

**UNIVERSIDAD POLITÉCNICA SALESIANA
SEDE CUENCA**

CARRERA DE INGENIERÍA DE SISTEMAS

*Trabajo de titulación previo a
la obtención del título de
Ingeniero de Sistemas*

PROYECTO TÉCNICO:

**“DISEÑO Y DESARROLLO DE UN CRAWLER SEMÁNTICO
PARA LA GENERACIÓN DE POBLADORES DE ONTOLOGÍAS”**

AUTOR:

GEOVANNY ALEXANDER QUIÑONEZ LAMBERT

TUTOR:

VLADIMIR ESPARTACO ROBLES BYKBAEV, PhD.

CUENCA - ECUADOR

2020

CESIÓN DE DERECHOS DE AUTOR.

Yo, Geovanny Alexander Quiñonez Lambert con documento de identificación N° 0703932814, manifiesto mi voluntad y cedo a la Universidad Politécnica Salesiana, la titularidad sobre los derechos patrimoniales en virtud de que soy autor del trabajo de titulación: **“DISEÑO Y DESARROLLO DE UN CRAWLER SEMÁNTICO PARA LA GENERACIÓN DE POBLADORES DE ONTOLOGÍAS”**, mismo que ha sido desarrollado para optar por el título de: *Ingeniero de Sistemas*, en la Universidad Politécnica Salesiana, quedando la Universidad facultada para ejercer plenamente los derechos cedidos anteriormente.

En aplicación a lo determinado en la Ley de Propiedad Intelectual, en mi condición de autor, me reservo los derechos morales de la obra antes citada. En concordancia, suscribo este documento en el momento que hago entrega del trabajo final en formato digital a la Biblioteca de la Universidad Politécnica Salesiana.

Cuenca, febrero del 2020



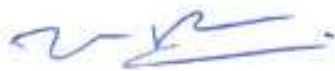
Geovanny Alexander Quiñonez Lambert

C.I. 070932814

CERTIFICACIÓN.

Yo, declaro que bajo mi tutoría fue desarrollado el trabajo de titulación:
“DISEÑO Y DESARROLLO DE UN CRAWLER SEMÁNTICO PARA LA GENERACIÓN DE POBLADORES DE ONTOLOGÍAS”, realizado por Geovanny Alexander Quiñonez Lambert, obteniendo el *Proyecto Técnico*, que cumple con todos los requisitos estipulados por la Universidad Politécnica Salesiana.

Cuenca, febrero del 2020.



Ph.D Vladimir Espartaco Robles Bykbaev

C.I. 0300991817

DECLARATORIA DE RESPONSABILIDAD.

Yo, Geovanny Alexander Quiñonez Lambert con documento de identificación N° 0703932814, autor del trabajo de titulación: **“DISEÑO Y DESARROLLO DE UN CRAWLER SEMÁNTICO PARA LA GENERACIÓN DE POBLADORES DE ONTOLOGÍAS”**; certifico que el total contenido del *Proyecto Técnico*, es de mi exclusiva responsabilidad y autoría.

Cuenca, febrero del 2020.



Geovanny Alexander Quiñonez Lambert
C.I. 070932814

AGRADECIMIENTOS.

Quiero expresar el más sincero agradecimiento a mi Tutor de proyecto técnico Ph.D. Vladimir Robles Bykbaev; además un agradecimiento a quien formo parte en la asesoría de este proyecto, al Ing. Diego Quisi, y a todos quienes durante este tiempo han formado parte del Grupo de Investigación en Inteligencia Artificial y Tecnologías de Asistencia (GIIATA) de la Universidad Politécnica Salesiana – Sede Cuenca y a la Cátedra UNESCO por brindarme su apoyo de manera desinteresada durante el desarrollo del presente trabajo.

Geovanny Alexander Quiñonez Lambert.

DEDICATORIA.

Esta tesis la dedico a mis abuelas que estuvieron siempre a mi lado brindándome sus oraciones y sus bendiciones, dándome a cada instante su aliento para llegar a culminar mi profesión, a cada uno de los miembros de mi familia que supieron darme su apoyo, a mis queridos amigos, que son hermanos y hermanas de otras madres, que supieron aguantar con paciencia mi apasionado e imposible carácter, haciéndome entrar en conciencia cuando no encontraba el camino.

Geovanny Alexander Quiñonez Lambert.

Índice de contenido

| | |
|--|----|
| 1. Resumen..... | 1 |
| 2. Abstract | 2 |
| 3. Introducción..... | 3 |
| 4. Objetivos. | 4 |
| 4.1. General..... | 4 |
| 4.2. Específicos | 4 |
| 5. Estado del Arte..... | 5 |
| 5.1. Uso de ontologías existentes en proyectos y ámbitos de desarrollo..... | 5 |
| 5.2. Tecnología semántica y Crawlers Semánticos dentro del campo de la investigación. | 6 |
| 6. Diseño de la propuesta planteada. | 7 |
| 6.1. Web Semántica..... | 7 |
| 6.1.1. Búsquedas y tecnología en la Web Semántica. | 8 |
| 6.1.1.1. XML (Extensible Markup Language)..... | 8 |
| 6.1.1.2. RDF (<i>Resource Description Framework</i>) | 8 |
| 6.1.1.3. OWL (Web Ontology Language) | 9 |
| 6.1.1.4. SPARQL (Protocol and RDF Query Language)..... | 9 |
| 6.1.1.5. HTML (HyperText Markup Language)..... | 10 |
| 6.2. La Web Semántica Y Ontologías..... | 10 |
| 6.3. Ontologías. | 11 |
| 6.4. MODS (Marco Ontológico Dinámico Semántico) | 11 |
| 6.5. <i>Crawler</i> | 12 |
| 6.6. Procesamiento de lenguaje Natural..... | 12 |
| 7. Desarrollo de la propuesta planteada..... | 13 |
| 7.1. Arquitectura del sistema. | 13 |
| 7.1.1. Capa de interacción..... | 14 |
| 7.1.2. Capa de servicios. | 15 |
| 7.1.2.1. Módulo de <i>Crawler</i> | 15 |
| 7.1.2.2. Módulo extracción de datos. | 16 |
| 7.1.2.3. Módulo de representación gráfica. | 16 |
| 7.1.3. Capa de módulo inteligente. | 16 |
| 7.1.3.1. Módulo PNL (<i>Natural Language Processing</i>) | 16 |
| 7.1.3.2. Módulo de gestión ontológica..... | 17 |
| 7.1.3.3. Módulo de análisis estadístico..... | 17 |
| 7.1.3.4. Módulo de consulta SPARQL..... | 17 |

| | |
|---|----|
| 7.1.4. Capa de datos. | 18 |
| 7.1.4.1. Módulo de consulta. | 18 |
| 8. Funcionamiento de la interfaz (SPELTA-ADaS) | 18 |
| 9. Experimentación con Razonadores | 22 |
| 9.1. Experimentación. | 22 |
| 9.2. Recomendaciones generales | 25 |
| 10. Conclusiones | 26 |
| 11. Trabajos a Futuro | 27 |
| 12. Referencias | 28 |

Índice de Figuras

| | |
|---|----|
| Figura 1: estructura RDF básica. [23]..... | 9 |
| Figura 2: Código de etiquetas HTML [23]..... | 10 |
| Figura 3: Código de etiquetas HTML..... | 10 |
| Figura 4 : Marco Ontológico Dinámico Semántico [27]..... | 12 |
| Figura 5 : Propuesta de la Arquitectura de SPELTA-ADAs (SPELTA-Automatic Data Searcher)..... | 13 |
| Figura 6 : Modelo MTV en Django..... | 15 |
| Figura 7 : Funcionamiento de Django - Scrapy | 16 |
| Figura 8 : Imagen extraída del artículo "An ontology-based expert system to generate therapy plans for children with disabilities and communication disorders", esta es una captura parcial de la ontología original..... | 19 |
| Figura 9 : Menú del entorno web, izquierda un navegador en Gnu-Linux y a la derecha un navegador en android. | 19 |
| Figura 10 : Pagina de configuración del SPELTA-ADaS | 20 |
| Figura 11 : Pantalla de Consultas SPARQL | 20 |
| Figura 12 : Administración De Actividades | 21 |
| Figura 13 : Pantalla donde el botón de proceso referencia. | 21 |
| Figura 14 : Grafica de palabras que aparecen con mayor frecuencia..... | 22 |
| Figura 15 : Sección de la ontología que presenta el módulo de representación gráfica. | 22 |
| Figura 16 : Protégé y la ontología de terapia Lenguaje importada en él..... | 23 |
| Figura 17 : Resultado del razonador SWRL..... | 23 |
| Figura 18 : Desordenes..... | 24 |
| Figura 19 : Habilidades declaradas..... | 24 |
| Figura 20 : Error de tipo de dato, en la ontología base..... | 25 |

1. Resumen.

El lenguaje y la comunicación se consideran pilares para desarrollar habilidades cognitivas, sociales y psicológicas en los niños. Sin embargo, la Terapia del Habla y Lenguaje (THL) no se ha abordado adecuadamente desde el punto de vista de las herramientas de soporte inteligentes, así como del modelado del dominio del conocimiento. Actualmente, en varios países existe una falta de estructuras adecuadas, personal y tecnologías de asistencia para apoyar los procesos de diagnóstico e intervención de niños con y sin discapacidades, y trastornos de la comunicación.

Por otro lado, actualmente mil millones de personas viven en el mundo con algún tipo de discapacidad de este grupo, un número significativo de personas presenta problemas de lenguaje. Según las últimas estimaciones de la Organización Mundial de la Salud, solo el 15% de las personas que necesitan tecnologías de asistencia pueden acceder a ellas. Por estas razones, en este proyecto de investigación, se desarrolla un prototipo para extraer información relacionada con las actividades de terapia del habla y lenguaje de las páginas web.

La herramienta informática se basa en procesamiento de lenguaje natural (PLN), ontologías, rastreo web y servidores web para buscar actividades de THL y poblar automáticamente una ontología. La herramienta tiene varios módulos organizados en capas. Esta arquitectura nos permite modificar y ajustar cualquier módulo sin afectar la funcionalidad de los otros módulos y capas.

Finalmente, para la fase de pruebas se desarrolla un entorno basado en arquitectura web en donde un terapeuta puede almacenar y buscar actividades de acuerdo a un conjunto de tesauros almacenados. De la misma manera, el sistema genera automáticamente un gráfico de la ontología, así como un gráfico de estadísticas que contiene las palabras más relevantes de la página web rastreada. Los resultados alcanzados son alentadores, dado que ha sido posible realizar búsquedas automatizadas y recopilar procesos de información relacionados con el campo THL. Finalmente, es importante mencionar que hemos utilizado herramientas de código abierto como Python, Django, RDF2DOT, MariaDB, entre muchas otras.

