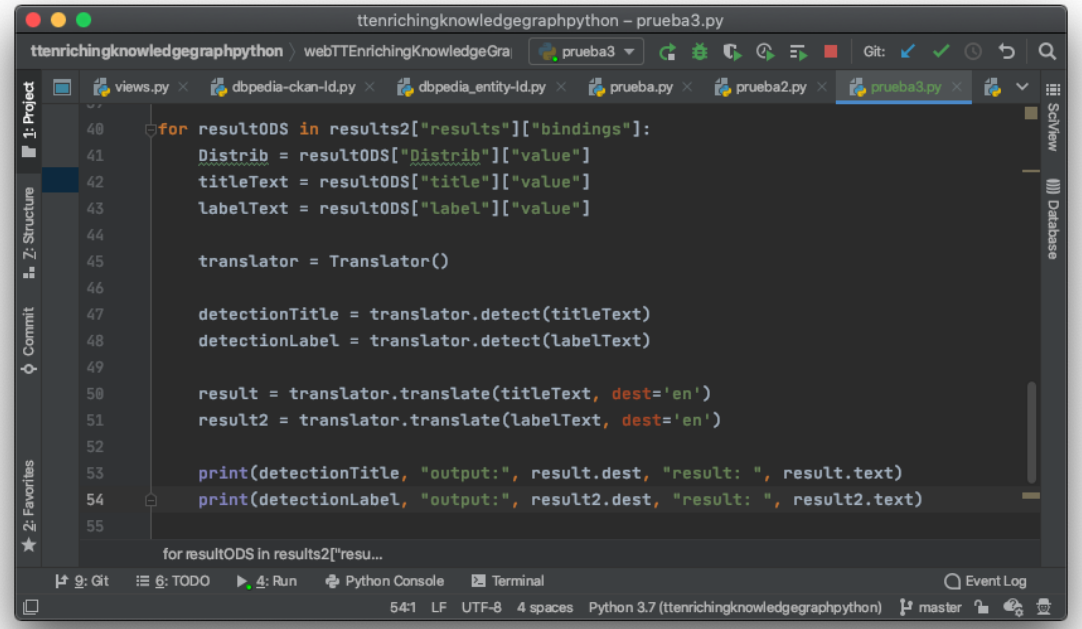
Conceptualmente, la traducción es la comunicación del significado de un texto en el idioma de origen, por medio de un texto equivalente en el idioma de destino. En este trabajo debido a que los recursos provienen desde diferentes fuentes de información y en varios idiomas, a cada recurso se aplicó un proceso de traducción. Se ha realizado un proceso de análisis y interpretación de idioma con el framework de google traductor (googletrans[17]). Esta herramienta permite realizar una interpretación del texto que entra y así determinar el idioma en que se encuentra, dando como resultado la similitud de idioma y traducción. Como parte del alcance de este trabajo se usó la traducción al idioma inglés, debido a que este idioma es el más usado en los datos utilizados.

En la **Figura 22** se aprecia un fragmento del código de la parte de la traducción que se realiza previo al análisis de traducción.

[17] https://pypi.org/project/googletrans/

Figura 22

Traducción de texto con google traductor api



3.2.2 ***NER y NLP***

El Reconocimiento de Entidades Nombradas (NER[18]) permite la extracción de información que busca localizar y clasificar en categorías predefinidas, tales como: personas, organizaciones, lugares y las entidades nombradas encontradas en un texto. Lo que se busca es identificar diversas características que tiene la información que se esta enriqueciendo de tal manera que en el NLP sea más rápido y el porcentaje de similitud sea mayor. Y por otra parte lo que busca el procesamiento de lenguaje natural es el reconocimiento de entidades en el cual quiere idear búsquedas lingüísticas específicas y detectar similitudes entre entidades.

[18] Reconocimiento de Entidades Nombradas.

3.2.2.1 FastText

Teniendo estas premisas, para el NLP en primera instancia se uso fasttext como herramienta para este proceso, el cual busca analizar los datos y determinar cual es el rango de similitud entre un texto con base de conocimiento que fue entrenada previamente y aplicado el NER, en la **Figura 20** se presenta un fragmento de la base de conocimiento de los ODS[19].

Con la base de conocimiento previamente estructurada y revisada se procede a realizar el entrenamiento con la finalidad de obtener un rango de similitud para esto en la documentación de fasttext muestra que se debe usar para wordNgrams=1 y se tiene que trabajar con un rango de aprendizaje entre 0.5 a 0.6, cuando se usa wordNgrams=2 el rango debe ir entre 0.7 a 1.0 como mejor modelo de entrenamiento. En este trabajo se usó el wordNgrams=1 con el fin de obtener un acercamiento más real a la información obtenida y el entrenamiento con 0.5 en el rango de similitud siendo esto la media para determinar si ese valor es similar o no. La **Figura *23*** muestra un fragmento del código en Python del proceso que se usó para el entrenamiento supervisado, para esto se realizó la herramienta de fasttext.

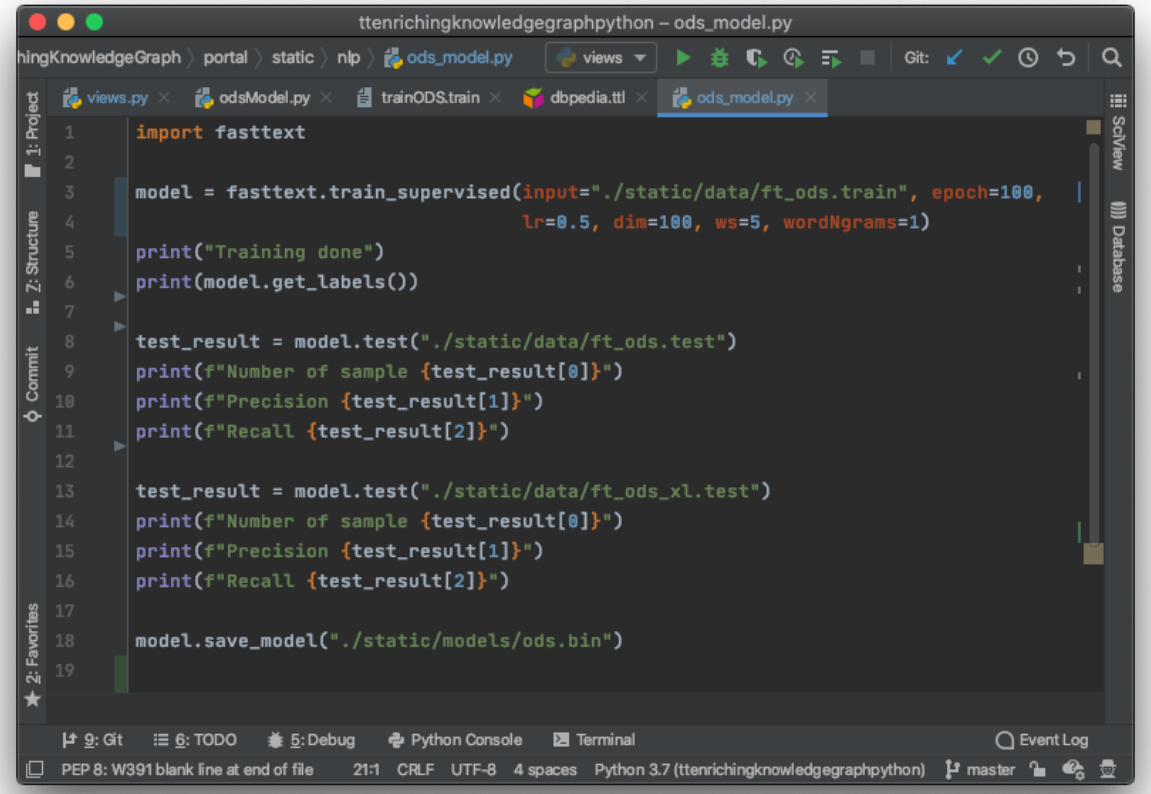
En el **Apéndice 1** se presenta un fragmento del modelo resultante del entrenamiento.

Una vez realizado este proceso nos da como resultado un modelo, el cual será la base en la que dando los parámetros en texto los analizar y tomará de referencia para determinar a que se puede relacionar o que pueda presentar alguna similitud que sea igual o mayor a 0.5 hasta 1, con esto se busca que el resultado de la interpretación sea más preciso.

[19] Objetivos de Desarrollo Sostenibles

Figura 23

Proceso de entrenamiento de los ODS



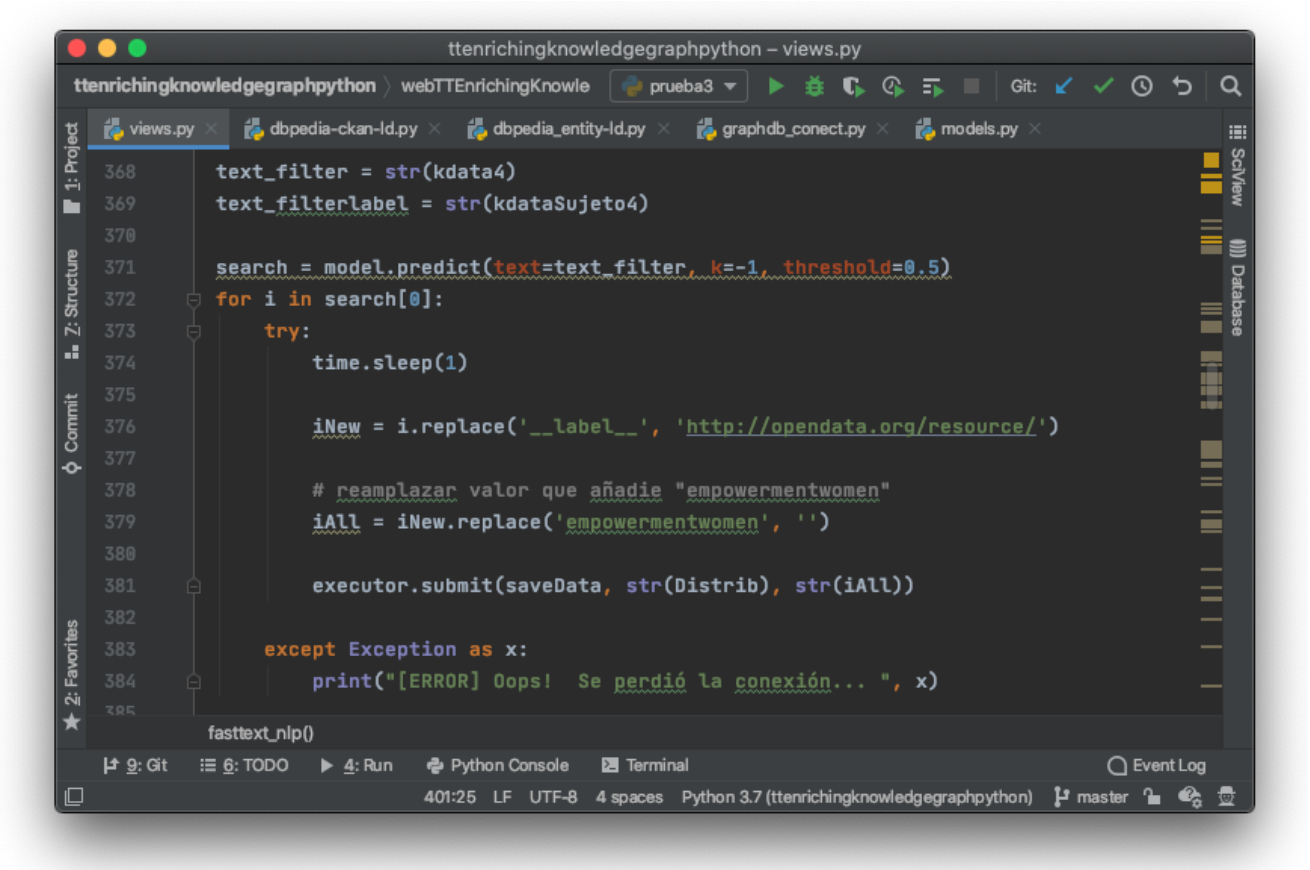
Para la detección de las similitudes entre los recursos y los ODS, se realiza un análisis previo con fasttext el cual me permite interpretar el contexto y por consiguiente obtener una interpretación mejor. Antes de empezar la detección de similitudes se hizo la traducción del contenido con la finalidad de que todos los recursos sean analizados y de esta manera obtener una mayor calidad de los resultados.

En el análisis se tiene que cargar primero el modelo entrenado el cual servirá para el proceso de NLP supervisado, lo que se realiza en esta parte es adecuar la información que le llega al analizador, luego de esto se crea un variable donde se almacena el resultado que es un arreglo, el cual tiene como salida la predicción, el label (url) y el porcentaje de similitud siendo este mayor a 0.5.

La ***Figura 24*** presenta un fragmento del código que permite la detección y obtener de la similitud del recurso con un ODS.

Figura 24

Proceso de predicción con fastText



Una vez obtenido el resultado del análisis se procede a interpretar el resultado. Lo que se busca con esto es la interpretación del contenido y saber que información es la que me sirve o no, y así enviar a un método que se lo denomina *saveData* el cual realiza el proceso de almacenar la información directamente en la base de datos semántica en este caso en GraphDB en un formato de tripletas.

Luego de haber sido guardada la información esta ya puede ser consultada y consumida por una API para luego su explotación. Lo que se busca con esta herramienta de fasttext es que se pueda analizar cada recurso sea con el label, title o description y como resultado nos devuelva a que Objetivos de Desarrollo Sostenible se relaciona y de esta manera enriquecer el grafo principal con una mayor información.

En el **Apéndice 2** se presenta un fragmento del resultado final del enriquecimiento con fasttext.

3.2.2.2 DBpedia spotlight

Primero la DBpedia es un proyecto para la extracción de datos desde la Wikipedia y ellos proponen una versión de la web semántica en el contenido. En el proceso de estructuración del sitio hacia la web semántica, Dbpedia desarrollo una API que la denominan dbpedia-spotlight el cual es una herramienta para anotar automáticamente las menciones de los recursos de DBpedia en un texto. Y así proporcionar una solución para vincular fuentes de información no estructuradas a la nube Linked Open Data a través de DBpedia.

Esta herramienta funciona con un enfoque de cuatro pasos establecidos por la DBpedia. DBpedia Spotlight realiza la extracción de entidades nombradas, incluyendo la detección de entidades y la resolución de nombres. También se puede utilizar para el reconocimiento de entidades nombradas, entre otras tareas de extracción de información.

Sabiendo los beneficios de este framework se procede a realizar la primera actividad que es el NLP y NER, en este caso usando spacy y de esta existe una extensión en la que también esta trabajando dbpedia-spotlight la cual se denomina *spacy_dbpedia_spotlight*, con esta biblioteca de software de código abierto de procesamiento avanzado de lenguaje natural, esta busca la interpretación sintáctica y semántica de los parámetros que se le pase desde los recursos que se analizó.

Lo que se busca al usar spacy es que el procesamiento de texto en nuestro caso el texto del label, title o Description. Estos textos se encuentran sin procesar, la mayoría de las palabras son raros y es común para las palabras que se ven completamente diferente a significar casi lo mismo. Las mismas palabras en un orden diferente pueden significar algo completamente diferente. Incluso dividir el texto en unidades útiles similares, estos pueden resultar difícil en muchos idiomas.

Si bien es posible resolver algunos problemas comenzando solo con los caracteres en bruto, generalmente es mejor usar el conocimiento lingüístico para agregar información útil.

Para poder identificar las palabras que están relacionadas entre si, spacy realiza una tokenización, con el fin de que esta pueda analizar y etiquetar un documento determinado. Aquí es donde entra el modelo estadístico o en otras palabras el modelo de etiquetas previamente entrenado, que permite a spacy hacer una predicción y el resultado que es mayor a 0.5 siendo esta la media que permita etiquetar que palabra se relaciona con la probabilidad del contexto. Un modelo consta de datos binarios y en diferentes idiomas.