Palabras claves: Ontologías, Crawler, PNL, SPARQL, Owlready2, Python, Django

2. Abstract

The language and communication are considered mainstays to develop cognitive, social, and psychological skills in children. However, the Speech-Language Therapy (SLT) has not adequately addressed from the viewpoint of the intelligent support tools as well as the knowledge domain modeling. Currently, in several countries exist a lack of adequate structures, personnel, and assistive technologies to support the processes of diagnosis and intervention of children with/without disabilities and communication disorders.

On the other hand, nowadays one billion of persons live in the world with some form of disability. From this group, a significant number of persons present language impairments. According to latest estimates of the World Health Organization, only the 15% of persons that need assistive technologies can access to them. For these reasons, in this research project, we have developed a prototype to extract information related to speech-language therapy activities from web pages.

With this aim, we have developed an informatics tool based on Natural Language Processing (NLP), ontologies, web crawling, and web servers to search SLT therapy activities and automatically populate an ontology. The tool has several modules organized in layers. This architecture allows us modifying/tuning any module without affecting the functionality of the other modules and layers.

To test the real feasibility of our proposal, we have developed a web-based environment where a therapist can search activities accordingly to a set of activities stored in a relational database. In the same way, the system automatically generates a graphic of the ontology as well as a statistics plot containing the most relevant words of the crawled web page. The achieved results are encouraging, given that it has been possible to perform automated searching and collecting processes of information related to the SLT field. Finally, it is important mentioning that we have used open source tools such as Python, Django, RDF2DOT, MariaDB, among many others.

3. Introducción.

La web está diseñada y enfocada al intercambio de información masivo de datos que fluye a través de ella para el beneficio del mundo; al contener y alimentarse de información desde varias fuentes y diferentes medios provoca un desfase que se genera por los diferentes formatos utilizados desde el inicio de su estructura. Este aspecto genera que se produzca caos. Es por ello que la web semántica reemplaza a la web tradicional¹ y se enfoca en la interpretación de datos mediante etiquetamiento, pudiendo interpretar y subdividir a detalle nuestros datos personales, académicos, comerciales, etc., Con ello, es factible obtener no solo un orden en su estructura, sino que las búsquedas en la web semántica sean fáciles de realizar utilizando palabras clave y básicas. Al emplear la web semántica, se puede ejecutar una búsqueda más elaborada y precisa, posibilitando el realizar deducciones lógicas para resolver las consultas requeridas por el usuario [1].

Los buscadores actuales no pueden analizar toda la información existente en la web por ello, solo una pequeña porción de esta información es interpretada por los mecanismos de inteligencia utilizados para las búsquedas, esto se debe a que en el inicio de su implementación los archivos y ficheros fueron cargados a la misma, sin ningún tipo de estándar formal. *“Hablar de Accesibilidad Web es hablar de un acceso universal a la web, independientemente del tipo de hardware, software, infraestructura de red, idioma, cultura, localización geográfica y capacidades de los usuarios”* [2] esta accesibilidad web, engloba a todas las personas e incluye a las que sufren de algún tipo de discapacidad o no, también *“[...] se puede definir la accesibilidad web como la posibilidad de que un producto o servicio web pueda ser accedido y usado por el mayor número posible de personas, indiferentemente de las limitaciones propias del individuo o de las derivadas del contexto de uso”* [3].

Debido a esto, apoyaremos en la ingeniería del conocimiento (IC), que es parte de la inteligencia artificial (IA) y cuyo objetivo principal es extraer, articular y computarizar el conocimiento de un experto. La IA trabaja con el modelado de objetos del mundo real o representación del ser de los objetos (ontología) que es una especificación conceptualizada, explícita y formal de un objeto [4]. Estas ontologías son piezas importantes para la solución de problemas en la IC, siendo ellas el pilar fundamental del área del conocimiento al cual está enfocando este trabajo.

¹ web tradicional. - Se denominada a la web en sus inicios de implementación carente de semántica o a los usos tradicionales en el internet.

4. Objetivos.

4.1. General

Desarrollar un *crawler* semántico para la generación automática de pobladores de ontologías.

4.2. Específicos

- Realizar un estado de arte de los *crawlers* semánticos.
- Diseñar e implementar un módulo basado en ontologías de un dominio específico para llevar a cabo búsquedas semánticas dentro de páginas web.
- Diseñar y desarrollar un módulo que implemente técnicas de Procesamiento del Lenguaje Natural (PLN) a fin de extraer contenido textual de páginas web.
- Dada una ontología básica de un dominio específico, desarrollar un algoritmo que permita poblar dicha ontología de forma automática.
- Diseñar consultas basadas en SPARQL o lenguajes similares a fin de extraer los contenidos almacenados en la ontología.
- En base al corpus construido realizar pruebas con diferentes tipos de herramientas que permiten realizar inferencias (*reasoners*, razonadores) dentro de la ontología que se ha poblado.

5. Estado del Arte.

5.1. Uso de ontologías existentes en proyectos y ámbitos de desarrollo.

De forma global las ontologías han pasado de un concepto metafísico a una nueva forma de conceptualizar información de un dominio de discusión propio. Por citar un ejemplo, la ontología *International Classification of Diseases* (ICD-10-CM), que es una base en la identificación de estadísticas y patrones de salud alrededor del mundo, siendo un estándar internacional para la clasificación de diagnósticos para todos los fines clínicos o de investigación en afecciones, enfermedades, trastornos, o lesiones a la salud. Esta ontología ha tenido un gran impacto a nivel industrial por su integración con sistemas de información, lo que facilita la recopilación de datos médicos, basándose en ICD-10 debido a que especialistas se ayudan en la confirmación de los diagnósticos, en sus revisiones [5].

Dadas ontologías como ICD-10-CM que facilitan al desarrollo y crecimiento de la información acerca de enfermedades, se han generado proyectos de ayuda en el diagnóstico de enfermedades, creando para ello modelos de conocimiento y sistemas de inferencia, que podrían ser utilizados por médicos e investigadores en el ámbito de la salud. Por ellos la viabilidad de web semántica para diversas áreas de la ciencia está ganando aceptación, y esto da paso a que se abran investigaciones en el uso eficaz de tecnologías semánticas [6].

Las ontologías se utilizan de forma frecuente en el ámbito de la investigación como en este caso el proyecto Onto-SPELTRA. Dentro de este proyecto se cuenta con un módulo de soporte a la toma de decisiones que puede adaptarse a diferentes estructuras de información. Este proyecto se sustenta en una ontología y un sistema experto para determinar qué ejercicios y actividades terapéuticas son adecuadas para niños con discapacidades y trastornos de comunicación [7].

Las ontologías pueden emplearse en proyectos y tecnologías desarrolladas con el fin de apoyar en los procesos de inclusión educativa o rehabilitación de niños, jóvenes o adultos con discapacidades o Necesidades Educativas Especiales. Un ejemplo de ello, es el proyecto "*An intelligent system based on ontologies and ICT tools to support the diagnosis and intervention of children with autism*", que tiene como objetivo brindar soporte a niños con Trastorno del Espectro Autista (TE). A través de un ecosistema inteligente basado en la web semántica, las ontologías y las TIC² inteligentes [8].

En la misma línea, dentro del artículo titulado "*A multilayer mobile ecosystem to support the assessment and treatment of patients with communication disorder*", se propone un ecosistema que ha sido diseñado con el objetivo de proporcionar un conjunto completo de funcionalidades para representar y gestionar información relacionada con la terapia del lenguaje. El ecosistema utiliza ontologías, arquetipos y vocabularios estandarizados para describir protocolos médicos, terapéuticos y

³ Tecnologías de Información y Comunicación.

estrategias de intervención aplicando a un entorno móvil multicapa para apoyar la evaluación y el tratamiento de los pacientes con trastornos de la comunicación [9].

FOAF (Friend Of A Friend) es una ontología popular, legible para las máquinas, esta describe la representación de información de perfil personal y relaciones entre personas con objetos. Construida en el lenguaje para la web semántica utiliza el marco de descripción RDF y lenguaje de marcado OWL, FOAF provee una representación de una red de relaciones sociales como una estructura gráfica [10].

5.2. Tecnología semántica y Crawlers Semánticos dentro del campo de la investigación.

Enfocándonos en tecnologías semánticas, se observa que la premisa “*un buscador semántico puede ser la mejor prueba del concepto real de la web semántica*” se cumplió de acuerdo al experimento realizado en 2014 entre tres buscadores influyentes, donde se concluye que Bing® recupera información más relevante en comparación con Google® y Yahoo®. En términos de rendimiento general, Bing recuperó documentos más relevantes en comparación con todos los motores de búsqueda seleccionados por lo que reduciría el problema de sinónimos y de palabras polisemia [11].

Estudios recientes sobre multimedia han recurrido al análisis semántico de alto nivel a partir del análisis semántico fundamental; el objetivo del análisis semántico de alto nivel sobre video es extraer la información semántica en el vídeo. Por ello, la construcción de una biblioteca o repositorio de información semántica completa y razonable se convierten en la base del estudio sobre video semántico, utilizando un web *crawler* y el análisis semántico concluyen afirmando que una biblioteca semántica puede mejorar la precisión del análisis semántico y la profundidad de la minería, para proporcionar soporte adicional en el área de vídeo [12]. “*Los crawlers son un mecanismo, un método o una pieza de código cuyo objetivo principal es navegar a través de la web, con la intención de obtener datos relevantes ejecutándose en un ciclo infinito durante meses con el objetivo de recopilar información relevante*” [13].

Actualmente, las ontologías han permitido mejorar la estructura taxonómica de los *crawlers* básicos mediante la adición de la semántica del tema de búsqueda. Esto es en virtud de que la ontología juega un papel importante en la semántica asociada con el dominio del tema, la ontología se combina con otros enfoques para obtener resultados más eficientes [15].

En esta línea de investigación el estudio titulado “*A Linked Data Based Personal Service Data Collection and Semantics Unification Method*” presentado y realizado por Yixue Zhao y sus colaboradores, en el cual proponen un método de recopilación de datos y unificación semántica (API abierta, rastreo web e importación manual) y luego, en base a algunos servicios de uso frecuente, se crea una ontología unificada de datos personales en forma de Datos Vinculados para facilitar la unificación semántica [16].

En este ámbito, se considera de interés mencionar el prototipo EGOKI, que especifica cómo se puede llegar a utilizar un *crawler* y ontologías actualmente. Este trabajo se enfoca en la automatización de una interfaz accesible para el usuario diseñadas y pensada para permitir que las personas con discapacidad accedan a servicios ubicuos, EGOKI integra ontologías para el almacenamiento de modelos de enfoque obtenidos a través de *web crawling*, con esto selecciona los recursos y modalidades adecuados de interacción en función de las capacidades de los usuarios. El prototipo fue probado en dos escenarios: uno dedicado a las personas con discapacidad visual y el otro con personas con impedimentos cognitivos. El resultado del prototipo muestra que la herramienta fue capaz de generar automáticamente interfaces de usuario accesibles y operables [17].

De esta manera el enfoque de los nuevos prototipos y desarrollos toma en consideración la adaptabilidad y la búsqueda de patrones que ayudan al aprendizaje de forma didáctica cómo es el caso de RAMSES, es un proyecto inteligente que se basa en un conjunto de aplicaciones móviles y en una interfaz robótica para otorgar soporte durante el proceso de SLT, con el objetivo de proporcionar un control completo de la terapia personalizada [18].

6. Diseño de la propuesta planteada.

6.1. Web Semántica.

Fue Tim Berners Lee, considerado el fundador de la Web y parte de World Wide Web Consortium (W3C), quien acuñó el término “*Web Semántica*”, iniciando una revolución en el manejo de información. Lee junto con algunos compañeros y colaboradores, propusieron las principales ideas sobre la Web Semántica, donde se plantea que la nueva web debía ser “*una extensión de la Web tradicional en la cual la información tiene un significado bien definido, propiciando el trabajo cooperativo entre computadora y personas*”. De igual forma se destaca que “*la Web Semántica aporta estructura al contenido significativo de las páginas Web, creando un entorno donde los agentes de software se moverán de una página a otra fácilmente llevando a cabo tareas sofisticadas para los usuarios*” [19].

En virtud del crecimiento exponencial de la cantidad de información existente en la Web tradicional, surge la necesidad de gestionar la gran cantidad de datos; proceso que se ejecutaba recuperando la información de forma no estructurada por falta de relación entre la información; se necesitaba encontrar una solución que resolvería la limitación mencionada que presenta la Web tradicional para obtener un mejor procesamiento automatizado de la información e incluso una indexación que permitía manejarla de forma eficiente los datos. A esta nueva forma de interpretar y estructurar los datos dentro de la web tradicional se le denomina web semántica.

6.1.1. Búsquedas y tecnología en la Web Semántica.

La Web Semántica dotada de procesos que ayudan a encontrar en la Internet respuestas más rápidas y sencillas, al utilizar información mejor definida, se estandariza de forma semántica, siendo el resultado una mejor interconectividad entre procesos de búsquedas y datos detallados que hace posible compartir, procesar y transferir. Esta web se apoya en lenguajes universales mejorando los problemas existentes en la web tradicional; haciendo un cambio de paradigma, la web semántica es capaz de procesar, razonar la búsqueda, relacionarla y realizar deducciones lógicas [20].

La recuperación de información de datos no estructurados, así como la falta de relación entre ellos hace que se requiera tecnologías para realizar estos fines, por esta razón presentamos las tecnologías más relevantes.

6.1.1.1. XML (Extensible Markup Language)

XML proporciona una sintaxis que marca a un documento, creando una relación sencilla y a la vez, otorgándole de una estructura que mediante etiquetas lo hacen legible al procesamiento de máquina.

XML se plantea como un estándar para el intercambio de información estructurada, cumpliendo a cabalidad las funciones para las que fue diseñado, Sin embargo, es limitado en el manejo de metadatos, para ello se crea un estándar que vincula a XML y los metadatos otorgando información sobre los propios datos volviéndolos altamente estructurados, al describir su contenido y siendo utilizados para la expresión del conocimiento objetivo en agentes de búsqueda semánticos. Los metadatos se vuelven un elemento importante en estructuras carentes de semántica dando así RDF (*Resource Description Framework*), como resultado [21].

6.1.1.2. RDF (Resource Description Framework)

RDF es un modelo estándar creado para el intercambio de información sin pérdida de significado, facilitando la representación de conocimiento en un entorno distribuido, haciendo posible relacionar objetos con el estándar URI (*Uniform Resource Identifier*). Dicho estándar se emplea en gran parte de los archivos o elementos alojados en la web a fin de asignarles una identificación y ubicación. De igual forma, URI dota de características como protocolos, nombre del servidor, ruta de direcciones y nombre del archivo, permitiendo ser un localizador URL (*Uniform Resource Locator*), ya que en éste último puede variar o cambiar de ubicación con el tiempo [20] [21].

RDF se basa en un formato serializable sencillo y funcional de etiquetas llamado tripletas o declaraciones, con esto se logra una descomposición del conocimiento a una forma básica; la Figura 1, representada en sujeto, predicado y

objeto sirve como lineamiento para la descripción de un dato, haciendo su almacenamiento y procesamiento fácil [20] [23].

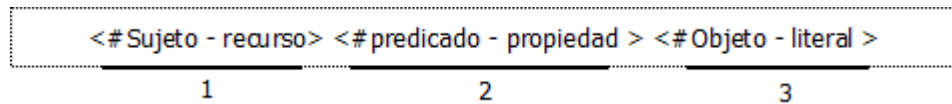


Figura 1: estructura RDF básica. [23]

6.1.1.3. OWL (Web Ontology Language)

Se crea como una extensión del modelo de datos de RDF, describiendo su significado implícito y manejando una conceptualización del contexto de los elementos. Al proveer semánticas en la construcción de modelos de datos complejos, vocabularios y lógica de software, esta se define como un conjunto de funciones orientadas a objetos, que relaciona las tripletas RDF a clases, dando asociaciones y relaciones complejas [21].

OWL gestiona objetos como clases, sub-clases e instancias y relaciones que permiten dar más significado a los predicados dentro de RDF, también entre otras operaciones permite definir relaciones entre clases, cardinalidad de atributos, equivalencia de clase o propiedades, clases enumeradas y proporciona propiedades a los atributos que agregan valor semántico.

De esta forma y en este OWL se define una ontología o “*vocabulario de términos que se formalizan a menudo cubriendo un dominio específico y que es compartido por una comunidad de usuarios*” [23] [24], siendo este un objetivo de la web semántica compartir, integrar y reutilizar los datos para su interoperabilidad mediante las ontologías.

6.1.1.4. SPARQL (Protocol and RDF Query Language)

SPARQL es un lenguaje estandarizado y creado especialmente para la consulta sobre recursos RDF, este lenguaje describe cómo realizar consultas y cómo recuperar sus resultados. Es flexible, ya que permite consultar patrones requeridos y opcionales en grafos (conjunto de tripletas RDF) de diferentes fuentes almacenadas a través de RDF.

Al soportar adición, subconsultas, negación, este lenguaje diseñado para consultas que forma parte de la web semántica, junto con sus conjunciones y disyunciones permite la combinación dinámica de varias fuentes de información creando nuevas fuentes de información, permitiendo que mediante los grafos RDF la consulta se acerque a una naturaleza semi-estructurada del mundo real, brindando la principal forma de hacer una consulta sencilla con sintaxis igual a SQL a una gran cantidad de datos sobre recursos RDF [24].

6.1.1.5. HTML (HyperText Markup Language)

HTML es la estructura propuesta por W3C, crea una relación de hipertextos basados en identificadores únicos, resolviendo el problema de pérdida de información, los hiperenlaces identifican y relacionan a un recurso dentro de una red mediante una cadena de caracteres o dirección URI estas direcciones estructuran una comunicación hacia otros documentos, páginas o incluso a otros servicios [26].

Utilizado como el lenguaje de estructuras gráficas, es una parte importante de la misma, debido a su simplicidad y estructura provee de una estructura visual a una página web, ayudando no solo en su estructura gráfica en los navegadores, sino que su código permite etiquetamiento semántico al momento de su desarrollo, indicando mediante etiquetas que tipo de contenido posee, haciendo que los buscadores semánticos puedan realizar un procesamiento de clasificación de mejor forma [26] [23].

```
<!DOCTYPE html>
<html Lang="es">
  <head></head>
  <body>
    <nav></nav>
    <section></section>
    <aside></aside>
    <footer></footer>
  </body>
</html>
```

Figura 2: Código de etiquetas HTML [23].

En base a ello, se hace de manera transparente la integración de RDF y HTML para el usuario mediante RDFa (Resource Description Framework in Attributes) que permite enriquecer la información de una página web, proporcionando atributos de marcado legibles para el lenguaje de máquinas, ayudando a los desarrolladores de páginas web a brindar un adecuado etiquetamiento y sin ambigüedad, permitiendo que se utilice una semántica significativa, con este proceso los motores de búsqueda pueden clasificar de mejor manera la información contenida en una página web [23].

6.2. La Web Semántica Y Ontologías.

El etiquetado en la web semántica, sirve para procesos de análisis y búsqueda de contenidos utilizando el etiquetamiento para precisar la recolección de la información, tomando en cuenta que una gran parte de información publicada, subida y alojada específicamente en la última década en la web, tiene algún tipo de referencia semántica, haciéndola interpretable desde su creación u origen; la web semántica

mejora el referenciación, al apoyarse en lenguajes estructurados como XML y RDF [23].

La web semántica tiene como objeto primario compartir datos y su integración a través del uso de ontologías, se tiene que aclarar que no todo el trabajo está enfocado en una sola ontología universal de búsqueda que abarca todo el vocabulario que se encuentran en la misma, sino que se reutiliza una gran cantidad de pequeñas ontologías.

La definición de cada ontología es compleja, pero tomando en cuenta el modelo de datos al que se enfoca y el proceso OWL esto puede variar; “no existe una sola ontología correcta para cualquier dominio” [21] debido a esto su creación es un proceso creativo y científico que puede ser subjetivo.

6.3. Ontologías.

En menos de veinte años, el término "ontología", originalmente prestado de la filosofía metafísica, se ha popularizado considerablemente en algunas ciencias y en los sistemas de información. Probablemente la popularidad es su concepto evolutivo de lograr la interoperabilidad entre múltiples representaciones de la realidad, por ejemplo, datos o negocios y modelos de proceso que residen dentro de sistemas informáticos y entre dichas representaciones y realidades, es decir, humanos y su percepción de la realidad [23].

Actualmente las ontologías nos permiten representar un gran número de relaciones complejas que los sistemas de información tradicionales no soportan, facilitando el manejo de información y su relación, con esto se cubre la necesidad de que el conocimiento sea representado de una forma legible. Una ontología es un vocabulario consensuado y reutilizable de objetos que da comienzo a la representación del conocimiento a diferentes niveles de formalismo a los términos y relaciones entre ellos, permitiendo expresar algún concepto dentro de un dominio de interés [23].

6.4. MODS (Marco Ontológico Dinámico Semántico)

El MODS es una propuesta novedosa para procesos de análisis, debido a que permite realizar consultas en lenguaje natural en la web semántica; permitiendo formalizar e interpretar consultas por los usuarios.

La consulta procesada es un elemento cargado de una cantidad de información, el análisis de esta información es útil para llegar a tener un tipo de perfil de usuario, ya que con esto aproximar y de forma sucesiva encontrar las respuestas que satisfaga sus necesidades con respecto a la consulta o búsqueda realizada [27].