Las anotaciones lingüísticas están disponibles como atributos del token en la herramienta. Entonces, para obtener la representación de cadena legible de un atributo necesitamos agregar un guión bajo "_" a su nombre como por ejemplo:

- *token.lemma_*
- *token.pos_*
- *token.tag_*
- *token.dep_*
- *token.shape_*
- *token. kb_id_*
- *token.is_alpha*
- *token.is_stop*

Cada uno de estos traen un resultado como:

- Texto: el texto de la palabra original.
- Lemma: La forma básica de la palabra.
- POS: La etiqueta simple de parte del discurso de Universal POS tags.
- Tag: la etiqueta detallada de la parte del discurso.
- Dep: Dependencia sintáctica, es decir, la relación entre tokens.

- Shape: la forma de la palabra: mayúsculas, puntuación, dígitos.
- Kb id: URL resultado de las palabras similares.
- is alpha: ¿el token es un carácter alfa?
- is stop: ¿El token forma parte de una lista de *stop*, es decir, las palabras más comunes del idioma?

Entonces, la mayoría de las etiquetas se ven bastantes abstractas y varían entre los idiomas. De la misma manera trabaja la librería de dbpedia-spotlight pero en este caso usando la herramienta de spacy ayudará a la interpretación del contexto del contenido entrante lo que hace *spacy_dbpedia_spotlight* trabajo con las anotaciones lingüísticas y con reconocimiento de entidad nombrada.

Lo que spacy hace con el NER es reconocer una entidad nombrada, el cual es un "objeto del mundo real" al que se le asigna un nombre, por ejemplo, una persona, un país, un producto, el título de un libro y en nuestro caso la URI del recurso que se encuentre estrechamente relacionada al recurso analizado para esto se usa la etiqueta *ent.kb_id_.* Lo que se busca es poder identificar las entidades que se encuentran en el texto y una vez reconocidas, estas pueden ser interpretadas y ver que información contiene y con esto crear el enriquecimiento del recurso con el recurso de la DBpedia. A continuación se describe un ejemplo para tener una mayor claridad del proceso.

De todos los recursos que se tiene se tomó un recurso de ejemplo, el cual dice "*Distribution of: Costs of proceedings - Australia*", este habla sobre costos de distribución en Austria, de este recurso se toma la **Description**, la **Figura *25*** muestra el texto de entrada sin procesar, en este caso será el texto de entrada que recibe spacy para su análisis e interpretación.

Dentro de la documentación de spacy especifica las variables que serán usadas para la interpretación de texto de entrada entre estas se encuentra *ent.text* (texto resultado), *ent.label_* (Valor de análisis sea verbo o algún resultado que nos de spacy) y *ent.kb_id_* (token resultado) los cuales son los principales recursos que voy a tomar en

cuenta del resultado. Con estas variables de análisis se activa el método para que comience a trabajar e interpretar el texto.

La **Figura *26*** presenta el valor resultante de manera visual, para obtener esto se realiza un mapeo con las anotaciones lingüísticas y el reconocimiento de entidades nombradas que se encuentre en el texto. Con estas nuevas etiquetas ya se puede saber que información me sirve para la creación de la nueva relación o el enriquecimiento de los datos y este nuevo recurso para si presentación se lo puede identificar con la URL de la DBpedia en el resultado y también lo presenta un como *dbr* que se hace referencia la URI “http://dbpedia.org/resource/NameResource” o “http://dbpedia.org/page/NameResource” , y de esta manera se obtenga el resultado dbr y enviado para su almacenamiento. La **Figura *27*** muestra un fragmento del código que nos permite realizar este proceso.

En el **Apéndice 3** se presenta un fragmento del resultado final del enriquecimiento usando spacy Dbpedia-spotlight.

Figura 25

Fragmento de texto de entrada ejemplo

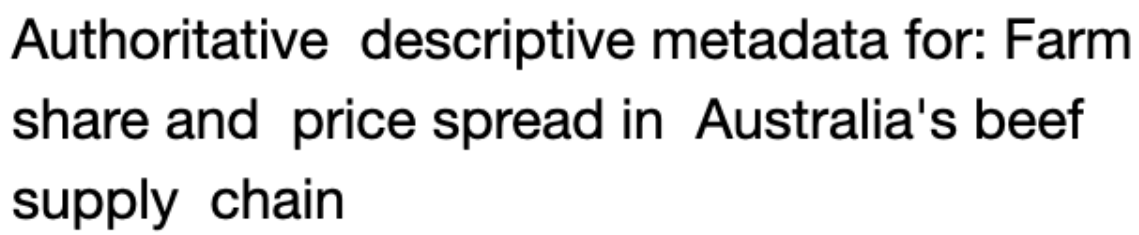


Figura 26

Ejemplo de interpretación del NER y el NLP con spacy Dbpedia-spotlight

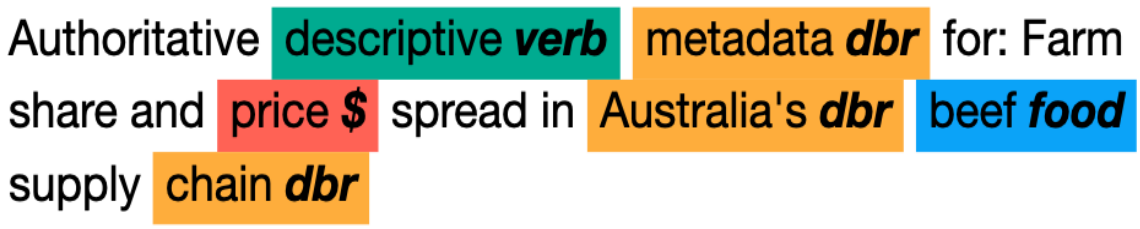
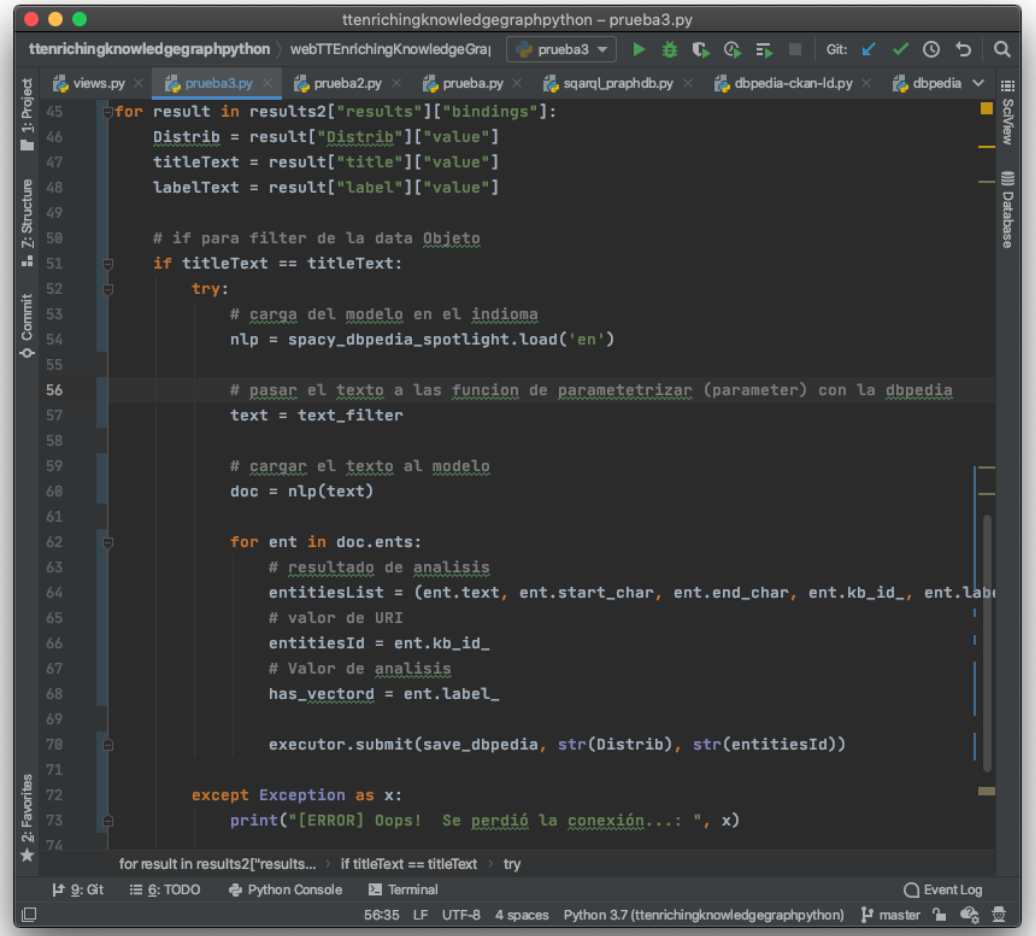


Figura 27

Proceso de predicción con dbpedia-spotlight



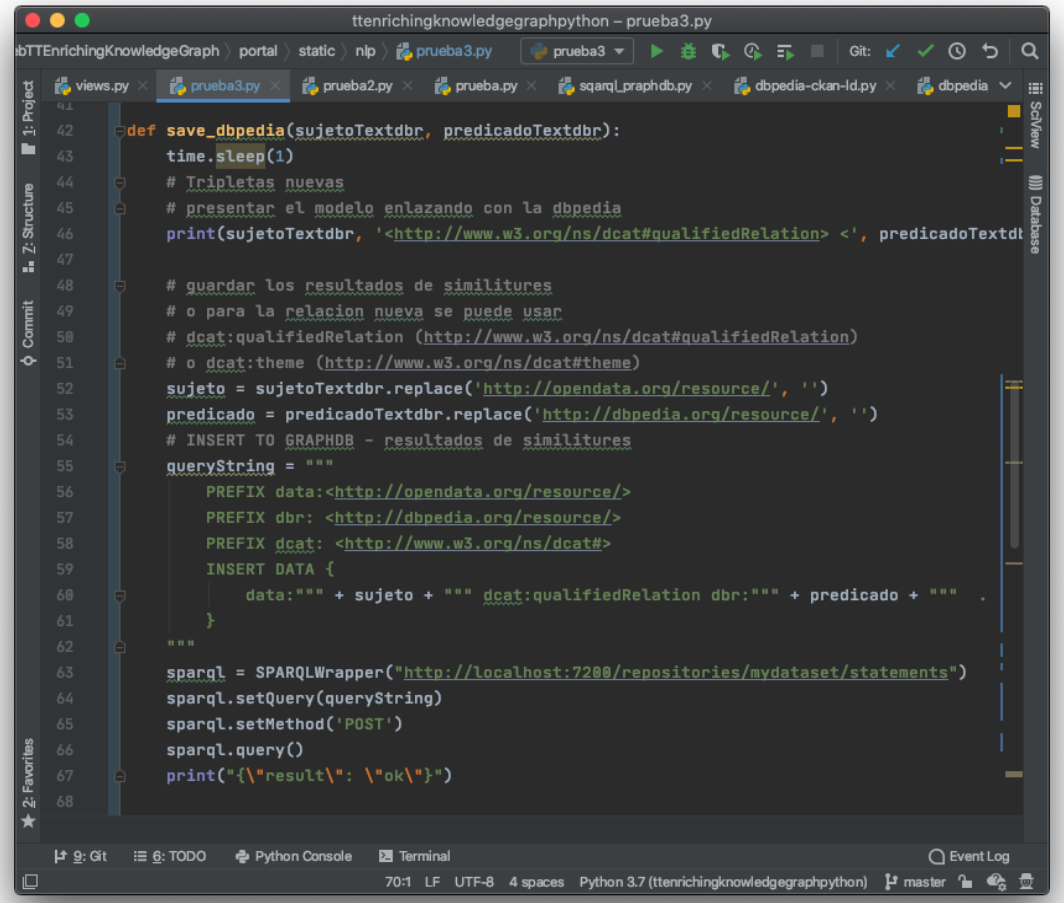
*3.2.3 **Creación de nuevas relaciones***

Para la creación o enlazamiento de nuevas entidades, se creó un método en el que se envía los parámetros resultantes de los análisis realizados con las dos herramientas. En este método se crea un proceso de inserción de datos directamente a la base de datos semántica, en este caso GraphDB, para esto se crea una consulta SPARQL de inserción para que se pueda almacenar el resultado final. Los parámetros que recibe son la URI del recurso interpretado más en resultado obtenido del análisis es decir la URI del nuevo

recurso que se asocia a este. La **Figura *28*** presenta el proceso para almacenar la información en el triple store.

Figura 28

Método de almacenamiento en GraphDB



Esto se aplica para los dos procesos de análisis sea para los ODS o la Dbpedia y adicionalmente se lo almacena en una base de datos SQL para revisar el contenido que esta siendo almacenado permitiendo analizar si lo obtenido está acorde al contenido analizado o no.

En el **Apéndice 5** se presenta de manera visual un fragmento de todos los datos que fueron insertados en GraphDB.

3.3 **Visualización**

La visualización de los datos enriquecidos se presenta en la sección 4.2 con los casos de aplicación en la detección de similitudes del dataset y los ODS. El resultado de la visualización se presenta con el sitio web piloto desarrollando.

Capítulo Cuatro

Pruebas y Resultados

En este capítulo se especifica un caso de aplicación para la detección de posibles relaciones con diferentes umbrales de entrenamiento en el proceso de NLP. Además, se evalúa el sitio piloto web. La sección 4.1 describe el caso de aplicación para el proceso de creación de enlaces entre los ODS y el dataset. Y en la sección 4.2 describe el resultado final de TT que es la visualización de los datos enriquecidos.

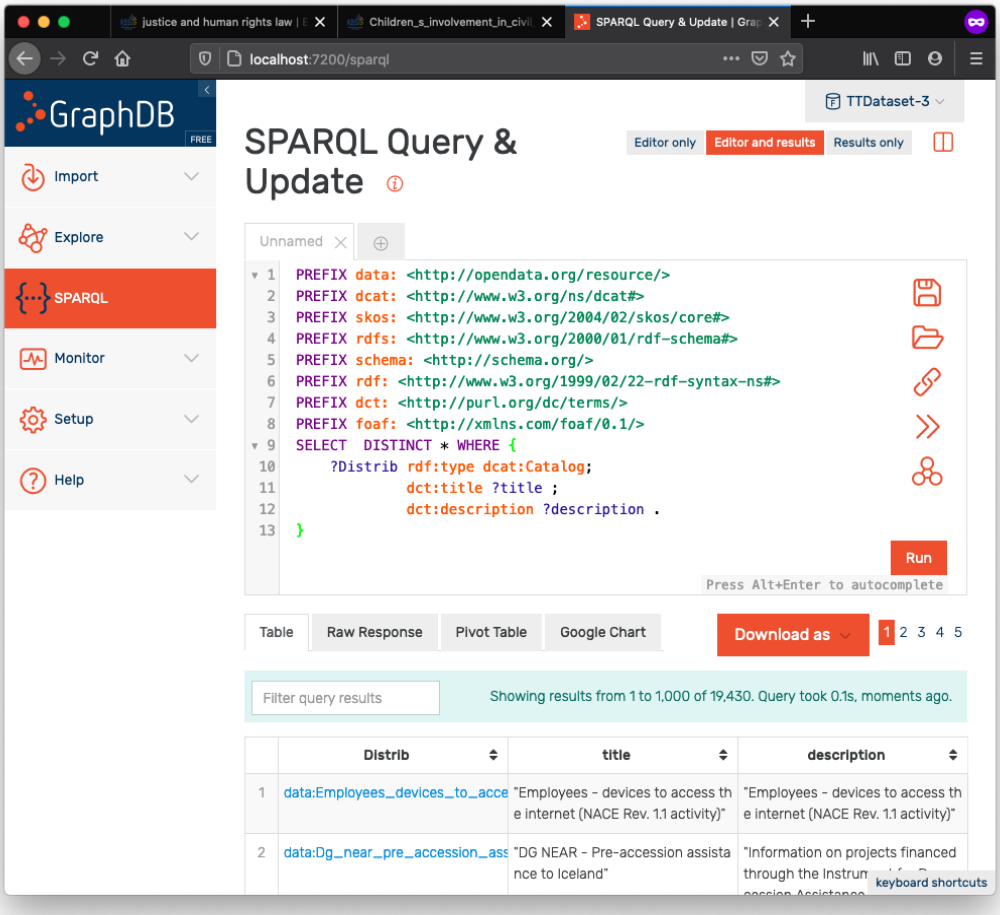
4.1 Casos de aplicación

Esta sección pretende mostrar varios ejemplos de algunos datasets que se tiene y presentar el proceso de enriquecimiento de los datos CKAN con la Dbpedia y con los Objetivos de Desarrollo Sostenibles.