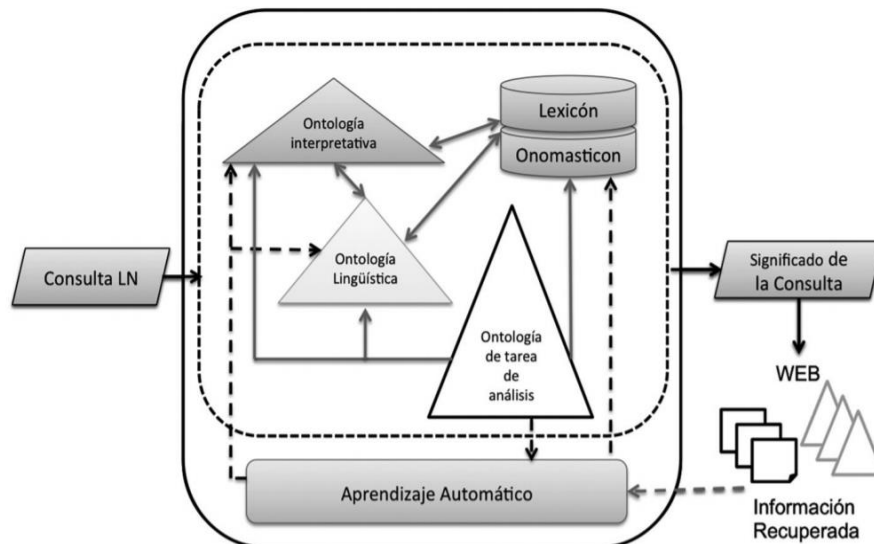


Figura 4 : Marco Ontológico Dinámico Semántico [27].

MODS no solo transforma la consulta realizada en una consulta fácil de interpretar por el lenguaje OWL, también su desafío es interpretar y formalizar la consulta hecha por el usuario en lenguaje natural y refinarla, explotando su contenido semántico para procesos de razonamientos automáticos con el fin de optimizar procesos de búsqueda [27].

6.5. **Crawler.**

Un crawler es un sistema que consta de un algoritmo de análisis, metódico y automatizado que tiene como fin obtener la información de hipervínculos y meta descripciones, contenido en un sitio en la web; este sistema busca páginas de alta relevancia para tener un mayor nivel de precisión. El objetivo es priorizar las páginas ya encontradas y gestionando la exploración de hipervínculos. Un *crawler* ayuda a la descarga del contenido de una página priorizada o la página web de relevancia para un tema en particular, la relevancia de una página se determina según los parámetros semánticos otorgados, al obtener los datos puros de una página (la estructura html o código fuente de la página), con estos datos se puede aplicar y ejecutar sobre ellos, nuevamente el proceso básico inicial del *crawler*, analizar y recolectar según los datos obtenidos según los parámetros semánticos datos.

6.6. **Procesamiento de lenguaje Natural.**

Al hablar del procesamiento de lenguaje natural, obligatoriamente nos adentramos en lenguaje natural (LN), ya que es el medio que utilizamos de forma repetitiva y cotidiana para establecer nuestra comunicación. Siendo una herramienta para razonamiento con gran función y valor, sin embargo, la sintaxis de un LN puede ser modelada fácilmente por un lenguaje formal [25].

El Procesamiento de lenguaje natural (PLN) consiste en la utilización de un lenguaje natural (LN) para una comunicación hombre-máquina, de tal manera que una máquina sea capaz de entender las oraciones que se le proporcione, al utilizar un LN, se facilita gran parte del desarrollo de programas que realicen tareas con lenguaje natural, apoyándose en los mecanismos ya creados [25].

7. Desarrollo de la propuesta planteada.

7.1. Arquitectura del sistema.

La propuesta planteada tiene una arquitectura que se organiza en varias capas, módulos y servicios que se brindarán a través de nuestro sistema nombrado **SPELTA-ADAs (SPELTA-Automatic Data Searcher)**. Este aspecto permite realizar cambios y mejoras en cualquier elemento sin afectar las funcionalidades de las demás capas. En la Figura 6 se pueden apreciar los elementos más importantes que constituyen el ecosistema desarrollado y que se detallan seguidamente:

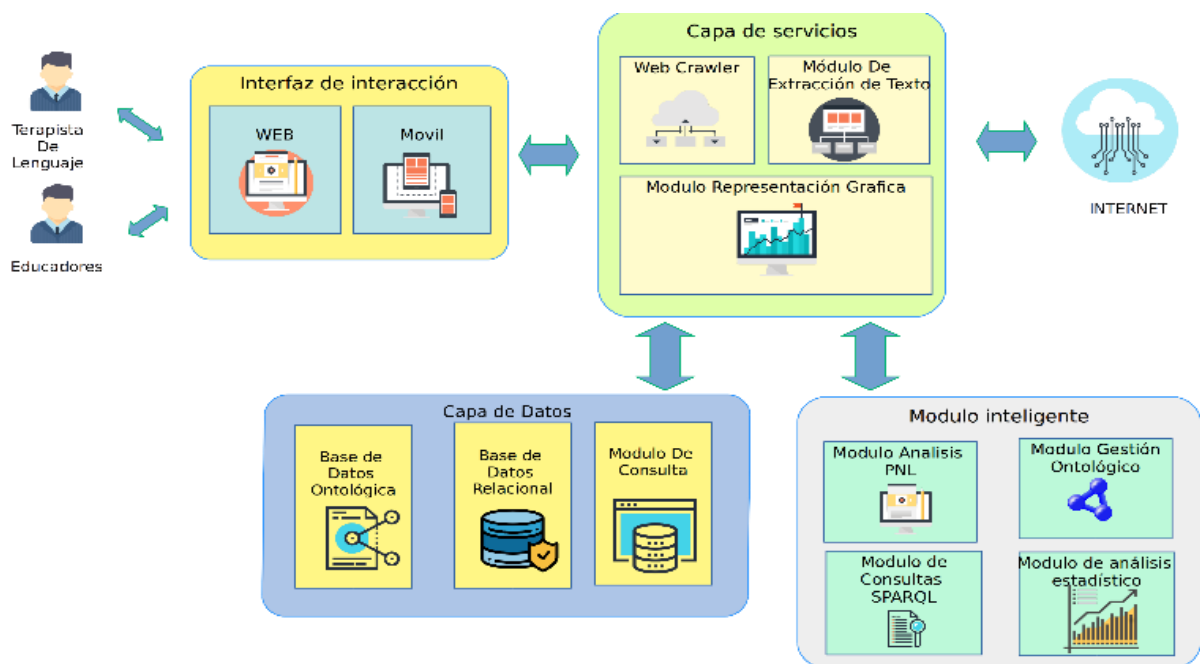


Figura 5 : Propuesta de la Arquitectura de SPELTA-ADAs (SPELTA-Automatic Data Searcher).

En la **capa de interacción** el ecosistema provee diversos servicios que permiten que los terapeutas de lenguaje, educadores, investigadores o usuarios expertos puedan acceder desde varios entornos (web, móvil) con el objetivo de dar versatilidad al sistema.

En la **capa de servicios** el ecosistema implementará una serie de funcionalidades que se utilizarán durante las diversas etapas de procesamiento y extracción de información con especial énfasis, en garantizar el correcto análisis de los datos obtenidos; teniendo así el **módulo de crawler** que permite llevar a cabo la búsqueda y obtención de información en la web semántica, mediante palabras clave,

en nuestro caso actividades presentadas mediante la capa de interacción, misma que obtiene los datos de una base de actividades previamente poblada. El *crawler* al lograr una recolección relevante de información da paso al **módulo de extracción de texto**, encargado de procesar en archivos las páginas web encontradas, almacenando su contenido semántico. En esta capa también tendremos el **módulo de representación gráfica** que utilizaremos posteriormente para obtener la gráfica de comprobación de nuestra ontología poblada.

En la **capa del módulo inteligente** implementa 4 módulos que permiten ejecutar acciones relacionados con el procesamiento de lenguaje, gestión de ontologías, análisis estadístico de los datos y consultas a la ontología poblada. El **módulo de procesamiento de lenguaje natural** realiza un análisis comparativo de las palabras claves buscadas y los datos almacenados mediante el *crawler*. El **módulo de gestión ontológica** implementa un modelo ontológico basado en la programación, que brindan soporte orientado a objetos empleado en lenguajes de programación de alto nivel. El **módulo de análisis estadístico** se implementa para la gráfica de patrones de palabras más utilizadas en el proceso de búsqueda. Por otra parte, el **módulo de consultas SPARQL** constituye un soporte en la consulta de información en nuestra ontología poblada.

La **capa de datos** de este ecosistema tendrá como funcionalidad el manejo y administración de datos, es integrada por el **módulo de consulta**, que expondrá mediante un *framework* web desde la capa de interacción que se ocupa de presentarlos de forma organizada; esta capa contiene también la **base de datos relacional**.

7.1.1. Capa de interacción.

Con el fin de darle mayor flexibilidad a la aplicación se crea la capa de interacción, la cual se encargará de la conectividad, siendo un *front-end* de entorno web, utiliza un *framework* de alto nivel basado en Python, que permite a los módulos interactuar entre sí.

Esta capa utiliza a Django, por ser un conjunto de módulos dedicados al desarrollo rápido y de diseño limpio sobre la web, teniendo como ventajas ser de libre distribución y de código abierto, facilita el desarrollo de nuestra aplicación, puesto que se encarga de gran parte de las dificultades que se pueden encontrar durante el desarrollo web, permitiendo crear cualquier aplicación sin necesidad de reinventar la rueda [28].

Este *framework* permite realizar nuestro desarrollo con una estructura MTV (*Model, Template, View*), la cual implementa capas según la necesidad del proyecto.

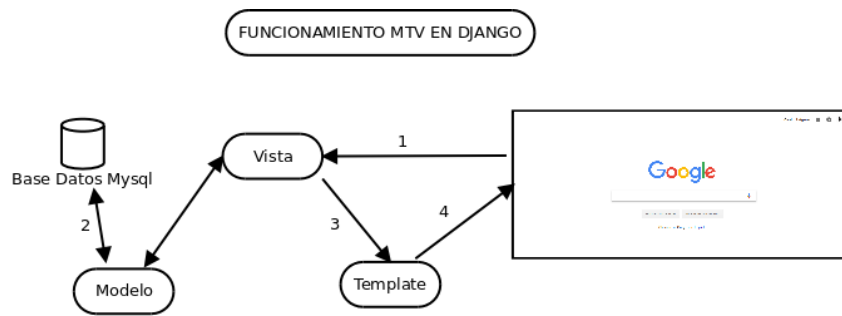


Figura 6 : Modelo MTV en Django

En la Figura 6 podemos observar el funcionamiento del modelo *MTV* antes mencionada.