Para el proceso de enriquecimiento con fasttext y la Dbpedia se trabajar con los 20 primeros dataset. A estos se les aplicará todos los procesos mencionados en las secciones anteriores como el NER y el NLP. Como primeros pasos se realiza una consulta básica con SPARQL para obtener los valores que van a ser analizados, en este caso serian el título y la descripción de recurso como se muestra en la **Figura 29**.

Figura 29

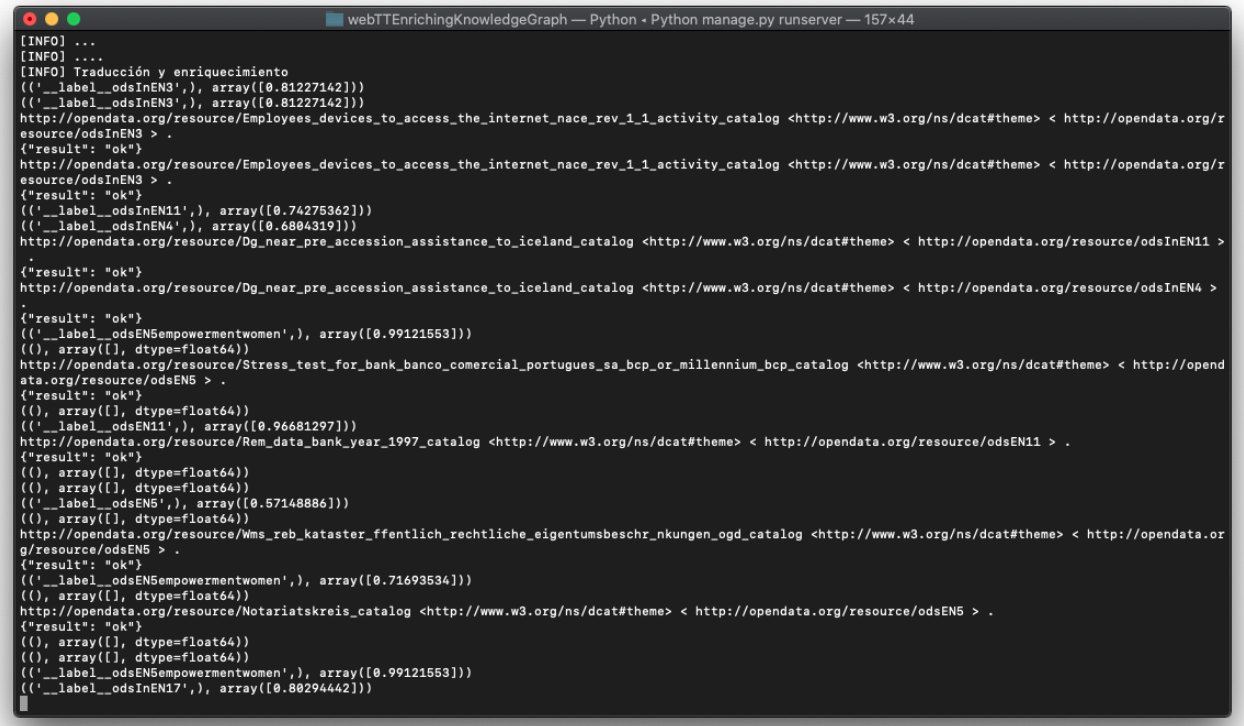
Consulta SPARQL para enriquecimiento



Usando el método de fasttext para la predicción de similitud que se presenta en la **Figura 24**, en este caso se trabaja con un rango de 0,5 y con ese rango obtendremos que algunas relaciones sean semejantes o no. La ***Figura 30*** muestra la ejecución del enriquecimiento de los datasets.

Figura 30

Captura del proceso de enriquecimiento con un rango de 0.5 de similitud

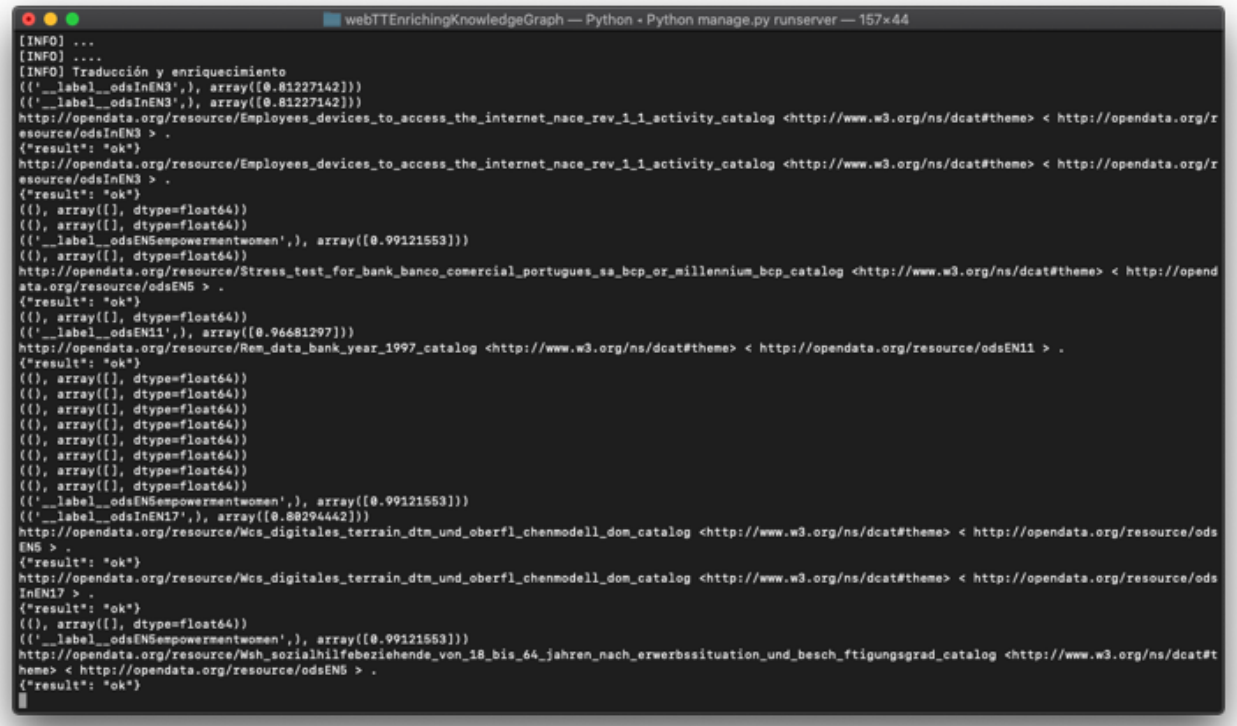


Con el rango de similitud de 0.5 se obtiene 28 nuevas tripletas de relación de los 20 recursos analizados. Se observa que un recurso puede tener más de dos relaciones alineadas a los ODS. Para esto se requiere otro análisis o interpretación humana con el fin de obtener una información más precisa o que tenga relación semánticamente.

A continuación, se ejecuta el mismo proceso de enriquecimiento con fasttext pero en este caso el rango de similitud se incrementa a 0.8 para obtener una mayor precisión.

Figura 31

Captura del proceso de enriquecimiento con un rango de 0.8 de similitud



Luego de análisis con el grado de similitud mayor, en este caso mayor a 0.8 de los 20 recursos analizados se implementan nuevas tripletas de relación. Esto nos asegura que las nuevas relaciones son más precisas para el conjunto de datos. Para asegurar esta similitud se requiere un análisis de un humano para verificar si las relaciones nuevas son las correctas en las que el valor de similitud se encuentre entre 0.8 hasta 0.89, y si esta es mayor o igual a 0.9 las relaciones son 95% más confiables y no requieren un análisis humano.

Una vez obtenido el resultado del análisis se envía al método save-Data, siendo este el que permita almacenar las nuevas relaciones obtenidas.

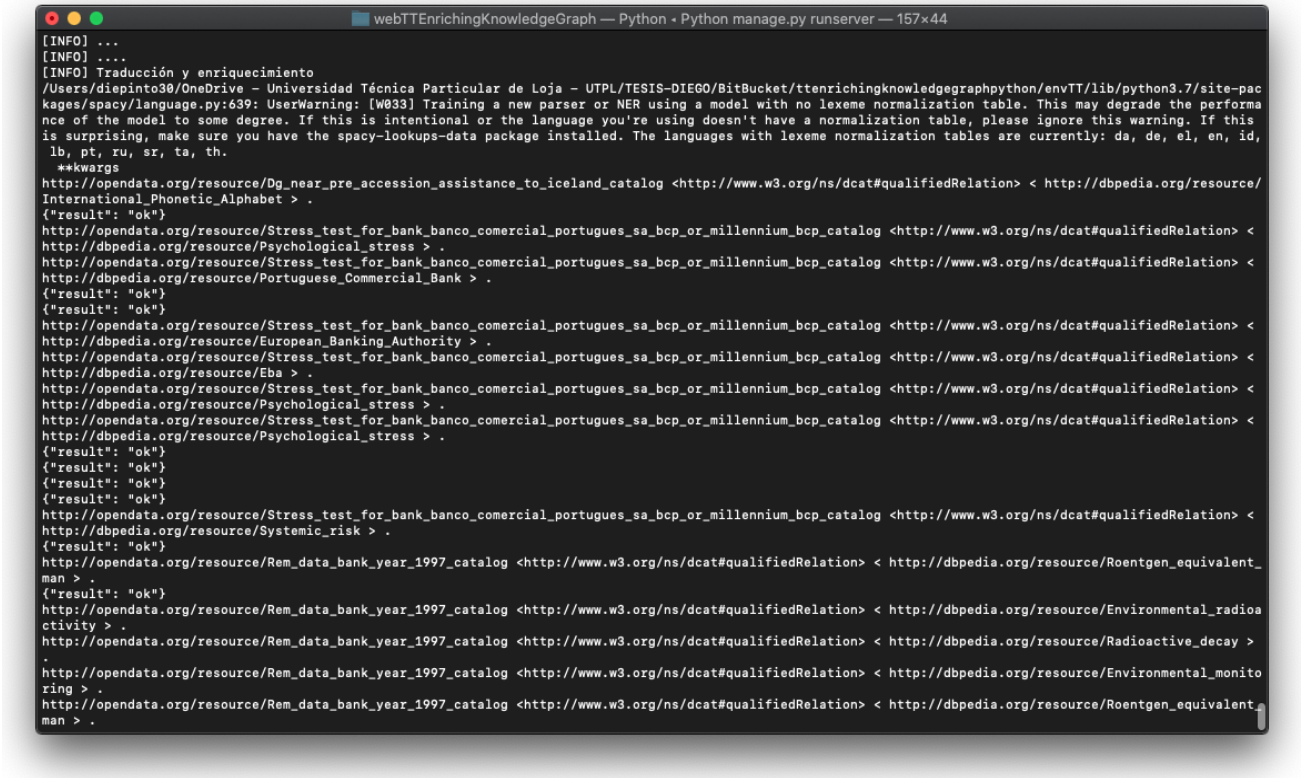
Con estos 20 datasets analizados con el rango de 0.5 se obtuvo un gran número de relaciones nuevas, en cambio con el rango de 0.8 las nuevas relaciones fueron menores debido a que el análisis tiene una mejor interpretación y por ende los resultados serán más

exactos en las relaciones de los ODS. Con el valor de similitud que sea mayor los datos serán más precisas y confiables, en cambio con el rango menor se requiere un análisis de una persona para validar cierta información.

El enriquecimiento de datos con *spacy_dbpedia_spotlight* se aplicará el en mismo método que se muestra en la **Figura *27***. En este modelo spacy define el rango de similitud acorde a la expresión de las entidades nombradas y las relaciones lingüísticas que obtiene. La **Figura 32** muestra el proceso de enriquecimiento que esta realizando *Dbpedia-spotlight.*

Figura 32

Captura del proceso de enriquecimiento de la Dbpedia-spotlight



4.2 **Visualización de los resultados obtenidos del enriquecimiento datos**

La visualización del grafo enriquecido se puede presentar de varias maneras como el mismo grafo general, pero esto sería muy difícil de entender a simple vista. Se ha pensado en crear un sitio web piloto que permitirá visualizar la información y la interpretación de los datos para el usuario de mejor manera. A continuación, se detalla el uso de las interfaces o vistas de usuario creadas el sitio web piloto y los resultados obtenidos del proceso de implementación de la fase de desarrollo.

Se tiene que tomar en cuenta que para la visualización se realizó consultas SPARQL en tiempo real para recuperar la información directamente desde el triple Store y así esta sea consumida por un api/json, siendo este formato uno de los mas fáciles para usar, mantener y consumir la información que devuelve la consulta.

Dentro de esta API se toman en cuenta los valores que se quiere recibir para la presentación en el JSON. Mediante este formato se hace el consumo con JavaScript (js) para su presentación en el sitio web.

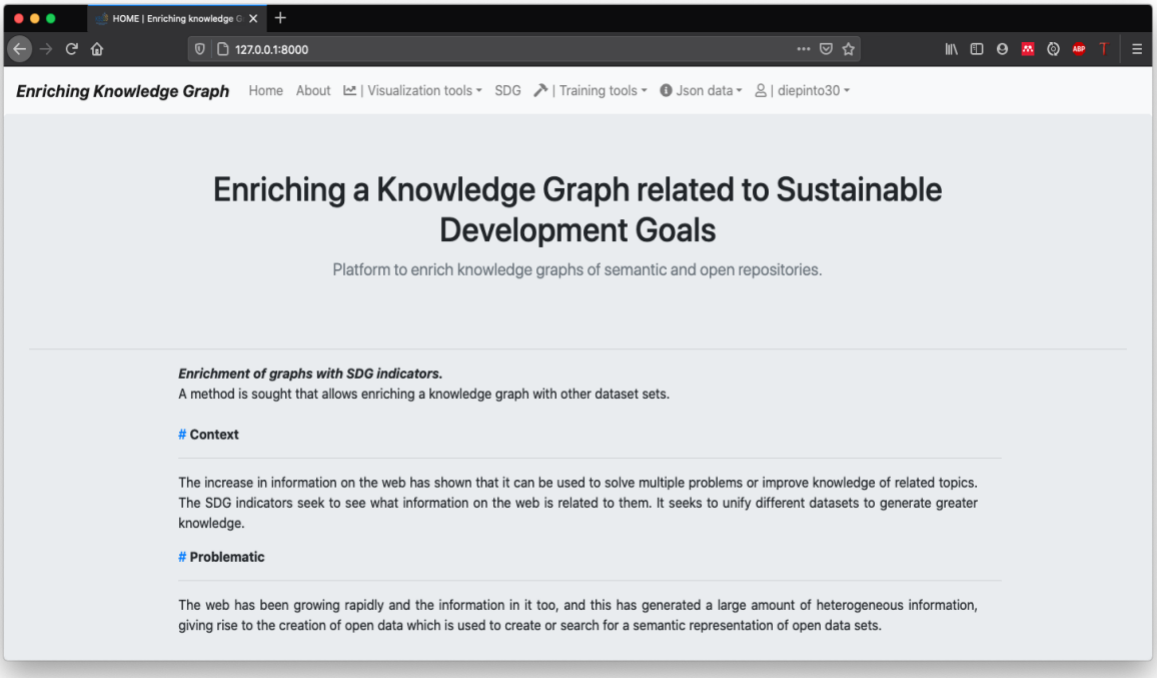
En el **Apéndice 6** se presenta un fragmento de una de las consultas SPARQL que se hizo para obtener los recursos que se esta usando. Y en el **Apéndice 7** se muestra una de las api/json, esta api se la proceso para obtener la información precisa y estructurar el contenido que se muestra en la misma, siendo así usado para el consumo con js.