- El navegador envía un *request* o solicitud a la vista.
- La vista interactúa con el modelo para obtener o procesar la orden.
- La vista hace una llamada a la plantilla.
- La plantilla renderiza un *response* o respuesta a la solicitud realizada en el paso 1.

7.1.2. Capa de servicios.

Esta capa contiene un conjunto de módulos, los cuales inician el proceso, al realizarse una petición mediante la interfaz web. A continuación, explicaremos los módulos que intervienen:

7.1.2.1. Módulo de Crawler.

El método de *crawling* de datos es una base fundamental para nuestro proyecto, debido a que se plantea utilizar estratégicamente esta herramienta en el análisis de información a gran escala, para ello utilizaremos la siguiente herramienta.

Scrapy.

Scrapy es un *framework* de aplicaciones que está escrito en Python con *Twisted*, este último es un popular marco de trabajo de *networking* basado en eventos; a *Scrapy* se lo utiliza para rastrear sitios web. En este proyecto lo utilizaremos para extraer datos estructurados que pueden ser utilizados en una amplia gama de análisis y como minería de datos para el procesamiento de la información obtenida [29], la Figura 7 nos demuestra gráficamente como interactúa el modulo con la capa de interconexión.

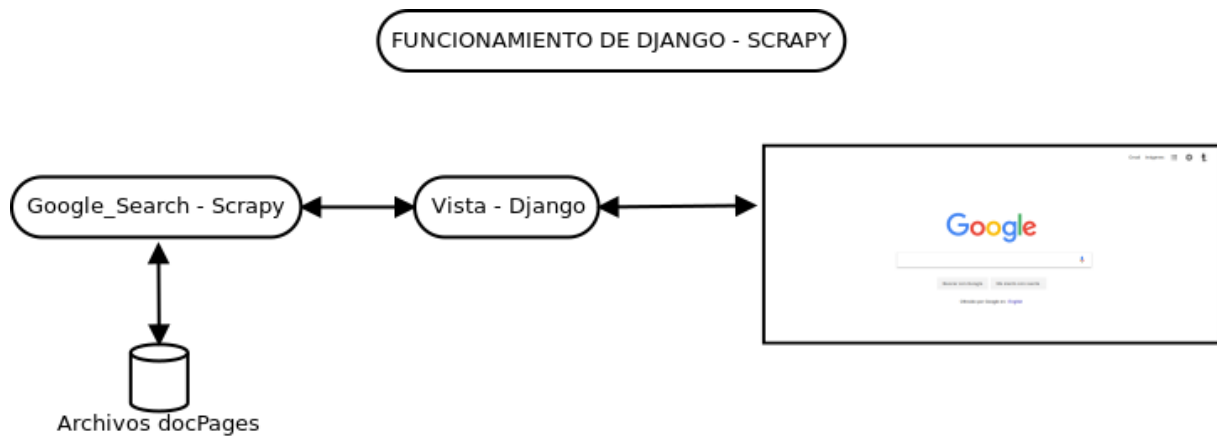


Figura 7 : Funcionamiento de Django - Scrapy

7.1.2.2. Módulo extracción de datos.

Mientras el módulo de *crawler* se ejecuta, el módulo de extracción de datos permite obtener el contenido semántico de una página web, encargándose de limpiar las etiquetas HTML existentes y almacenar este contenido de forma estructurada en archivos de texto que posteriormente se utilizan en el análisis semántico.

7.1.2.3. Módulo de representación gráfica.

Ayudará a representar de forma gráfica la ontología poblada, utilizando **RDF2Dot** esta simple librería basada en Python, es utilizada en gnu-linux para convertir un archivo RDF a dot. Al manejar dot podemos utilizar el lenguaje descriptivo para texto plano, que facilita de forma simple crear gráficas que pueden ser entendibles por humanos y maquinas.

7.1.3. Capa de módulo inteligente.

Esta encargada del proceso analítico semántico con la información recolectada anteriormente.

7.1.3.1. Módulo PNL (*Natural Language Processing*)

El mismo que permite un análisis mediante librerías de procesamiento multilingüe automático, que proporcionará información relacionada a las palabras claves buscadas, realizando un procedimiento matemático y estadístico, buscará similitudes semánticas en el repositorio de archivos generado por el módulo de extracción de texto; analizará todos los archivos dando como resultado un **ranking** que contiene un vector con varios pesos estadísticos de similitud y su respectiva oración asociada al mismo, estas oraciones contendrán una relación con las palabras claves utilizadas al inicio del proceso.

NLTK (Natural Language Toolkit)

El idioma en el que estamos trabajando es español, por esto utilizaremos NLTK [31] el cual es una plataforma líder para desarrollar programas en Python especializado en análisis del lenguaje natural. NLTK proporciona interfaces fáciles de usar con más de 50 recursos de corpus de datos y un conjunto de bibliotecas de procesamiento de texto, brinda clasificación, tokenización, derivación, etiquetado, análisis y razonamiento semántico para bibliotecas de PNL [32].

Freeling.

Es una librería de código abierto para el procesamiento multilingüe, que proporciona una amplia gama de funcionalidades de análisis para varios idiomas, se acopla a cualquier entorno de desarrollo, al ser una librería que puede ser llamada desde cualquier aplicación de usuario que requiera servicios de análisis del lenguaje [33]; al utilizar el lenguaje español, *Freeling* presta las condiciones necesarias para ser incorporado en el desarrollo, por su soporte de análisis lingüístico en español, que fue abarcado de forma sólida desde su versión estable anterior 2.2.

7.1.3.2. Módulo de gestión ontológica.

La gestión ontológica permite la manipulación de una ontología ya sea de forma local o publicada en la web, razón por la que utilizaremos el siguiente modulo.

OWLready2.

Owlready2 es un módulo para Python que permita la programación orientada a ontologías. Este módulo permite la carga de ontologías en forma de clase objeto dentro de Python, con esto podremos modificar, guardar y realizar razonamientos a través de Hermit [30]. Por ello este módulo permite un acceso transparente a las ontologías en OWL, mediante lenguaje de alto nivel [34].

7.1.3.3. Módulo de análisis estadístico.

Analiza y procesa el **ranking** utilizando NLTK, este se genera en el proceso del módulo de PNL. Para obtener las palabras que aparecen con mayor frecuencia, se analiza sustantivos, verbos, predicado y adjetivos de las oraciones dentro del ranking, haciendo un proceso estadístico con ellas, al ejecutarse el módulo estadístico, como resultado tendremos una gráfica aseverando que, mediante un proceso de análisis semántico, se puede obtener las palabras que aparecen con mayor frecuencia.

7.1.3.4. Módulo de consulta SPARQL.

Dado que con owlready2 podemos gestionar una ontología, esta librería me permite hacer consultas de la misma. Utilizando el lenguaje SPARQL, se carga la

ontología poblada y mediante la capa de interacción podemos hacer consultas para confirmar la inserción de datos.

7.1.4. Capa de datos.

El objetivo de esta capa es manipular la información de la base de datos y dejar un entorno de comunicación con nuestra capa de servicios.

7.1.4.1. Módulo de consulta.

Al emplear una base de datos se necesita exponer el contenido del sistema, ya que requerimos de consultas, inserciones, actualizaciones o eliminación de registros; el módulo de consulta es quien gestionara estas acciones mediante peticiones en nuestro entorno web.

MySQL.

Es una base de datos utilizada para desarrollo web, como se desarrollará una aplicación sin fin de lucro el *MySQL* es gratuito, los alojamientos en la web manejan administradores de esta base por defecto [35].

Base de Datos Relacional.

La base de datos relacional principalmente, contiene registros que representan actividades de terapia de lenguaje y habilidades relacionadas con el lenguaje y la comunicación. Cada par de registro actividad-habilidad representa el proceso de intervención que se llevará a cabo dentro de un plan de terapia del lenguaje.

Base de Datos Ontológica.

La base de datos ontológica contiene las nuevas actividades que se buscan en Internet. Estas actividades se guardan en forma de tripletas RDF.

8. Funcionamiento de la interfaz (SPELTA-ADaS)

Como parte de la propuesta planteada se desarrolla un software que representa lo expuestos en la arquitectura.

Tomando en cuenta que los casos de estudio son muy amplios y las ontologías complejamente extensas citamos un ejemplo demostrado en la Figura 8. Esta imagen expone una ontología con un alto grado de complejidad referencial.

Por esa razón el desarrollo propuesto se delimita en el análisis y población de una pequeña porción de la misma; Relacionando Actividades y habilidades, tomando como referente inicial actividades dadas por un terapeuta obteniendo terapias relacionadas con la actividad inicial y alimentando la base de conocimiento establecida.

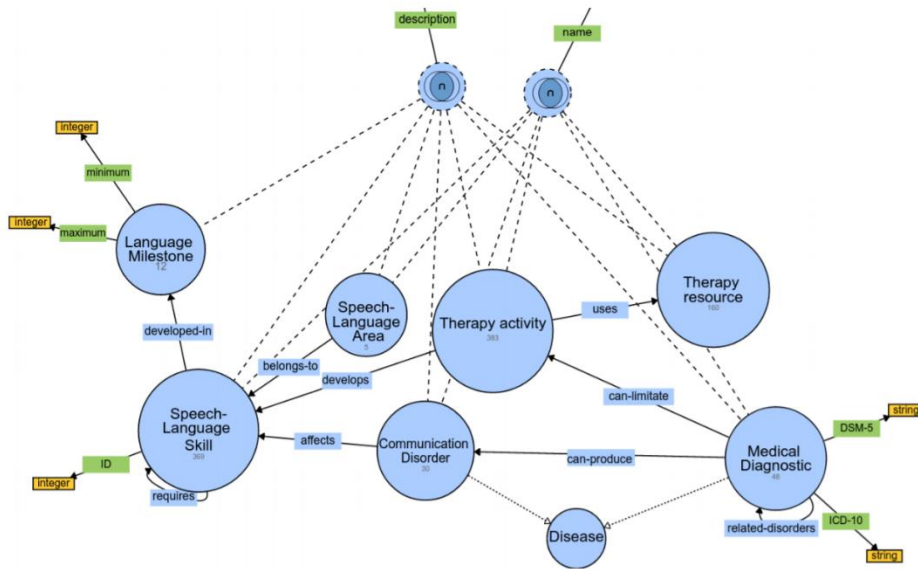


Figura 8 : Imagen extraída del artículo "An ontology-based expert system to generate therapy plans for children with disabilities and communication disorders", esta es una captura parcial de la ontología original.

Al ser un evento cíclico el análisis, extracción y obtención de datos de forma inferente para la población de la ontología, se tomará de referencia la base de las actividades de una terapia, por consiguiente, el software descrito a continuación nos ayudará a realizar los procesos necesarios para la población de una actividad en una ontología.

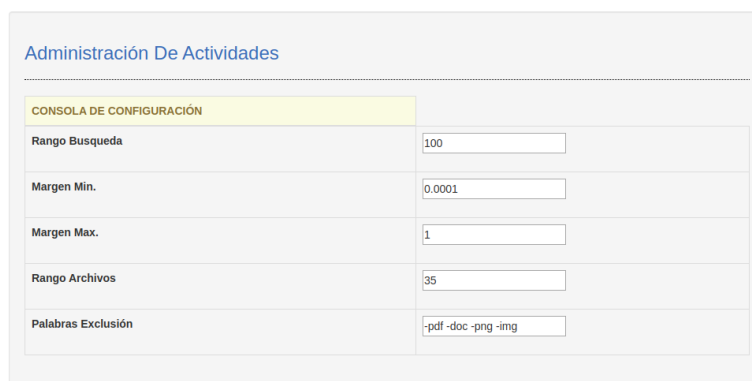
El entorno web ha sido diseñado para partir de un menú representado en la Figura 9 el cual permitirá al usuario una fácil navegación.



Figura 9 : Menú del entorno web, izquierda un navegador en Gnu-Linux y a la derecha un navegador en android.

A continuación, presento la funcionalidad de los menus.

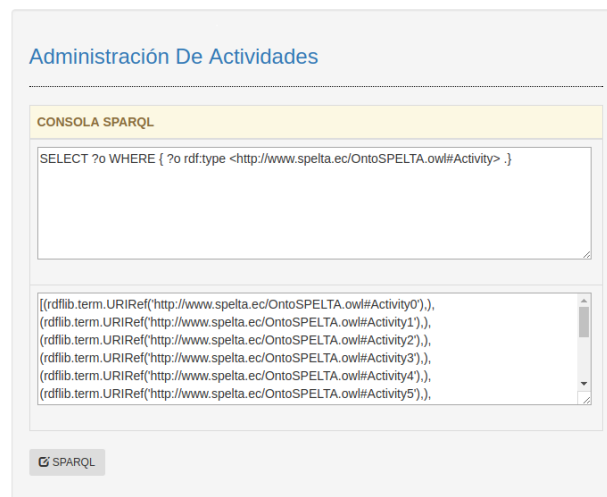
Configuración este botón permitirá configurar parámetros que se manejan dentro del sistema, como se observa en la Figura 10. La consola de configuración contiene un **rango de búsqueda** mismo que es el rango de páginas que limita la búsqueda de nuestro crawler, **margen mínimo y máximo** son los que se encargaran del umbral al momento de procesar el ranking, por contener pesos estadísticos necesita tener un rango configurable para su proceso. Las **palabras exclusión** son palabras o abreviaturas que deseamos excluir en nuestra búsqueda, en este caso pondremos como ejemplo “-pdf -doc -png -img” lo cual permitirá que nuestra búsqueda excluya a los archivos pdf, documentos, imágenes png y img



| CONSOLA DE CONFIGURACIÓN | |
|--------------------------|---------------------|
| Rango Busqueda | 100 |
| Margen Min. | 0.0001 |
| Margen Max. | 1 |
| Rango Archivos | 35 |
| Palabras Exclusión | -pdf -doc -png -img |

Figura 10 : Pagina de configuración del SPELTA-ADaS

Módulos: este botón nos lleva a la pantalla que podemos visualizar en la Figura 11, la cual permitirá realizar consultas SPARQL, sobre la ontología poblada.



```
SELECT ?o WHERE { ?o rdf:type <http://www.spelta.ec/OntoSPELTA.owl#Activity> . }
```

[[rdfib.term.URIRef("http://www.spelta.ec/OntoSPELTA.owl#Activity0"),
(rdfib.term.URIRef("http://www.spelta.ec/OntoSPELTA.owl#Activity1"),
(rdfib.term.URIRef("http://www.spelta.ec/OntoSPELTA.owl#Activity2"),
(rdfib.term.URIRef("http://www.spelta.ec/OntoSPELTA.owl#Activity3"),
(rdfib.term.URIRef("http://www.spelta.ec/OntoSPELTA.owl#Activity4"),
(rdfib.term.URIRef("http://www.spelta.ec/OntoSPELTA.owl#Activity5"),

SPARQL

Figura 11 : Pantalla de Consultas SPARQL

Módulos: botón de procesamiento, es quien nos permite ir a la Administración De Actividades, véase en la Figura 12, aquí es representado lo que realiza la capa de integración y la capa de consulta, ya que gracias a ellas se presenta los datos de forma organizada.

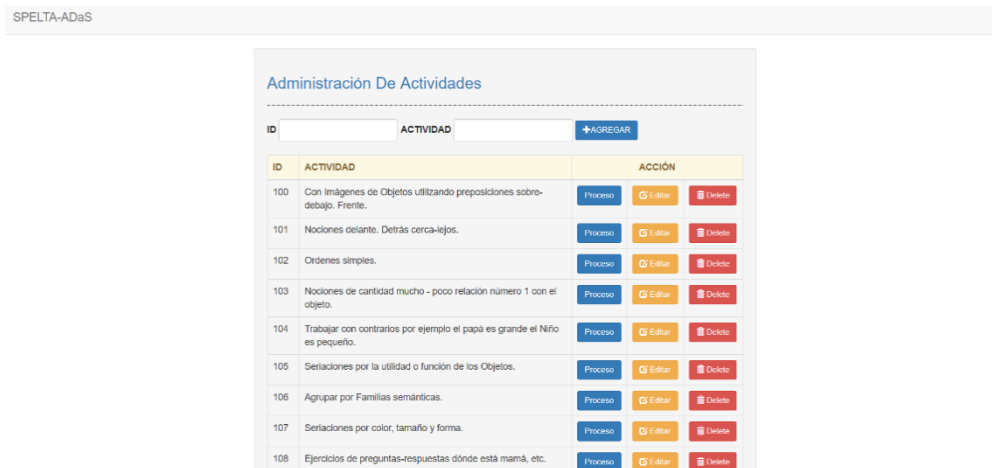


Figura 12 : Administración De Actividades

La página inicial presenta una intuitiva gestión de datos de las actividades en la cual podemos agregar, editar y eliminar una actividad, el botón de Proceso cumple otra función más avanzada ya que con el ingresamos a la siguiente pantalla, Figura 13.

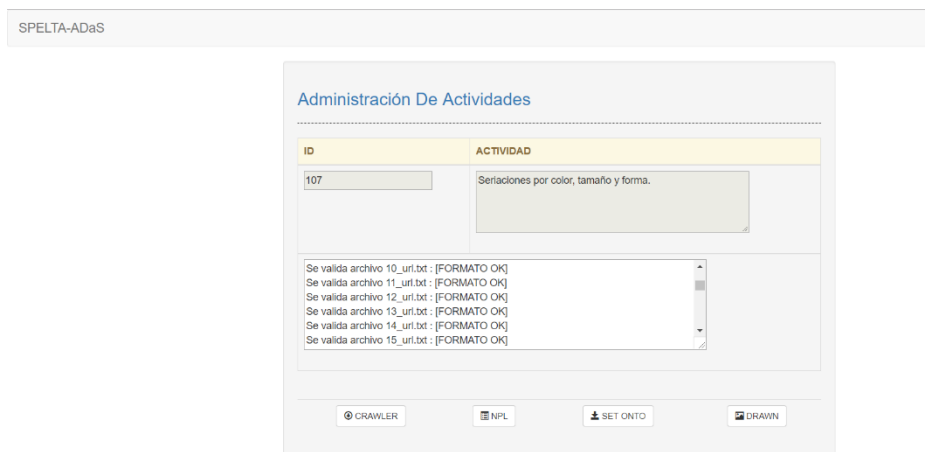


Figura 13 : Pantalla donde el botón de proceso referencia.

Para este procesamiento secuencial debe realizarse paso a paso los 4 botones creados, que representan un subproceso cada uno ejecutan los módulos del esquema propuesto.

CRAWLER: inicia el módulo de *crawler* este se encarga de recorrer los enlaces de las páginas web una a una de forma automática y sistemática almacenando por cada iteración un archivo con la página completa, así mismo el módulo de extracción de texto recorre los archivos almacenados y los deja listos para el modulo PLN

NPL: inicia el módulo de PLN sobre los archivos guardados analizándolos de forma semántica.

SET ONTO: ejecuta el módulo de gestión ontológica y módulo de análisis estadístico, ya que genera la gráfica de las palabras más utilizadas antes de ser almacenadas, en la Figura 14 se demuestra cómo se representa la gráfica y su análisis.

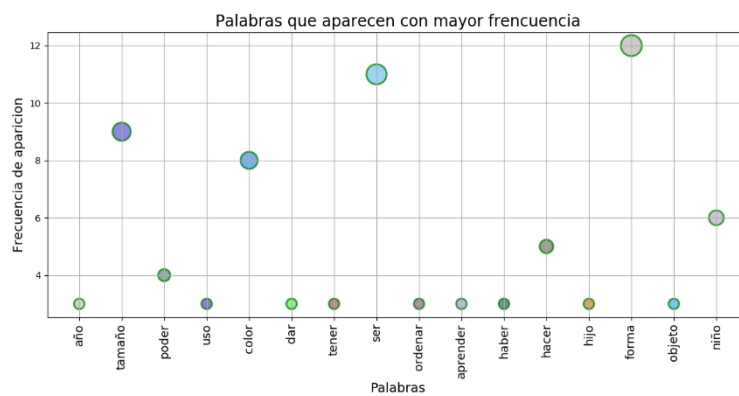


Figura 14 : Gráfica de palabras que aparecen con mayor frecuencia

DRAWN: ejecuta el módulo de representación gráfica, sin embargo, la gráfica de la ontología ocupa un gran espacio para poderla mostrar por ello trataremos de simplificar con la Figura 15, una sección de la ontología creada.

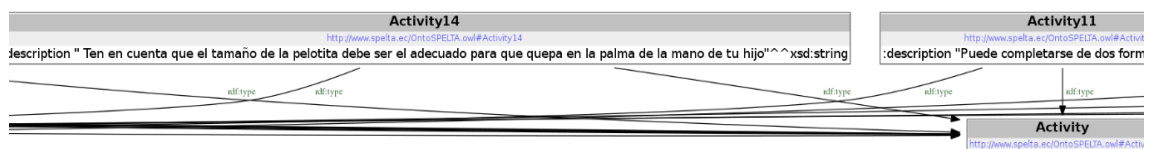


Figura 15 : Sección de la ontología que presenta el módulo de representación gráfica.

9. Experimentación con Razonadores

9.1. Experimentación.

Para nuestras pruebas hemos utilizado el proyecto *Protégé*, que es un framework hecho en Java que nos permite editar ontologías y construir sistemas inteligentes, en nuestro caso será usado para analizar una terapia del lenguaje.