4.2.1 ***Sitio web***

En las siguientes figuras se describen las características y las acciones que cada una de las interfaces que se han creado para el sitio web piloto. En la **Figura 33** muestra el diseño del home o página principal de la aplicación web. En esta se presenta los objetivos de la aplicación, la estructura general de arquitectura en la que se basa para el desarrollo y las opciones del menú de navegación que tiene el sitio.

Figura 33

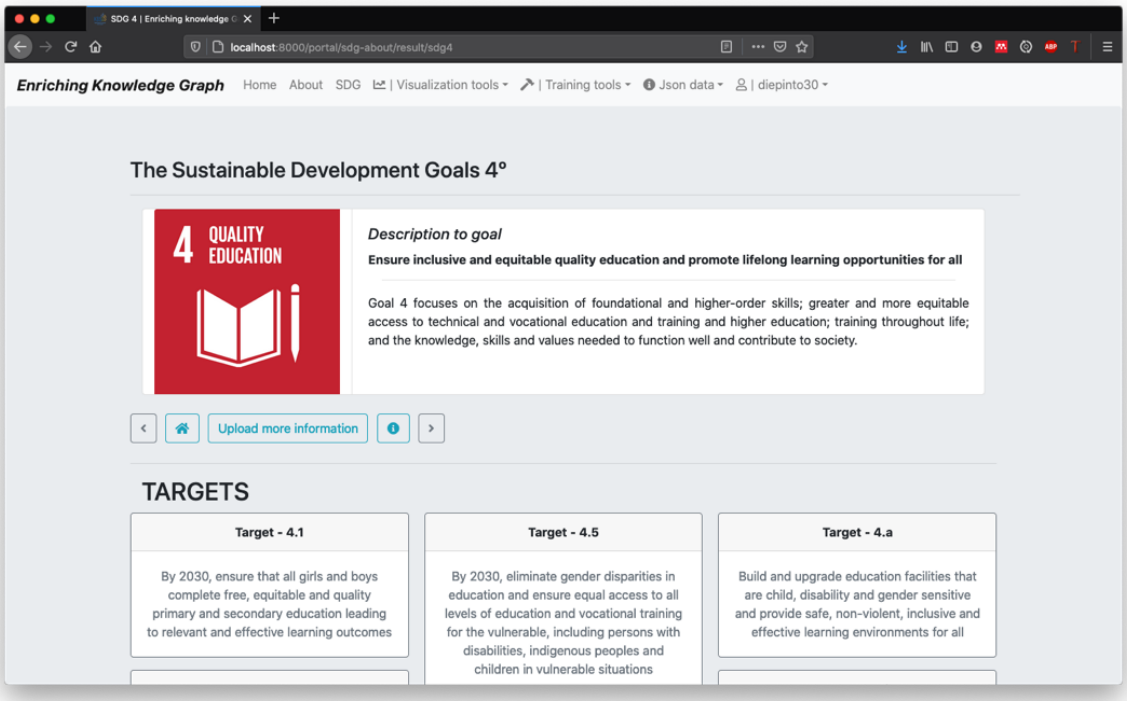
Página principal de la aplicación Web



Dentro del menú de navegación se tiene una opción que dice “SDG”, este punto se presenta cada uno los de los objetivos de desarrollo de sostenible. Al hacer clic en el icono del objetivo nos lleva a una página en la cual nos permite ver las características de sus indicadores y metas que estos objetivos quieren realizar. La **Figura 34** presenta este apartado.

Figura 34

Página de los ODS, en la aplicación Web



En este sitio se presenta una opción de “Upload more information” y presenta gráficas de información que los recursos que estén relacionados a cada objetivo en una Word Cloud y más gráficas que permitan ver los datos relacionados y enriquecidos. La **Figura 35** se muestra los resultados obtenidos y estos se los muestra en una nube de palabras que presenta los conceptos de los ODS, en esta gráfica se busca ver la mayor incidencia de palabras mencionadas en este ODS.

También en la misma sección tiene una opción que presenta una tabla con buscador (**Figura 36**), el cual permite ver todos los recursos que se encuentren relacionados a cada Objetivo de Desarrollo Sostenible y de esta manera poder apreciar más información enriquecida.

Esta tabla permite redirigir a otra página a visualizar toda la información del recurso asociado. La **Figura 36** se muestra un fragmento esta tabla.

Figura 35

Nube de palabras generada a partir de los conceptos de los ODS

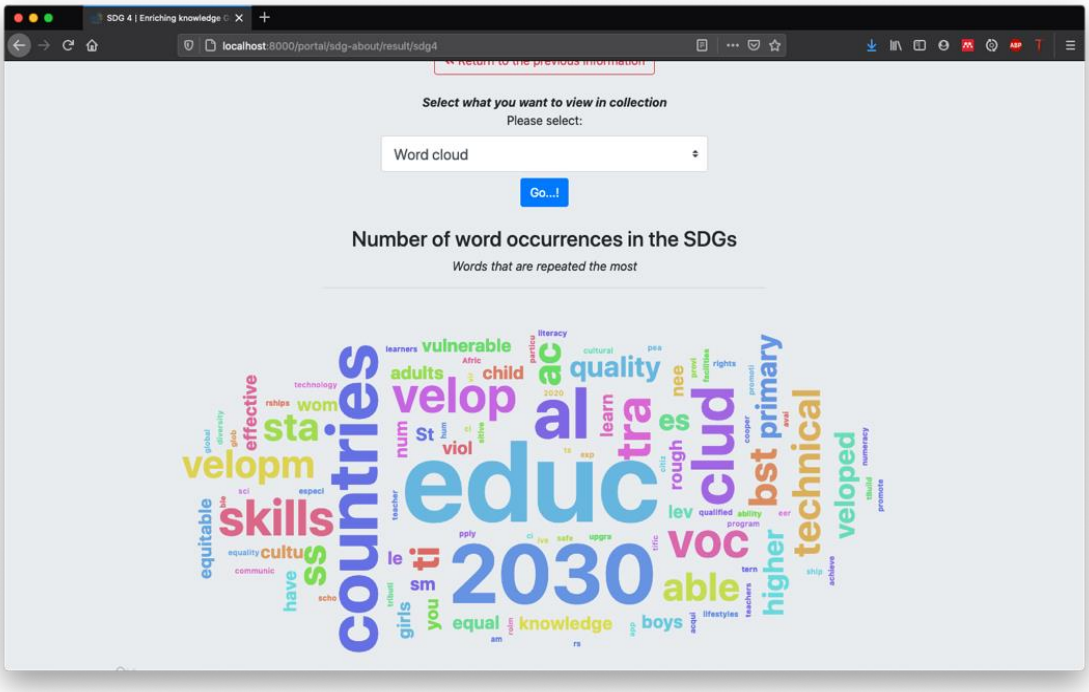
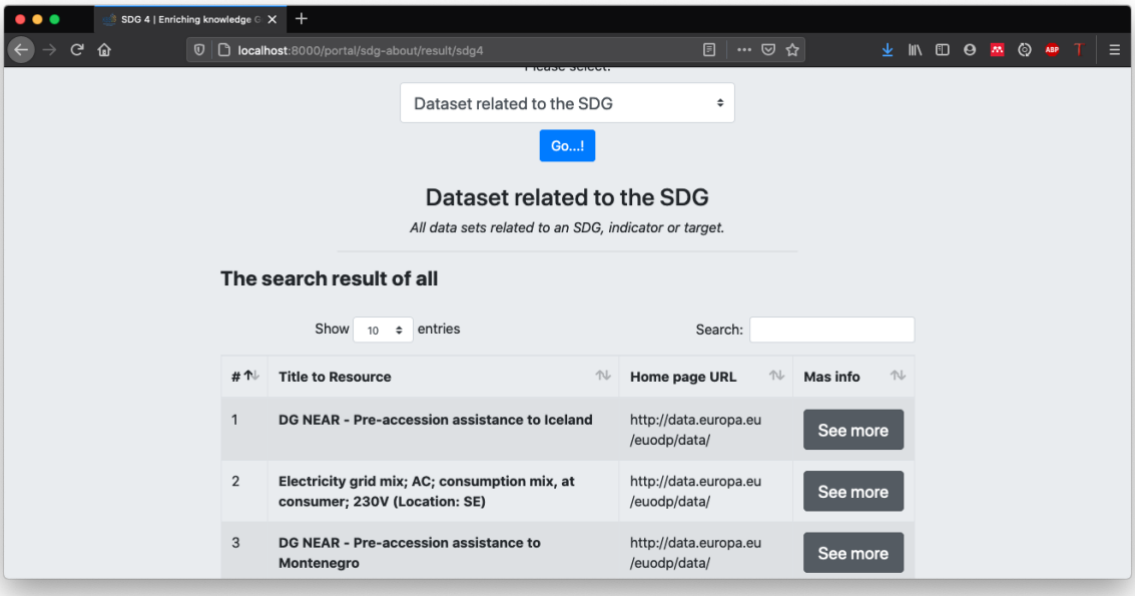


Figura 36

Recursos asociados a los ODS



Otra visualización que se tiene es con un buscador y diferentes gráficas de resumen de los datos. Si nos dirigimos al menú de navegación existe un apartado de “Visualization tools” y en esta existe dos herramientas de visualización de los datos enriquecidos, a continuación de las describe a cada una de ellas.

4.2.1.1 Search

En este apartado se muestra un buscador (**Figura 37**), el cual permite realizar consultas sobre la información de los recursos en este caso sobre el titulo, descripción y sobre los datos enriquecidos del recurso, en esta búsqueda se aplicará la traducción en caso de que el texto de la búsqueda sea diferente del idioma base que es el inglés y adicionalmente la corrección del texto si esta mal escrito dando lugar a la corrección de texto de búsqueda. El resultado obtenido de la búsqueda nos llevará a una nueva pestaña denominada browser.

El browser busca presentar todos los recursos que han sido ya enriquecidos; para esto se usa un API del resultado de la consulta en tiempo real hacia el tripleStore, de esta manera el resultado es consumido con Ajax para crear la tabla resultante del contenido.

La **Figura 38** presenta el resultado obtenido de la búsqueda, también cabe mencionar que en la parte derecha se encuentran algunos filtros que se aplican a los resultados existen de la búsqueda, los filtros implementados en el resultado son selección por página de destino, por ODS, por idioma del recurso y un buscador que ayuda filtrar la información resultado según lo que busco.

Figura 37

Buscador de palabras relacionadas

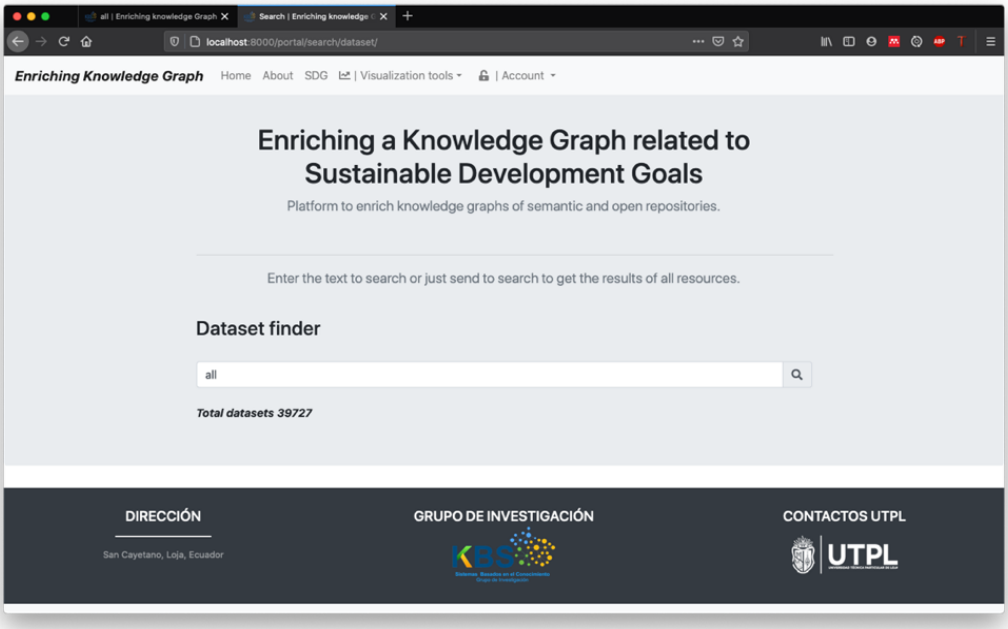
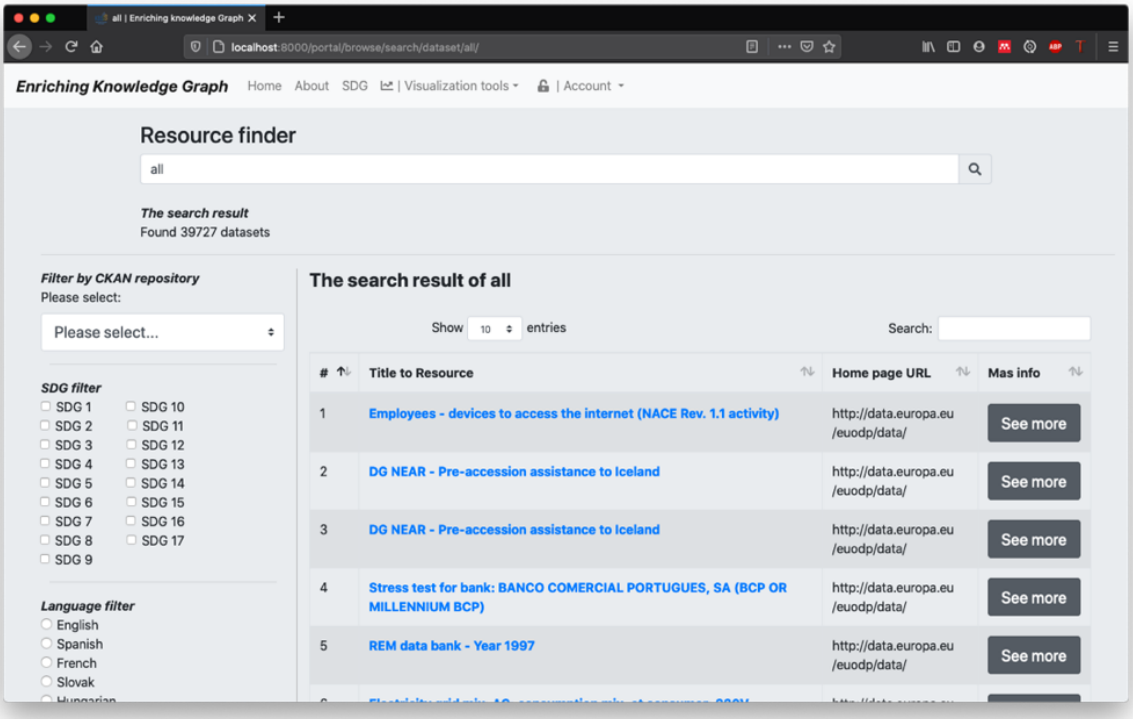


Figura 38

Buscador de recursos enriquecidos

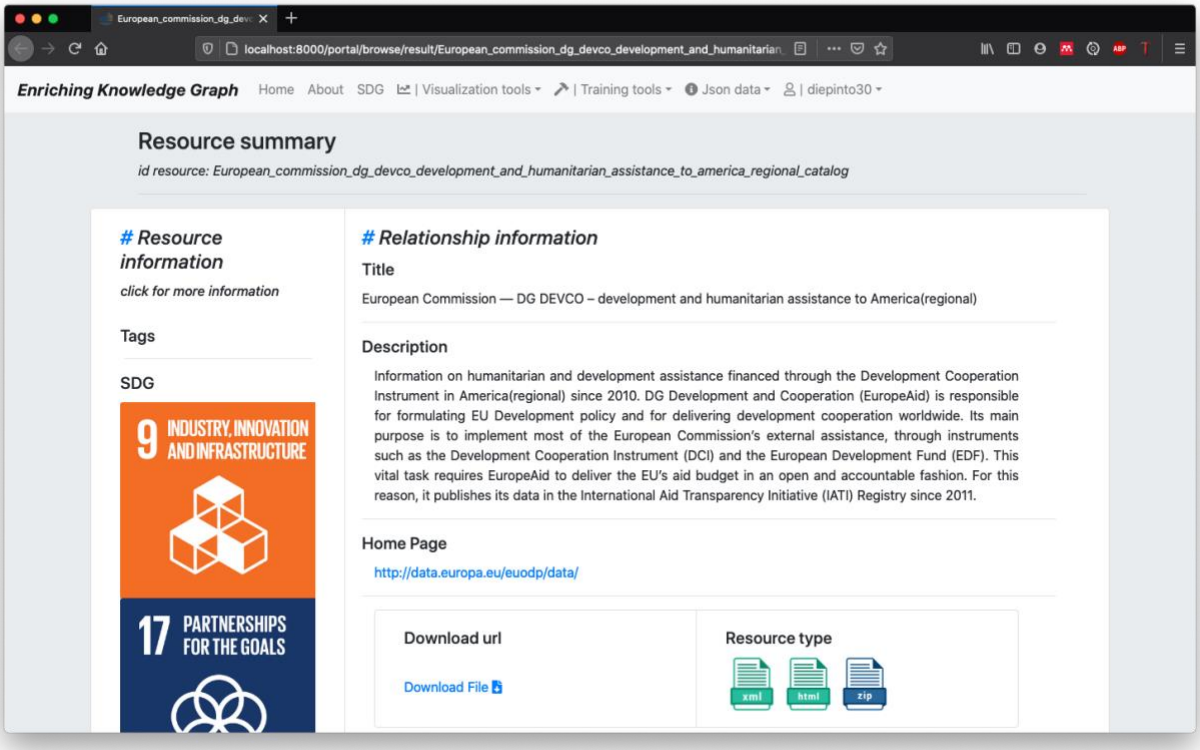


En los apéndices del 8 al 11 se presenta la evolución del buscador desde su primera versión hasta la versión 4 que incluye un buscador y más flitros de búsqueda.

Una vez que se tiene visible todos los datos, en el botón de “see more” o en el mismo título nos lleva a una nueva pestaña en que se cargan los datos del recurso al que se hizo clic, en la **Figura 39** se muestra el contenido resultante.

Figura 39

Resultado de búsqueda en nueva pestaña



4.2.1.2 Graphics Dataset New

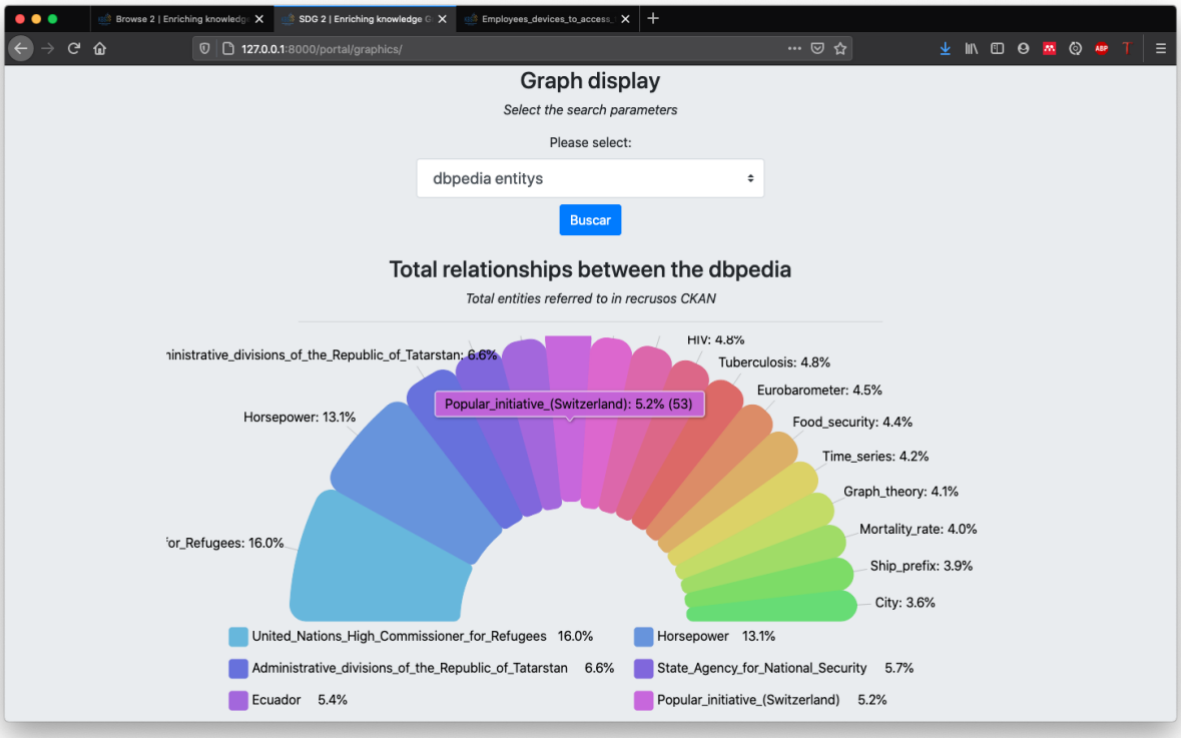
En esta opción se representa con unas gráficas varios resultados generales relacionados con las nuevas características del proceso de enriquecimiento de datos.

La **Figura 40** presenta un conteo total de las 25 entidades más nombras de la DBpedia que se encuentra en cada uno de los recursos. Esta gráfica denominada semicírculo de sectores, busca representar las etiquetas que sean más usadas en los

recursos y determinar de que temáticas están abordando estos recursos que fueron analizados y mapeados.

Figura 40

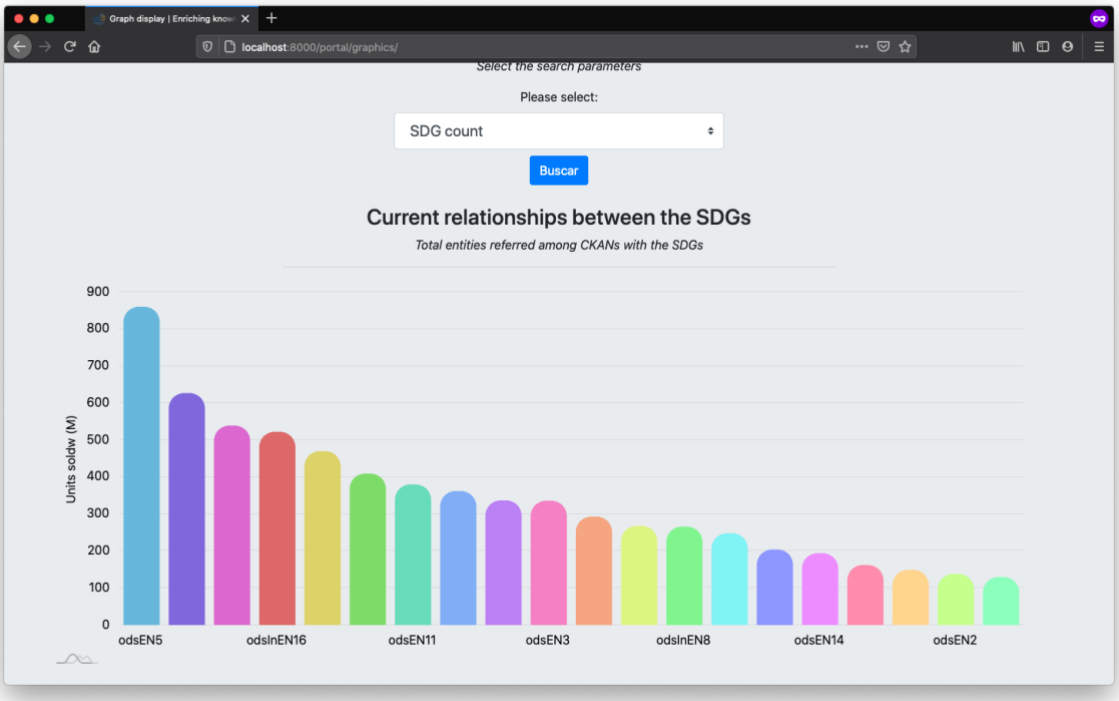
Semi-círculo de sectores de mayores entidades mencionadas de la dbpedia



La **Figura 41** presenta el total de recursos analizados y el total de enlaces que se tiene con cada ODS.

Figura 41

Total de entidades referenciadas entre los recursos CKAN con los ODS



*4.2.2 **Otros Resultados***

Una vez concluido el sitio web piloto y el proceso de enriquecimiento de los datos se pude resumir lo siguiente. De los 39727 recursos CKAN semánticos se ha podido enriquece más de un 36% de todo el contenido. También un 4% del dataset principal no se ha encontrado alguna coincidencia en los ODS y la Dbpedia. Y el 60% representa el resto de los datos que faltan por enriquecer. Este resumen se lo puede apreciar en la ***Figura 42***.

Después de pasar por el proceso de análisis mediante los ODS se puede mencionar que los ODS más usados en el enriquecimiento fueron los siguientes. El ODS 17 que representa a *Alianza para lograr los objetivos* en el cual se tiene más de 1300 recursos asociados al CKAN. Otro más usado es el ODS 5 que representa *la igualdad de género* con más 900 recursos enlazados y otro que es el más mencionado es el ODS 16 que es *Promover sociedades justas, pacíficas e inclusivas* con más de 500 recursos enlazados. En la figura se puede ver más datos de uso de los ODS.

Figura 42

Resultado de datos enriquecidos

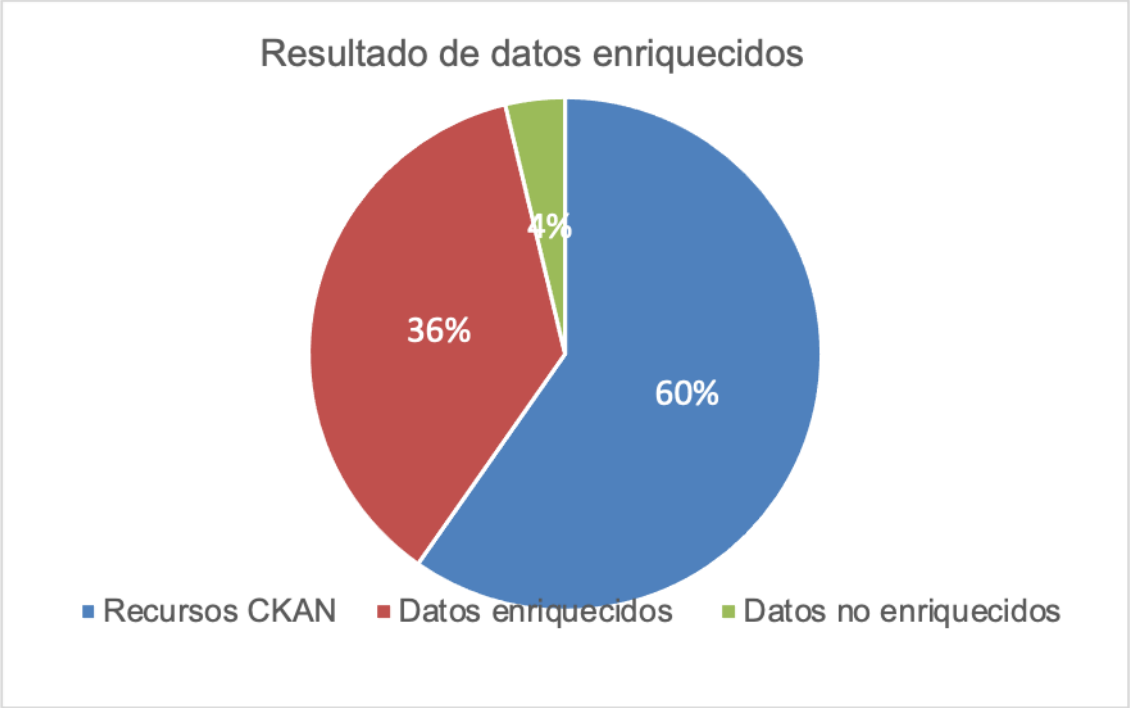
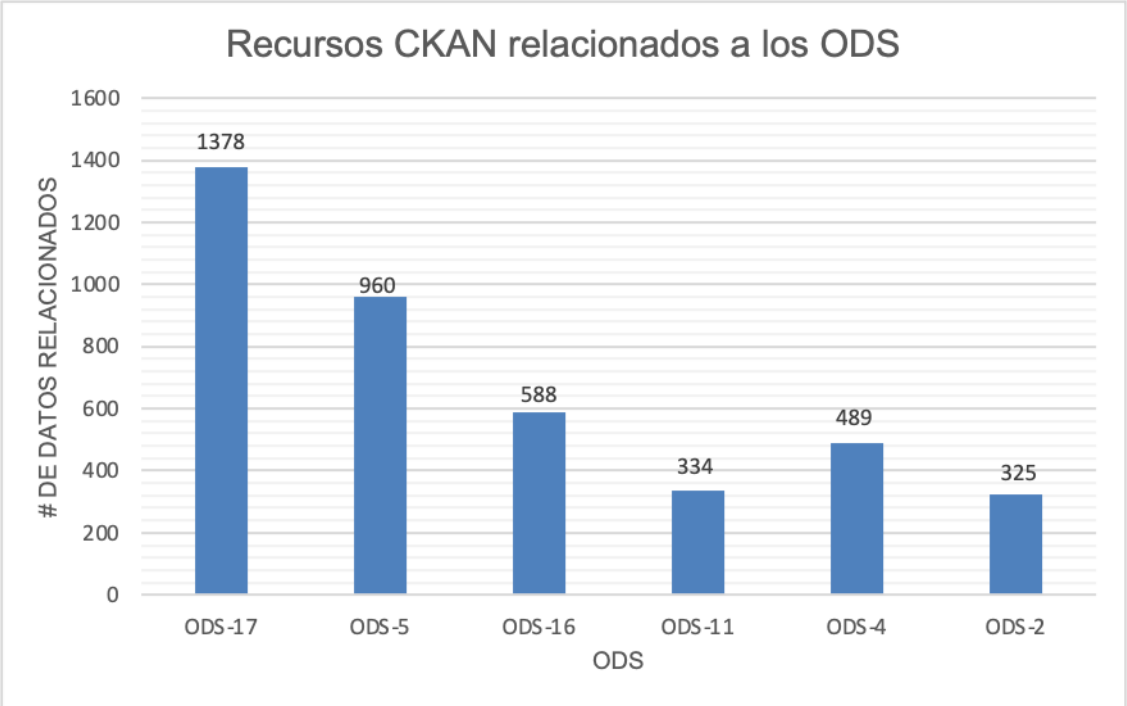


Figura 43

Recursos CKAN relacionados a los ODS



Finalmente, a través de la arquitectura o del método planteado es capas de acoplarse a diversos datos que requieran ser enriquecidos, tomando en cuenta la estructura que estos posean y de esta manera poder incrementar el conocimiento actual en los datos. El algoritmo es autónomo, esto quiere decir que es capaz de trabajar por si solo durante un tiempo prolongado.