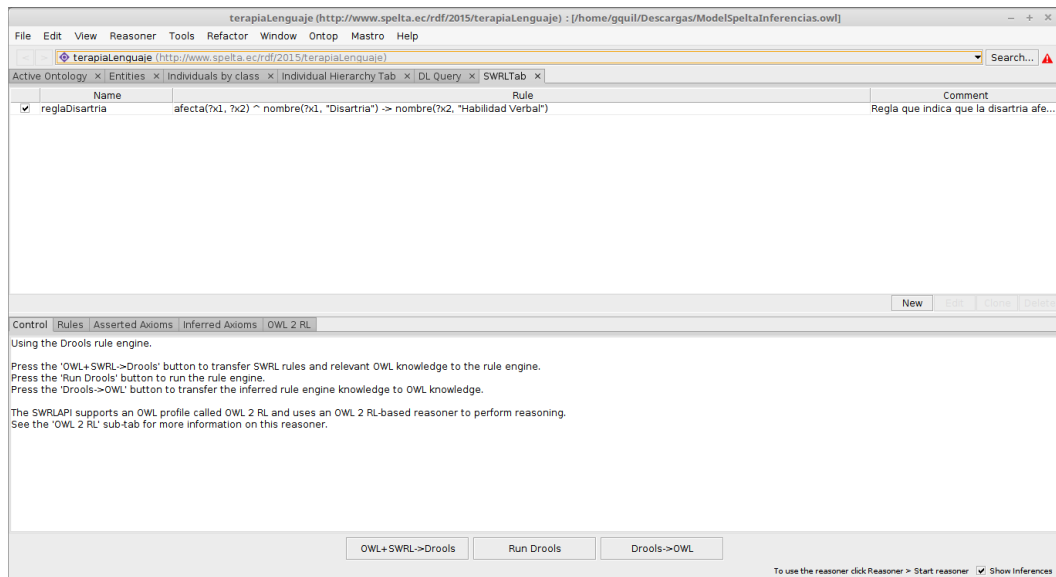


Figura 16 : Protégé y la ontología de terapia Lenguaje importada en él.

Protégé permite realizar análisis con razonadores, es por esta razón que creamos una regla para realizar la prueba, véase en la Figura 16; para esta prueba se utiliza **SWRL** (*The Semantic Web Rule Language*) que nos ayudara a determinar según la regla dada, determinar un axioma que este dentro de la misma. En la Figura 17 podemos apreciar el axioma.

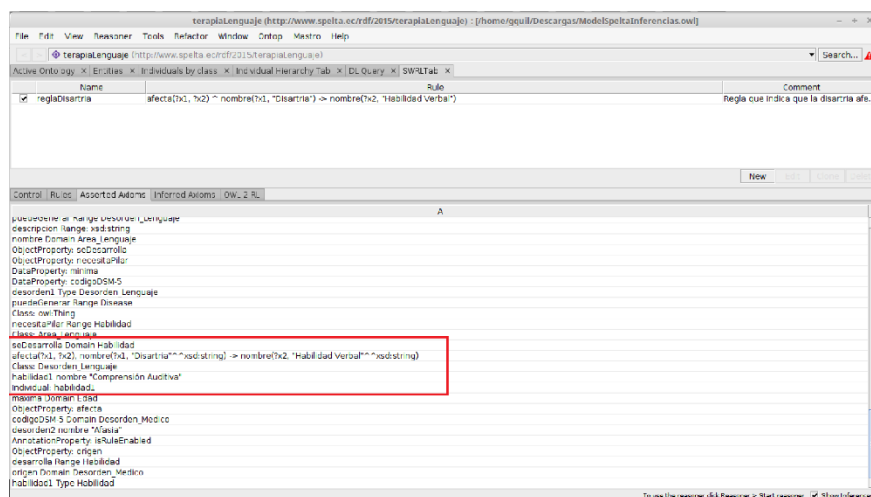


Figura 17 : Resultado del razonador SWRL

El axioma que el razonador determino dada la regla demuestra que la habilidad de comunicación afectada es comprensión auditiva.

También se realizó pruebas con el razonador **Hermit** dentro de la misma plataforma *Protégé* obteniendo como conclusión, de acuerdo a las reglas creadas, este razonador genera una asociación entre los desórdenes y las habilidades, en la Figura 18, se presentan los siguientes desórdenes declarados para la ejecución de pruebas.

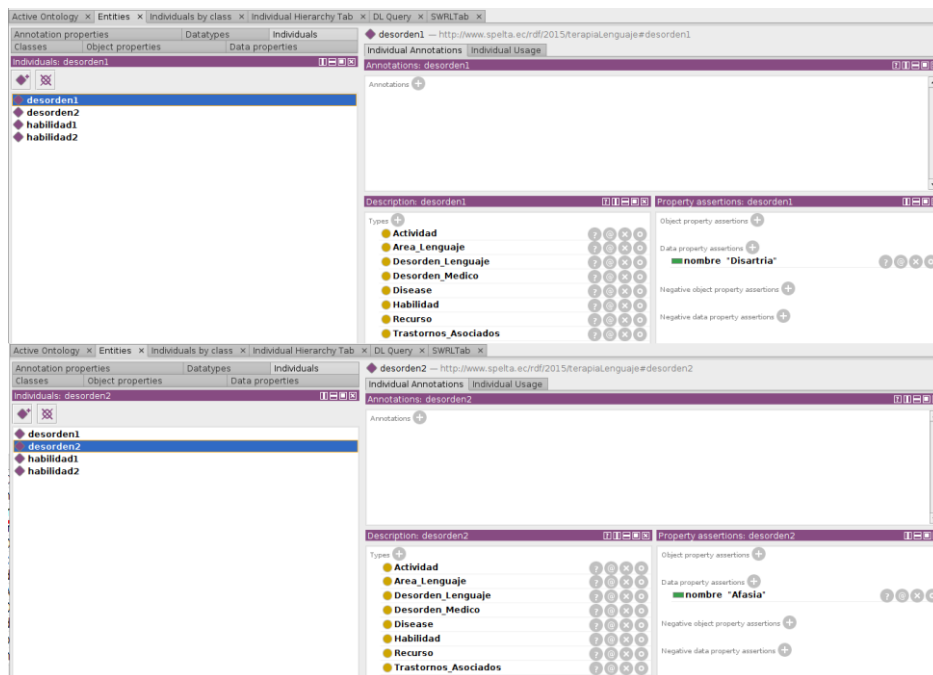


Figura 18 : Desórdenes.

En la Figura 19, se presenta dos habilidades declaradas ya que un desorden del lenguaje afecta a una habilidad de comunicación como se aprecia, en la etapa de pruebas.

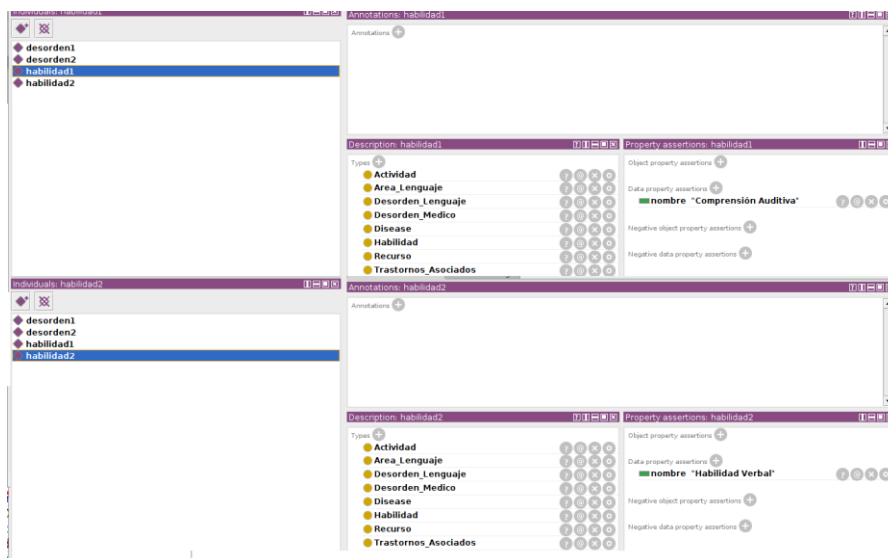


Figura 19 : Habilidades declaradas.

9.2. Recomendaciones generales

Al realizar la experimentación con **Hermit**, se pretendía hacer los razonadores de prueba en la ontología base, sin embargo, obtuvimos como resultado lo expuesto en la Figura 20; se revisó la estructura de la ontología y no se encontró ningún tipo de inconveniente, sin embargo, no se pudo realizar pruebas debido a una inconsistencia de tipos de datos. Esto ayuda a concluir que debemos tener cuidado al momento de manipular las ontologías por que podríamos estar afectando indirectamente o directamente su funcionamiento.

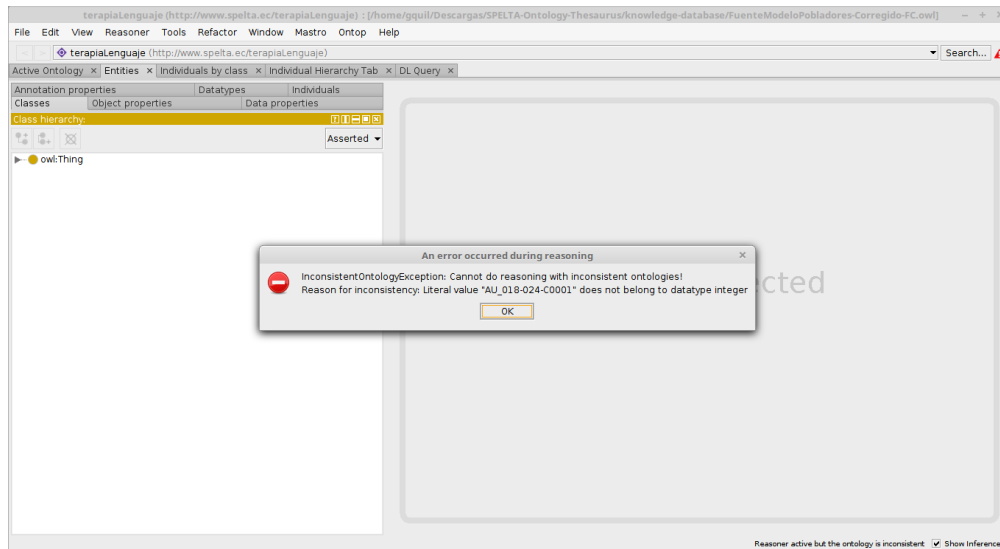


Figura 20 : Error de tipo de dato, en la ontología base.

10. Conclusiones

El estado del arte revela datos importantes. En donde los *crawlers* basan sus búsquedas en ontologías, utilizándolas para inferir información y tener mayor alcance. Sin embargo, aunque existen variedad de *crawlers* que utilizan diferentes modelos y métodos de búsqueda, en su mayoría dejan de lado la retroalimentación de las ontologías utilizadas, por esta razón el proyecto propuesto se encarga de su población y con ello cubrir nuevos conceptos y elementos de conocimiento.