4.3 **Discusión**

Después de terminar la aplicación se analiza lo siguiente, la solución propuesta para este trabajo cumple con todos los objetivos del trabajo de titulación. Para esto se logro proponer y crear un método de enriquecimiento de Grafos de Conocimiento mediante la arquitectura propuesta en la **Figura 11**. Este busca unir cualquier recurso abierto a uno y varios indicadores de ODS y también con otras fuentes de datos abiertos en este caso con los recursos la DBpedia.

En primera instancia se almacenó los datos en GraphDB la primera versión y una vez analizados los datos y procesados se generaron nuevas tripletas siendo estas el resultado del enriquecimiento de los nuevos datos y luego fueron almacenadas en la base de datos semántica GraphDB (segunda versión o datos enriquecidos). Esto permite el consumo en tiempo real de la información que se requiere, pero para esto se necesita generar consultas basadas en SPARQL de esta forma permitiendo visualizar el nuevo contenido añadido a los recursos CKAN.

El proceso de enriquecimiento de datos se lo realizó usando la herramienta de fasttext. La parte fundamental para el proceso es tener una base de conocimiento bien estructurada que permita un análisis lógico y semántico. Y también se uso la herramienta de la Dbpedia que se llama dbpedia-spotlight. Este Framework permitirá dar un valor agregado a la información que es enriquecida con el fin de determinar que palabras o frases tienen incidencia en el texto con la Dbpedia.

Por otra parte, con las herramientas usadas se lograron establecer nuevos enlaces entre diferentes datasets y enlaces con el dataset externos con la dbpedia siendo estos de acceso abierto y determinado su relación con los ODS. Toda esta información enriquecida

puede ser visualizada en el sitio web piloto mediante las diferentes APIs creadas. Este sitio permite que cualquier persona que pueda hacer uso de estos resultados, ayudar al análisis de las mismas y usarla fácilmente.

En estebuye con la capacidad de crear un método que permita en 5 fases (la extracción, NER, NLP, Flujo de trabajo y publicación) extraer información, analizar, crear nuevas relaciones con diferentes conjuntos de datos relacionados con la objetivos, metas e indicadores de los ODS, y publicar estos datos en la web.

Finalmente, este trabajo buscar ser la base para investigaciones futuras, y aportar a nuevas iniciativas para enlazar a múltiples conjunto de datos relacionados con la Agenda 2030 y asi mejorar la información que se encuentra en la web. Adicionalmente, los procesos realizados en la aplicación pueden ser adaptados para cualquier tipo de conjunto de datos semánticos y así mejorar el conocimiento actual e ir incrementado la información de forma rápida y segura.

Conclusiones

La motivación principal del trabajo de investigación es crear un método para enriquecer conjuntos de datos abiertos relacionados con la agenda 2030, extraídos desde plataformas CKAN, y que se han descrito como Grafos de Conocimiento Semántico anotados a una base de conocimiento de ODS y a recursos de la Dbpedia.

A partir de los ODS, indicadores y metas se ha creado una base de conocimiento (BC) descrita en RDF, que se usó para detección de similitudes, la recomendación y etiquetado semántico a través de *fasttext*. Esta base de conocimiento se creó a través de la combinación de procesos manuales (extracción) y automatizados (transformación). El idioma utilizado en la BC es el inglés, y contiene 852 entradas relacionadas con objetivos, metas e indicadores.

Se usó Dbpedia-spotlight para tareas de análisis semántico: Reconocimiento de Entidades Nombradas, Procesamiento de Lenguaje Natural y Mapeo Semántico. El análisis de texto ha resultado en la creación de más de 8 mil nuevas relaciones a partir de los datasets descritos como parte de este trabajo.

La creación del proceso el enriquecimiento ha sido evolutivo. La versión final de este proceso implica realizar un proceso de traducción para identificar el idioma del texto de entrada, en caso de que sea diferente del idioma inglés el texto es traducido al idioma base. A continuación, se realiza una limpieza de caracteres especiales. El siguiente paso consiste en aplicar procesamiento de lenguaje natural para reconocer entidades nombradas.

Para la creación de las nuevas relaciones se usó vocabularios semánticos, *dcat:theme* dando lugar a 6680 nuevas relaciones resultantes del uso de fasttext; *dcat:qualifiedRelation* para anotar 8054 nuevas relaciones resultantes del uso de dbpedia-spotlight y finalmente *dct:language* para la anotación de 7593 nuevas relaciones que determinan el idioma del grafo de conocimiento inicial (Eguiguren Palacios, 2019).

El enriquecimiento y la visualización del conjunto de datos abiertos enriquecidos dio lugar al desarrollo de una aplicación web. Dicha aplicación web permite navegar conjuntos de datos abiertos anotados con los ODS. El sitio es interactivo y accesible para todo tipo de usuario.

Recomendaciones

A continuación, se describen algunas recomendaciones al termino del presente trabajo y tomando en cuenta trabajos futuros.

Se recomienda realizar un análisis previo de los datos que se van a procesar con la finalidad de determinar los valores serán los usados para el enriquecimiento del grafo. Es importante tener en cuenta como adaptar sus estrategias de información con los datos con las nuevas relaciones y esas estén adaptadas al vocabulario del grafo principal.

El re-uso de recursos ontológicos es clave en el proceso de enriquecimiento semántico. Sin embargo, para ampliar las posibilidades de enriquecimiento de conjuntos de datos abiertos relacionados con los ODS, sería conveniente crear una ontología o vocabularios semánticos específicos para la anotación de recursos con elementos de información propios de la Agenda 2030.

Además de las tareas NER y NLP reportadas en este trabajo, consideramos importante combinar estos con SPACY con el fin de mejorar la calidad del enriquecimiento de los grafos de conocimiento.

En cuanto a la base de conocimiento usado para fasttext, como trabajo futuro se tiene previsto ampliarla a otros idiomas, esto aumentará la capacidad de anotar semánticamente conjuntos de datos que usan lenguas diferentes al inglés.

En el análisis que se realiza para identificar las entidades nombradas de los conjuntos de datos e interpretar el texto de entrada se recomienda que el índice de similitud sea superior a 0,8.

La propuesta de enriquecimiento de grafos semánticos presentada en este trabajo al usar servicios y recursos de información en la nube y en tiempo real (dbpedia-spotlight, googleTranslate, fasttext) demanda calidad de conexión a Internet. Este es uno de los desafíos que debe ser gestionado si se busca contar con un entorno basado en grafos de conocimiento semántico a gran escala.

Referencias

Alesso, H. P., & Smith, C. F. (2006). *THINKING ON THE WEB Berners-Lee, Gödel, and Turing*.

Berners-Lee, T. (2015). *5-star Open Data*. https://5stardata.info/en/

Berners-Lee, T., Connolly, D., Stein, L.A., Swick R. (2000). *The Semantic Web*. https://www.w3.org/2000/Talks/0906-xmlweb-tbl/text.htm

Bizer, C. (2009). The emerging web of linked data. *IEEE Intelligent Systems*, *24*(5). https://doi.org/10.1109/MIS.2009.102

Cheaney, A. (2012). *What is Web 2.0?* Quora. https://www.quora.com/What-is-Web-2-0

Danneels, L., Viaene, S., & Van den Bergh, J. (2017). Open data platforms: Discussing alternative knowledge epistemologies. In *Government Information Quarterly* (Vol. 34, Issue 3, pp. 365–378). Elsevier Ltd. https://doi.org/10.1016/j.giq.2017.08.007

Duus, R., & Cooray, M. (2016). *The importance of open data | World Economic Forum*. https://www.weforum.org/agenda/2016/02/the-importance-of-open-data/

Eguiguren Palacios, J. E. (2019). *Marco de trabajo para la integración semántica de portales de datos abiertos basados en CKAN Piloto de Datos Abiertos alineados a los indicadores ODS. (Trabajo de Titulación de Ingeniero en Sistemas Informáticos y Computación). UTPL, Loja*. http://dspace.utpl.edu.ec/handle/20.500.11962/24629

Escobar, P., Candela, G., Trujillo, J., Marco-Such, M., & Peral, J. (2020). Adding value to Linked Open Data using a multidimensional model approach based on the RDF Data Cube vocabulary. *Computer Standards and Interfaces*, *68*(September 2019), 103378. https://doi.org/10.1016/j.csi.2019.103378

Gavin, B. (2018). *What Is An XML File?* https://www.howtogeek.com/357092/what-is-an-xml-file-and-how-do-i-open-one/

Griggs, D., Stafford-Smith, M., Gaffney, O., & Rockström, J. (2013). *Sustainable development goals for people and planet*. 2013. https://www.nature.com/articles/495305a

Hazaël-Massieux, D., & Berners-Lee, T. (2003). *The Semantic Web and its applications at W3C*. https://www.w3.org/2003/Talks/simo-semwebapp/all.htm

Indicadores de ODS. (2010). http://www.endvawnow.org/es/articles/336-indicadores.html

Jabbar, J. A., & Bulbul, R. (2019). SEMANTIC ENRICHMENT of ROUTING ENGINES USING LINKED DATA: A CASE STUDY USING GRAPHHOPPER. *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences - ISPRS Archives*, *42*(4/W14), 111–117. https://doi.org/10.5194/isprs-archives-XLII-4-W14-111-2019

Kitchenham, B., & Charters, S. (2007). *Guidelines for performing Systematic Literature Reviews in Software Engineering*.

Lapuente, M. J., & Lamarca, C. (2018). *RDF*.

Miessler, D. (2020). *Difference Between a URI and a URL*. https://danielmiessler.com/study/difference-between-uri-url/

Okoli, C., & Schabram, K. (2010). *Working Papers on Information Systems A Guide to Conducting a Systematic Literature Review of Information Systems Research A Guide to Conducting a Systematic Literature Review of Information Systems Research*.

Ontotext. (2020). *GraphDB Documentation*. http://graphdb.ontotext.com/documentation/standard/

ONU. (2019). *UNSD SDGs API*. https://unstats.un.org/SDGAPI/swagger/#!/Indicator/V1SdgIndicatorByIndicatorCodeSeriesListGet

Open Data Handbook. (2019). *La importancia del Open Data*. https://www.unblogenred.es/la-importancia-del-open-data/

Pfenninger, S., DeCarolis, J., Hirth, L., Quoilin, S., & Staffell, I. (2017). The importance of open data and software: Is energy research lagging behind? *Energy Policy*, *101*, 211–215. https://doi.org/10.1016/j.enpol.2016.11.046

Quattrini, R., Pierdicca, R., & Morbidoni, C. (2017). Knowledge-based data enrichment for HBIM: Exploring high-quality models using the semantic-web. *Journal of Cultural*

Heritage, *28*, 129–139. https://doi.org/10.1016/j.culher.2017.05.004

Sánchez, A., Piedra, N., & Morocho, J. C. (2018). Using linked data to ensure that digital information about historical figures of Loja remains accessible and usable. *CEUR Workshop Proceedings*, *2096*, 21–34.

Sharma, V. (2015). *What are WEB 2.0 & WEB 3.0?* https://www.quora.com/What-are-WEB-2-0-WEB-3-0

Song, H. J., & Park, S. B. (2018). Enriching translation-based knowledge graph embeddings through continual learning. *IEEE Access*, *6*, 60489–60497. https://doi.org/10.1109/ACCESS.2018.2874656

Stauffacher, D., Hattotuwa, S., & Weekes, B. (2012). *The potential and challenges of open data for crisis information management and aid efficiency A preliminary assessment*. http://wbi.worldbank.org/wbi/open-aid-partnership

Techopedia. (n.d.). *What is the Open Data Platform (ODP)*. Retrieved January 8, 2020, from https://www.techopedia.com/definition/31438/open-data-platform-odp

W3C. (2015). *Ontologies*. https://www.w3.org/standards/semanticweb/ontology

Wonderflow. (2018). *12 NLP Examples: How Natural Language Processing is Used*. Wonderflow. https://www.wonderflow.co/blog/natural-language-processing-examples

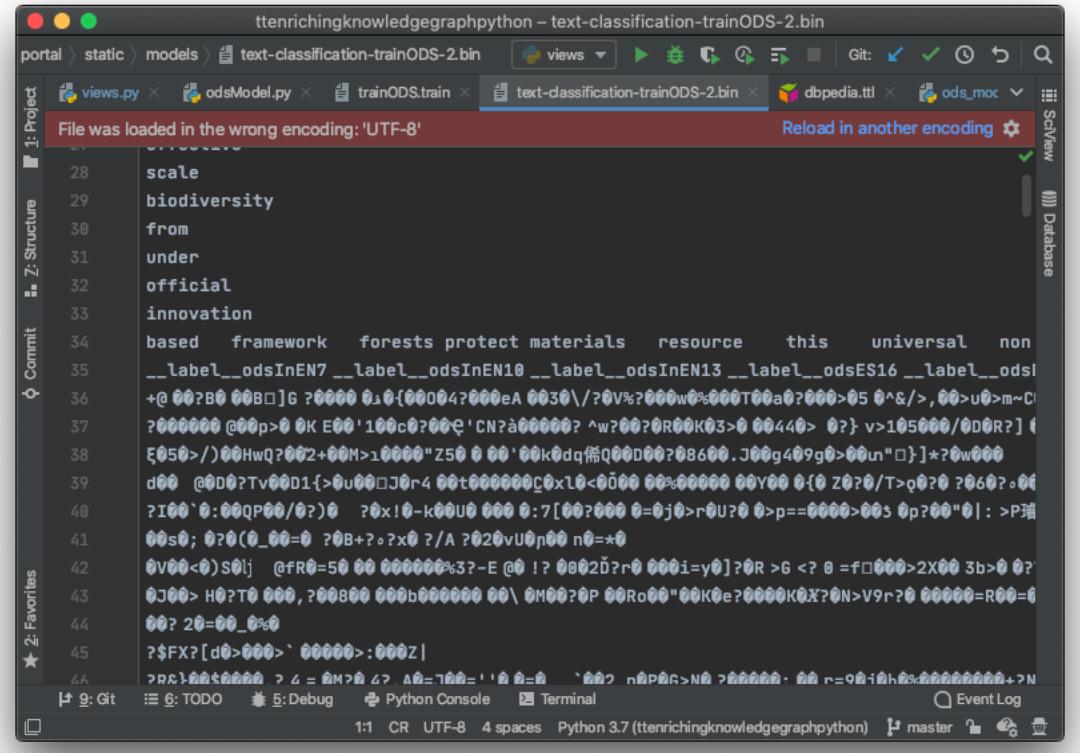
Apéndice

Se incluye de acuerdo al orden citado en el cuerpo del Trabajo de Titulación.

Apéndice 1:

Figura 44

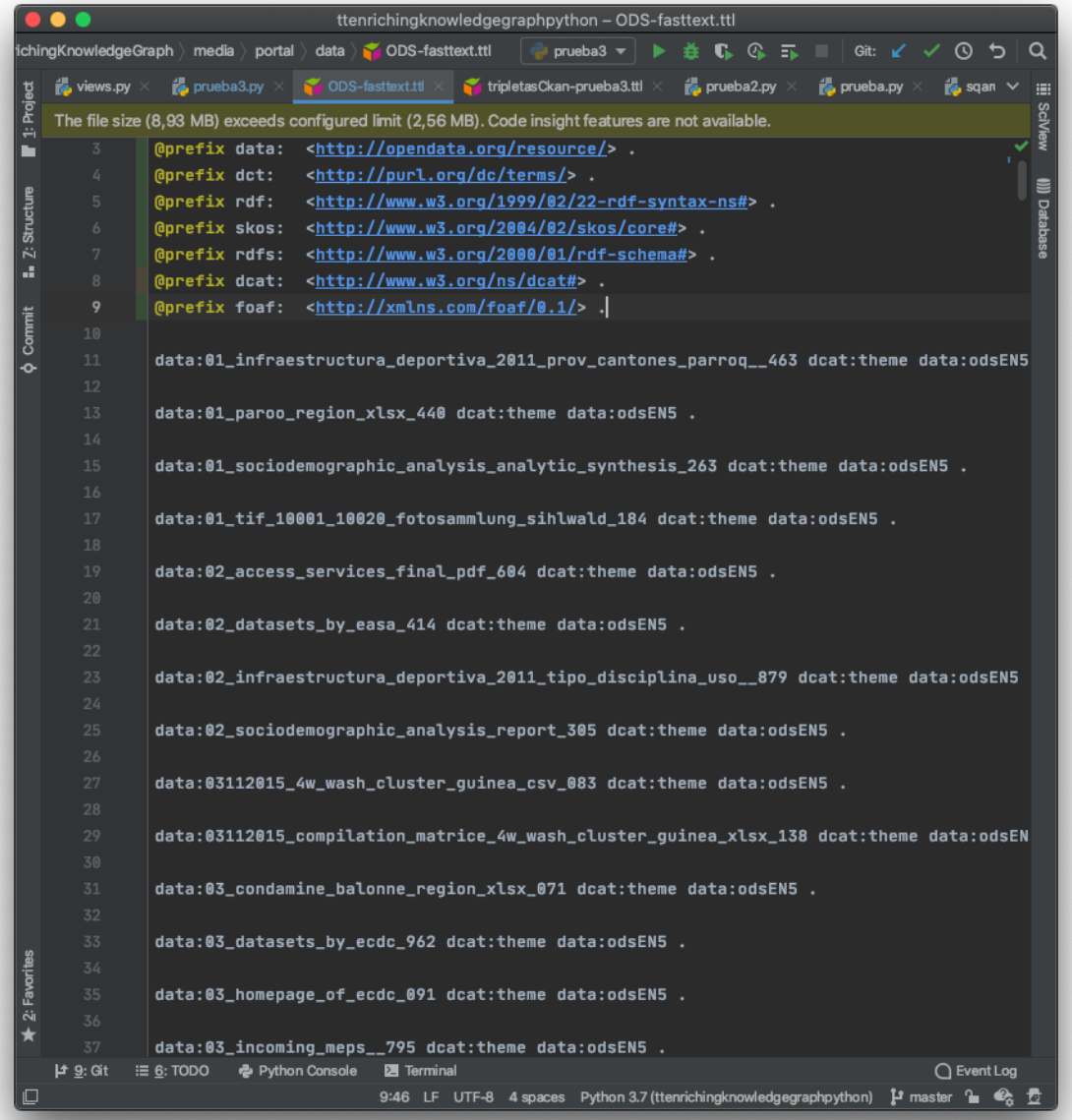
Modelo entrenado con Fasttext.



Apéndice 2:

Figura 45

Fragmento de las nuevas relaciones creadas con fasttext

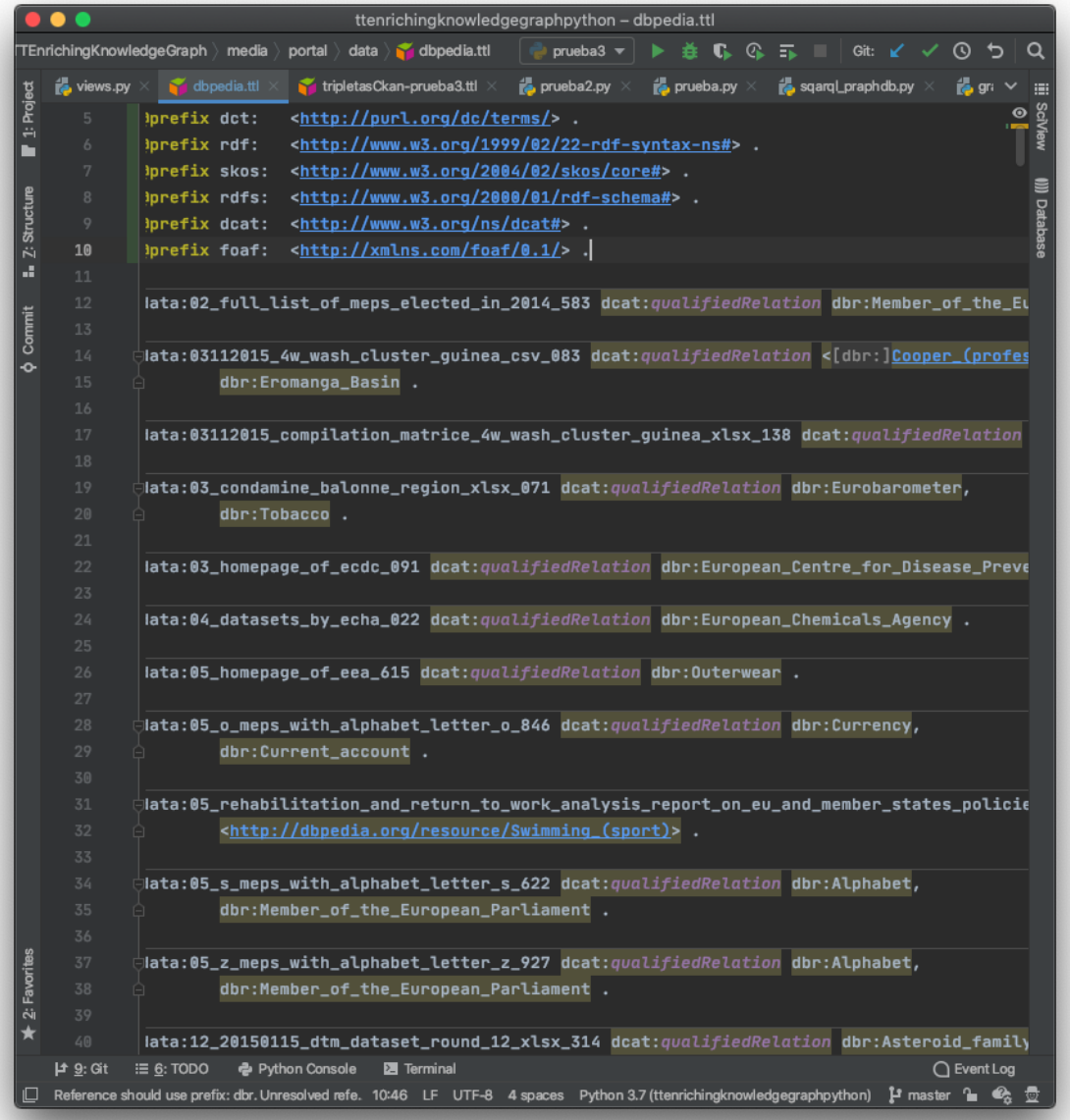


En esta imagen se aprecia en un formato turtle (ttl) en resultado de las nuevas relaciones entre los recursos y los ODS.

Apéndice 3:

Figura 46

Fragmento de las nuevas relaciones creadas con spacy.

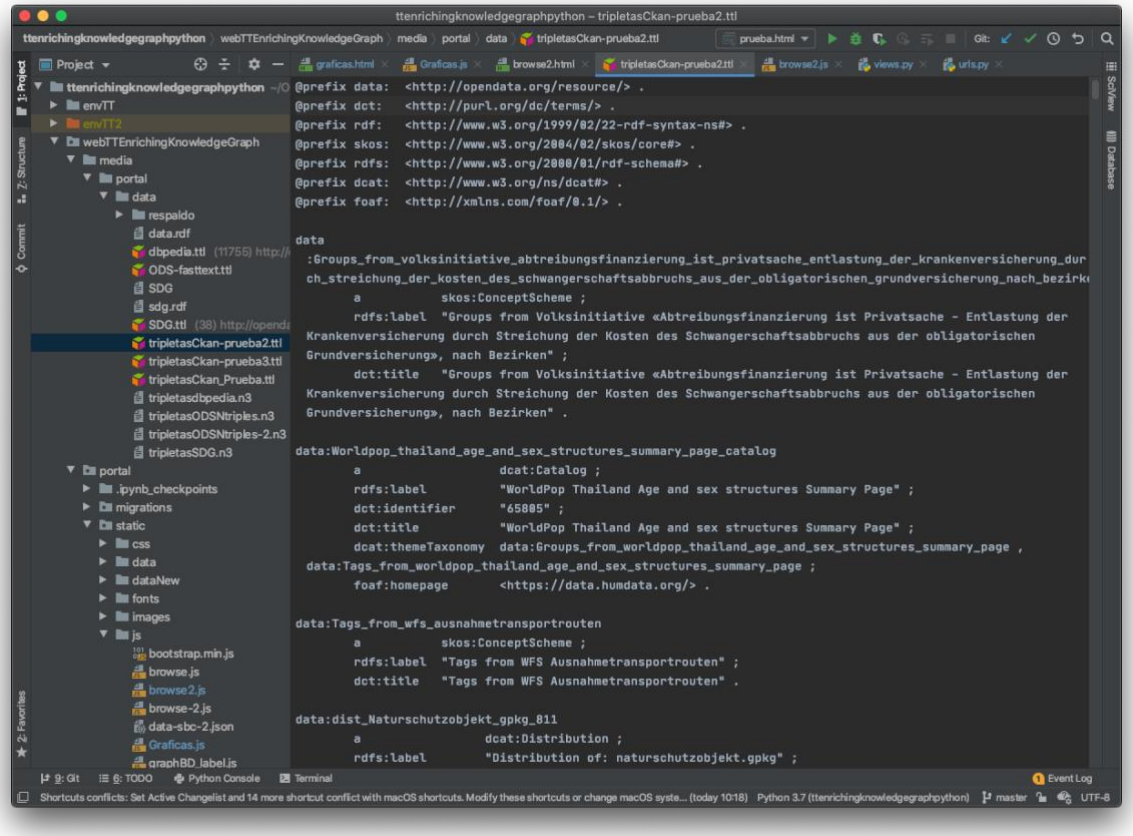


En esta imagen se aprecia en un formato turtle (ttl) en resultado de las nuevas relaciones entre los recursos y los recursos con la dbpedia.

Apéndice 4:

Figura 47

Fragmento de los datos antes de ser enriquecidos.

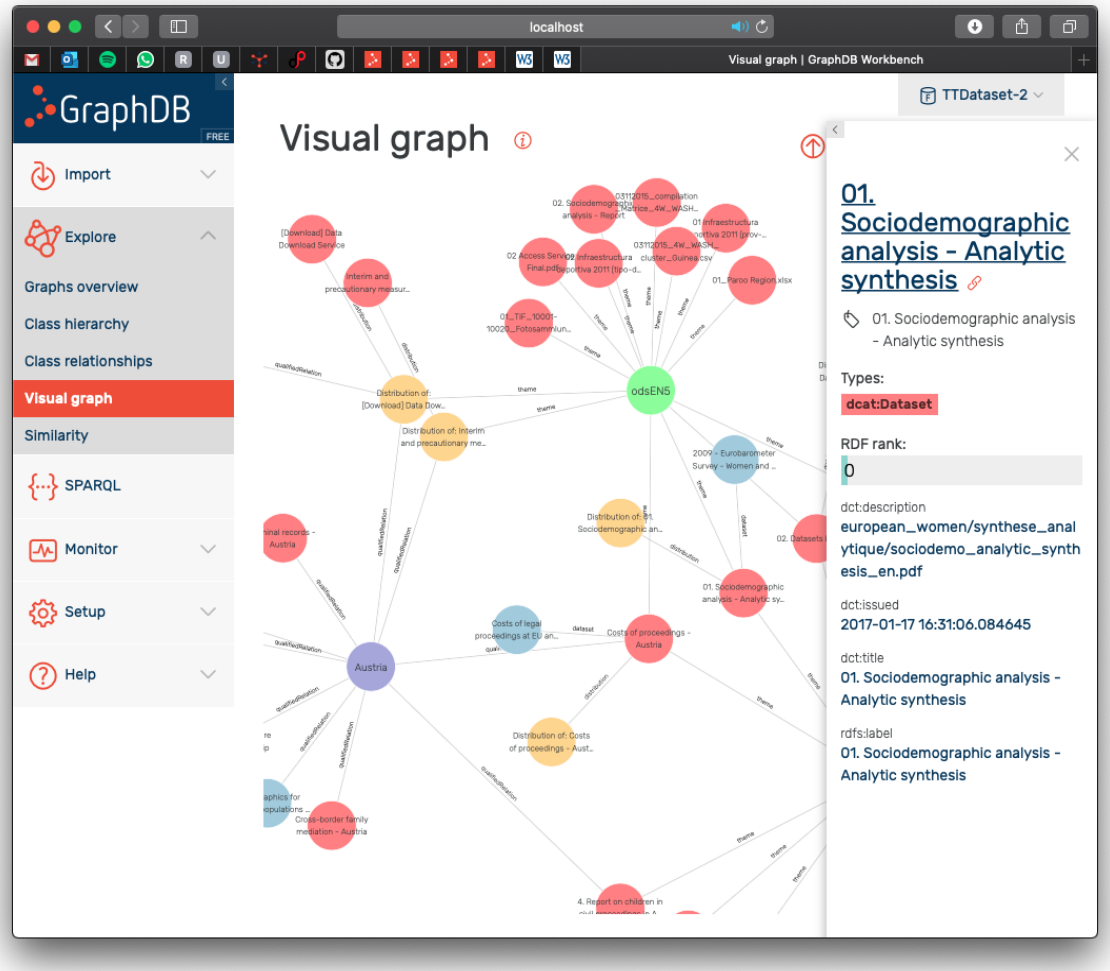


En esta imagen se aprecia en un formato turtle (ttl) los recursos CKAN base.

Apéndice 5:

Figura 48

Resultado del enriquecimiento con los nuevos enlaces de los ODS y de la dbpedia.

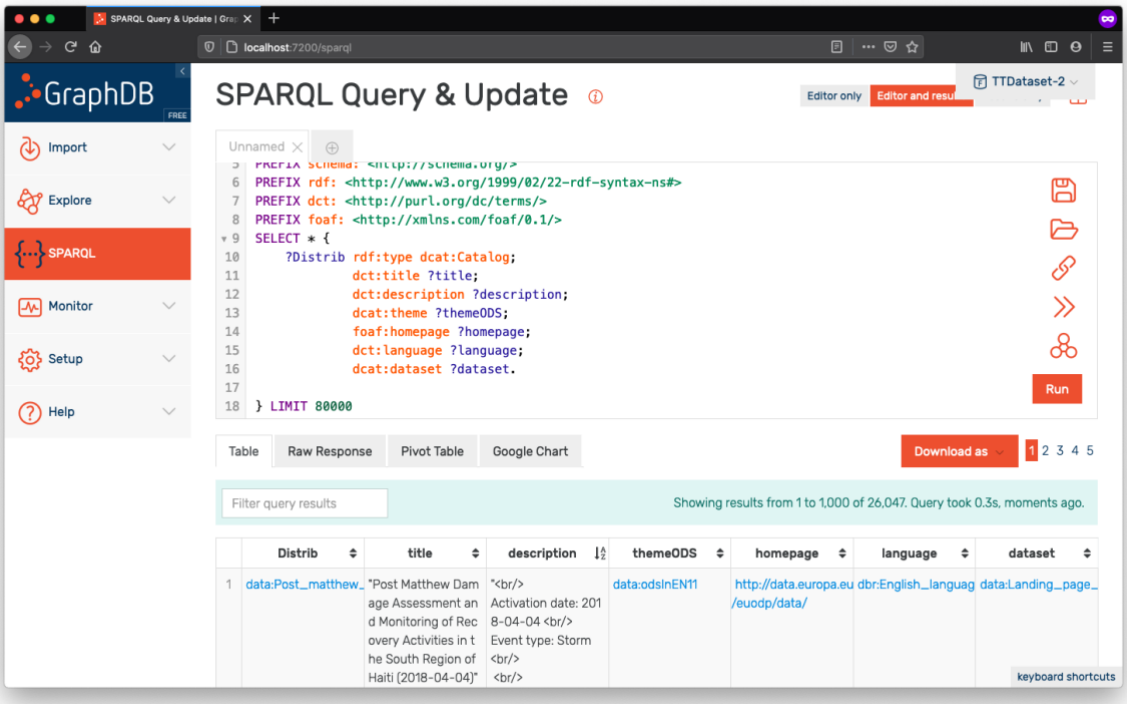


Este es el resultado de la visualización desde la aplicación de GraphDB.