Por esta razón se ha considerado que las ontologías tienen un papel importante ya que si se combinan con el análisis semántico se obtiene resultados más eficientes, mejorando significativamente la calidad de búsqueda de un *crawler*; uno de los inconvenientes encontrados al momento de realizar nuestro proceso de recolección de información, fue dado por falta de estándares los cuales limitan la abstracción del contenido semántico de una página web.

Los *crawlers* son orientados al análisis de contenidos documentales e hipertextuales, pero no consta con un análisis semántico en sus búsquedas ya que en la web actual encontramos todo tipos de estructuras hipertextuales sin estándares, el trabajo realizado ayudo a interpretar esta falencia, por esta razón uno de los problemas notorios fue obtener una data útil para el análisis y procesamiento semántico.

A partir del procesamiento de texto nos damos cuenta que, mediante análisis de lenguaje natural, en la tokenización se pueden eliminar palabras vacías o sin carga semántica para mejorar el rendimiento al momento de procesar similitud entre una o más oraciones, a esto agregamos que el análisis de lenguaje natural realizado mediante herramientas como *Freeling* que tiene un proceso de espera considerable, debido a que los análisis matemáticos y estadísticos utilizados revisan igualdad entre texto.

Se evidencia el potencial del módulo Owlready2 basado en Python que nos permitió manejar nuestra ontología desde una visión orientada a objetos, sin embargo, también se pudo concretar que se debe tener cuidado en la manipulación de la ontología; puesto podemos cambiar la estructura de los datos cambiando su naturaleza, cosa que afectaría el uso de la ontología predeterminado.

11. Trabajos a Futuro

Se desarrolla una herramienta de apoyo en base a un *crawler* semántico la arquitectura, funcionalidades expuestas en este trabajo acorde a las necesidades existentes.

Con los datos obtenidos el proyecto toma forma y nos permite un punto de partida en la relación de web semántica y ontologías, enfocándonos al futuro de este trabajo, cabe mencionar que se realiza un análisis semántico a profundidad con las búsquedas obtenidas por los *crawlers*.

A continuación, se describe hipótesis que pueden ser desarrolladas en trabajos futuros:

Los *crawlers* manejan consultas para una recolección de contenido hipertextuales, este proceso genera gran cantidad de tráfico en los servidores de búsquedas, estos servidores tienen restricciones una de ellas el tiempo entre búsqueda y búsqueda para no generar un bloqueo temporal o definitivo dentro del proceso, es importante tener un desarrollo o mecanismo que ayuda a evitar este inconveniente.

La solución dada por este trabajo es poblar una ontología con base a la búsqueda mediante un *crawler* semántico, a futuro se puede mejorar el proceso realizando mediante lotes, abarcando gran cantidad de datos en una sola ejecución, poblando varios segmentos de la ontología al mismo tiempo.

12. Referencias

- [1]. Sutton, A., & Samavi, R. (2017, October). Blockchain Enabled Privacy Audit Logs. In International Semantic Web Conference (pp. 645-660). Springer, Cham.
- [2]. Accesibilidad W3C.- Consulta: Febrero: Febrero-2018 [Online]. Available: <https://www.w3c.es/Divulgacion/GuiasBreves/Accesibilidad>
- [3]. Hassan Montero, Y., & Martín Fernández, F. J. (2003). Qué es la accesibilidad web. *No solo usabilidad*, (2).
- [4]. Kaya, S. A. K., Messaoudi, N., Bouhadida, M., & Bennour, A. (2016). Le Knowledge Management: Socle de Construction de Memoire de Projet. Réformes Economiques et Intégration en Economie Mondiale, (21), 187-207.
- [5]. Hedegaard, H. B., Johnson, R. L., & Ballesteros, M. F. (2017). Proposed ICD-10-CM Surveillance Case Definitions for Injury Hospitalizations and Emergency Department Visits. National health statistics reports, (100), 1.)
- [6]. Agarwal, P., Verma, R., & Mallik, A. (2016, August). Ontology based disease diagnosis system with probabilistic inference. In Information Processing (IICIP), 2016 1st India International Conference on (pp. 1-5). IEEE.
- [7]. Robles-Bykbaev, V., Arévalo-Fernández, C., Naranjo-Cabrera, E., Quito-Naula, P., Pauta-Pintado, J., Ávila, G., & Quezada, R. (2017, July). A Hybrid Approach Based on Multi-Sensory Stimulation Rooms, Robotic Assistants and Ontologies to Provide Support in the Intervention of Children with Autism. In *International Conference on Applied Human Factors and Ergonomics* (pp. 477-487). Springer, Cham.
- [8]. Galán-Mena, J., Ávila, G., Pauta-Pintado, J., Lima-Juma, D., Robles-Bykbaev, V., & Quisi-Peralta, D. (2016, June). An intelligent system based on ontologies and ICT tools to support the diagnosis and intervention of children with autism. In *Biennial Congress of Argentina (ARGENCON), 2016 IEEE* (pp. 1-5). IEEE.
- [9]. Robles Bykbaev, V. E., López Nores, M., Pazos Arias, J. J., García Duque, J. & Guillermo Anguisaca, J. (2015). A multilayer mobile ecosystem to support the assessment and treatment of patients with communication disorder, In 7th International Conference on E-Health (EH, co-located with MCCSIS). Gran Canaria, Spain.
- [10]. FOAF Vocabulary Specification 0.99.- Consulta: Febrero-2018 [Online]. Available: <http://xmlns.com/foaf/spec/>

- [11]. Khan, J. A., Sangroha, D., Ahmad, M., & Rahman, M. T. (2014, November). A performance evaluation of semantic based search engines and keyword based search engines. In *Medical Imaging, m-Health and Emerging Communication Systems (MedCom), 2014 International Conference on* (pp. 168-173). IEEE.
- [12]. Guo, G., & Wei, W. (2011, August). Video Semantic Information Architecture Based on Web Crawlers. In *Internet Technology and Applications (iTAP), 2011 International Conference on* (pp. 1-4). IEEE.
- [13]. Blázquez Ochando, Manuel. "Nuevos retos de la tecnología web crawler para la recuperación de información." *Métodos de información 4.7* (2014): 115-128.
- [14]. Rajarajeswari, S., & Chakraborty, S. (2017). Application of Formal Concept Analysis in Ontology Engineering. *Indian Journal of Public Health Research & Development*, 8(4), 1300-1306.
- [15]. Gaur, R., & Sharma, D. K. (2014, August). Review of ontology based focused crawling approaches. In *Soft Computing Techniques for Engineering and Technology (ICSCCTET), 2014 International Conference on* (pp. 1-4). IEEE.
- [16]. Zhao, Y., Wang, Z., Zou, L., Wang, J., & Hao, Y. (2014, May). A Linked Data Based Personal Service Data Collection and Semantics Unification Method. In *Service Sciences (ICSS), 2014 International Conference on* (pp. 118-123). IEEE. ISO 690
- [17]. Gamecho, B., Minón, R., Aizpurua, A., Cearreta, I., Arrue, M., Garay-Vitoria, N., & Abascal, J. (2015). Automatic generation of tailored accessible user interfaces for ubiquitous services. *IEEE Transactions on Human-Machine Systems*, 45(5), 612-623. ISO 690
- [18]. Robles-Bykbaev, V. E., Guamán-Murillo, W., Quisi-Peralta, D., López-Nores, M., Pazos-Arias, J. J., & García-Duque, J. (2016, October). An ontology-based expert system to generate therapy plans for children with disabilities and communication disorders. In *Ecuador Technical Chapters Meeting (ETCM), IEEE (Vol. 1, pp. 1-6)*. IEEE. ISO 690.
- [19]. Pino Toledano, David del. *Creación de un crawler semántico y distribuible para su aplicación en un buscador web*. MS thesis. 2014.
- [20]. Rodríguez García, Miguel Ángel. "Extracción semántica de información basada en evolución de ontologías." *Proyecto de investigación: (2014)*.
- [21]. Torres, D. M. B., Romero, A. C., & Sanabria, J. S. G. (2017). Web semántica, más de una década de su aparición. *Puente*, 8(1), 61-69.
- [22]. Sawant, V. V., & Ghorpade, V. R. (2014, December). Automatic semantic classification and categorization of web services in digital environment. In

- Computer and Communications Technologies (ICCCT), 2014 International Conference on (pp. 1-6). IEEE.
- [23]. Khan, J. A., & Kumar, S. (2014, October). Deep analysis for development of RDF, RDFS and OWL ontologies with protege. In Reliability, Infocom Technologies and Optimization (ICRITO)(Trends and Future Directions), 2014 3rd International Conference on (pp. 1-6). IEEE.
- [24]. Editores W3C, "OWL 2 Web Ontology Language Document Overview (Second Edition)", W3C, 2017, Available: <http://www.w3.org/TR/owl2-overview/>.
- [25]. Vásquez, A. C., Quispe, J. P., & Huayna, A. M. (2009). Procesamiento de lenguaje natural. *Revista de investigación de Sistemas e Informática*, 6(2), 45-54.
- [26]. Editores W3C, "Web Design and Applications – HTML & CSS", W3C, 2017, Available: <https://www.w3.org/standards/webdesign/htmlcss>
- [27]. Rodríguez, T., & Aguilar, J. (2014). Aprendizaje ontológico para el marco ontológico dinámico semántico. *DYNA*, 81, 56-63.
- [28]. Editores Django, "Meet Django", Django, 2017, Available: <https://www.djangoproject.com/>
- [29]. Wang, J., & Guo, Y. (2012, October). Scrapy-based crawling and user-behavior characteristics analysis on taobao. In Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC), 2012 International Conference on (pp. 44-52). IEEE.
- [30]. Editores Jean-Baptiste LAMY, "Owlready2 0.3 documentationS", Jean-Baptiste LAMY, Available: <http://pythonhosted.org/Owlready2/intro.html>
- [31]. Ankit Kumar, S. (2017, November). Multi-Lingual Sentiment Analysis of twitter data by using classification algorithms. In Electrical, Computer and Communication Technologies (ICECCT), 2017 Second International Conference on. IEEE
- [32]. Editores NLTK Project, "NLTK 3.2.5 documentation", NLTK Project, Available: <http://www.nltk.org/>
- [33]. PADRÓ, Lluís. Analizadores multilingües en freeling. *Linguamática*, 2011, vol. 3, no 1, p. 13-20.
- [34]. Lamy, J. B. (2017). Owlready: Ontology-oriented programming in Python with automatic classification and high level constructs for biomedical ontologies. *Artificial intelligence in medicine*, 80, 11-28.
- [35]. Arias, M. Á. (2017). Aprende Programación Web con PHP y MySQL: 2ª Edición. IT Campus Academy.