Apéndice 6:

Figura 49

Consulta SPARQL para obtener los resultados de análisis.

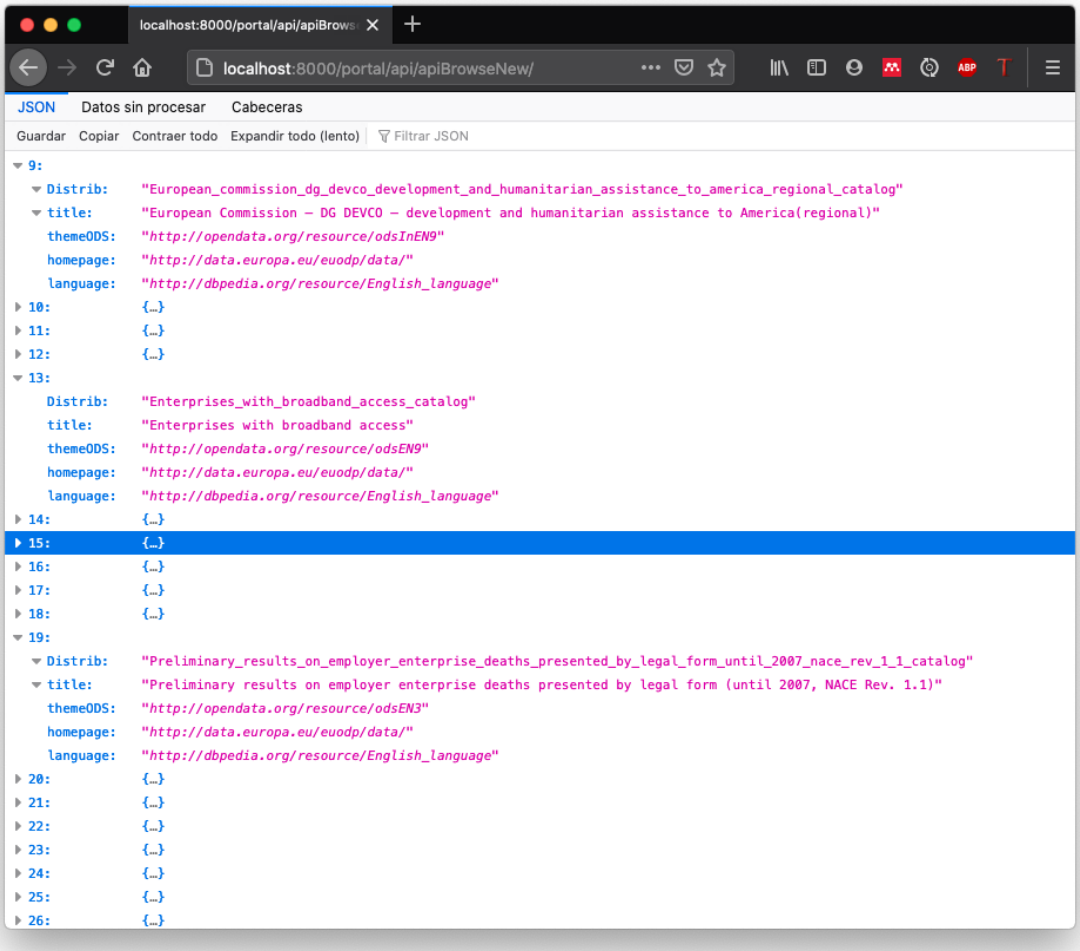


En esa consulta se busca obtener los resultados que nos permita saber sobre los recursos que están enriquecidos y saber más de esos recursos analizados.

Apéndice 7:

Figura 50

Api/Json de los resultados de la consulta Sparql en tiempo real.

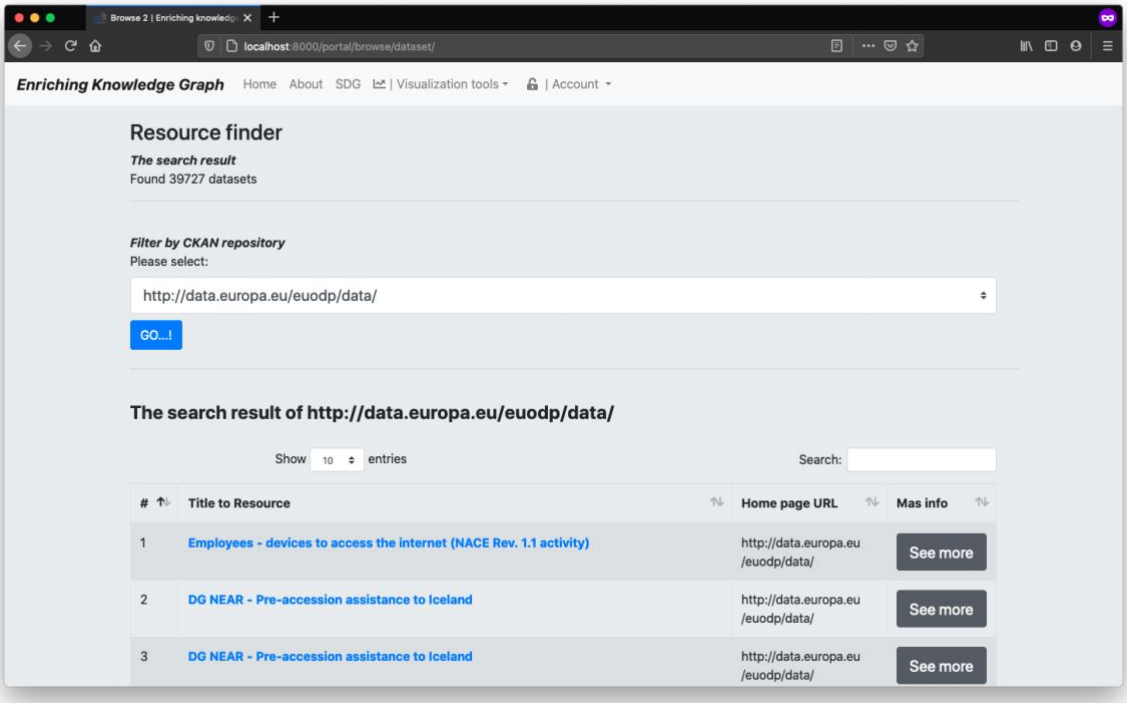


Este json busca estructurar y analizar los resultados de manera mas fácil para el consumo del mismo en el sitio.

Apéndice 8:

Figura 51

Buscador en la primera versión

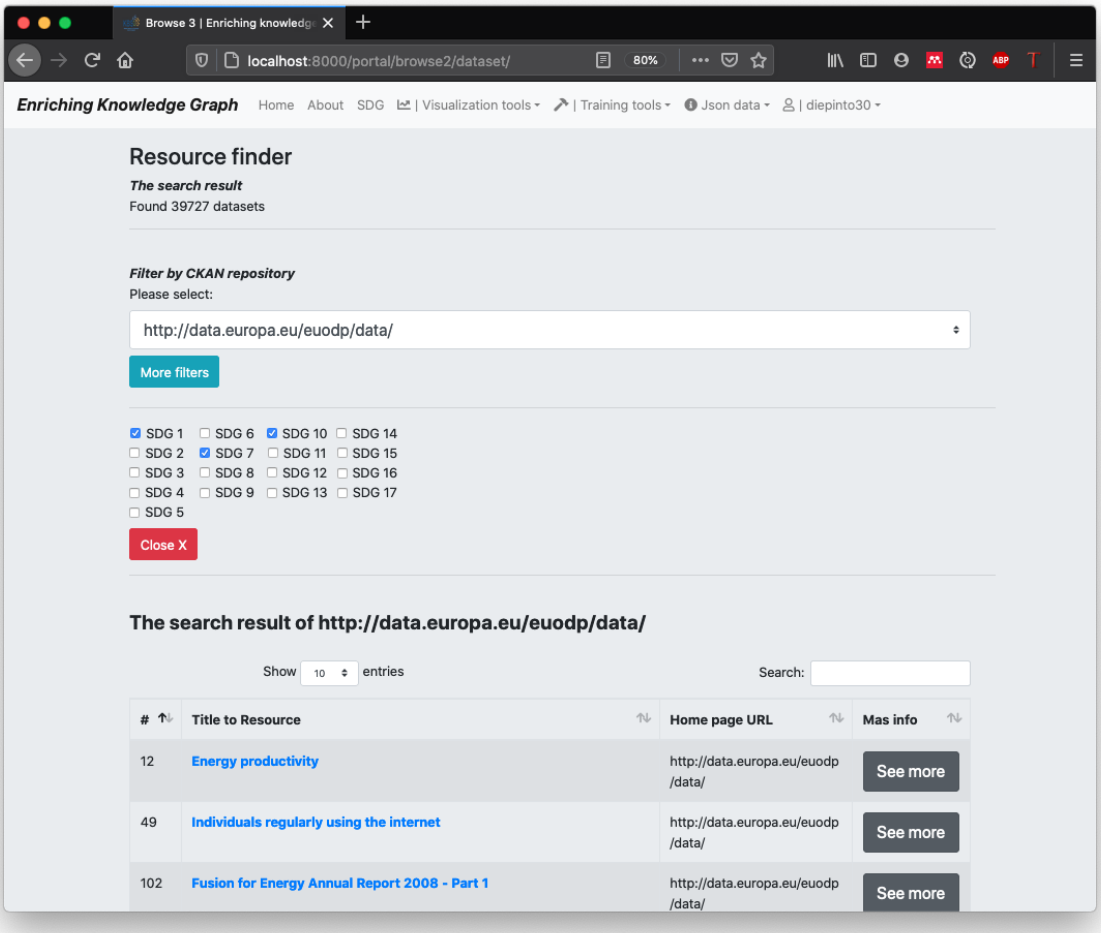


Este buscador tiene filtro por los homePage de los recursos y un buscador por entrada de texto.

Apéndice 9:

Figura 52

Buscador en la segunda versión

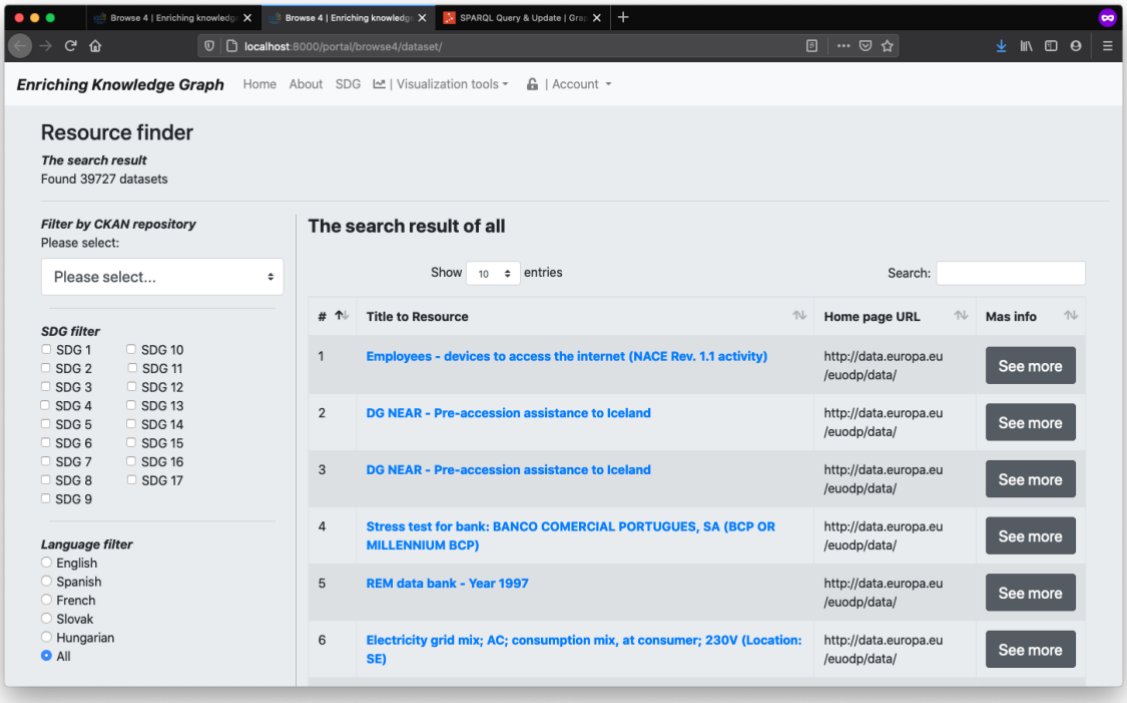


Este buscador tiene filtro por los homePage de los recursos, un checkbox para filtrar por los ODS relacionados con uno o varios recursos y un buscador por entrada de texto.

Apéndice 10:

Figura 53

Buscador en la tercera versión

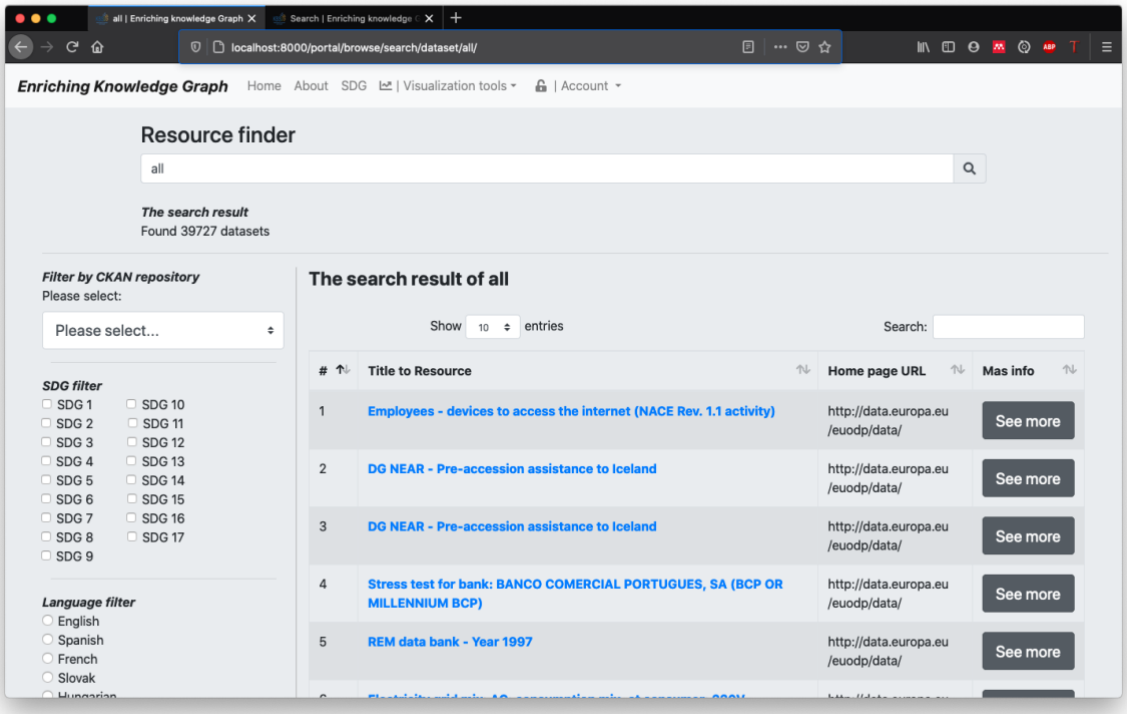


Este buscador tiene filtro por los homePage de los recursos, un checkbox para filtrar por los ODS relacionados con uno o varios recursos, un radiobox para filtrar por idiomas y un buscador por entrada de texto.

Apéndice 11:

Figura 54

Buscador en la cuarta versión



Este buscador tiene un buscador de coincidencia de palabras y filtros mejorados por el homePage de los recursos, un checkbox para filtrar por los ODS relacionados con uno o varios recursos, también otro con checkbox para filtrar por idiomas y un buscador por entrada de texto del resultado obtenido.