
# UNIVERSIDAD DE MURCIA

ESCUELA INTERNACIONAL DE DOCTORADO

Integroly: Sistema automático de población de un grafo de conocimiento a partir de datos sociales masivos en el dominio de marketing político

D. Héctor Hiram Guedea Noriega

2022


# UNIVERSIDAD DE MURCIA

ESCUELA INTERNACIONAL DE DOCTORADO

**Tesis Doctoral**

Integroly: Sistema automático de población de un grafo de conocimiento a partir de datos sociales masivos en el dominio de marketing político

D. Héctor Hiram Guedea Noriega

2022

Director

Dr. Francisco García Sánchez

## Agradecimientos

Dedicada a mi hija María Joaquina y a mi esposa Katy,
mi principal motor.

Profundo agradecimiento a mi director de tesis,
Dr. Francisco García Sánchez, por ser un tremendo
guía y apoyo durante todo este trayecto.

A mis padres y mi hermano,
por todos sus consejos.

# Índice de Contenido

# Índice de Tablas

# Índice de Figuras

# Lista de Acrónimos

| Término | Significado |
|---|---|
| API | Application Programming Interface |
| BI | Business Intelligence |
| CSS | Cascading Style Sheets |
| DA | Data Analysis |
| DBMS | Database Management System |
| DL | Description Logic |
| ETL | Extract, Transform, and Load |
| GDPR | Reglamento General de Protección de Datos |
| HTML | HyperText Markup Language |
| HTTP | Hypertext Transfer Protocol |
| IA | Inteligencia Artificial |
| IDE | Integrated Development Environment |
| KG | Knowledge Graph o Grafo de Conocimiento |
| KR | Knowledge Representation |
| ML | Machine Learning |
| MOP | Market-Oriented Party |
| MySQL | My Structured Query Language |
| NER | Named Entity Recognition |
| NLP | Natural Language Processing o Procesamiento de Lenguaje Natural |
| OE | Ontology Evolution/Enrichment |
| OL | Ontology Learning |
| OLAP | Online Analytical Processing |
| OP | Ontology Population |
| OWL | Web Ontology Language |
| PMont | Political Marketing Ontology |
| PoS | Parts of Speech |

| | |
|---|---|
| RDF | Resource Description Framework |
| SVG | Scalable Vector Graphics |
| SWRL | Semantic Web Rule Language |
| URI | Uniform Resource Identifier |
| W3C | World Wide Web Consortium |
| WebGL | Web Graphics Library |
| XML | eXtensible Markup Language |

# Resumen

Las campañas políticas modernas se sustentan en estrategias formuladas en base a lo que se ha dado en conocer como marketing político. El marketing político abarca desde la definición del producto político, a través de un minucioso análisis de las necesidades del electorado, hasta el desarrollo de campañas y la gestión de la comunicación política. La inteligencia de mercado, esto es, entender las demandas reales del mercado político, supondrá la primera etapa en este complejo proceso.

Uno de los principales problemas de la inteligencia de mercado político es el procesamiento de la diversidad de las fuentes de datos como son las encuestas, redes sociales, páginas Web, base de datos de instituciones, entre otras más. Las redes sociales se han convertido en una de las principales plataformas de conversación y canales para compartir experiencias y opiniones. Fomentan el discurso público y, en particular, cada vez más utilizadas para asuntos políticos, como es la participación ciudadana, el proselitismo o el debate político. Sin embargo, el monitoreo, la extracción, el procesamiento, el almacenamiento y el análisis de los datos masivos sociales en el ámbito político es una tarea complicada por su heterogeneidad propia de la estructura de la fuente de datos. A pesar de los retos y complicaciones, estas tareas son muy importantes para lograr encontrar las demandas a corto, medio y largo plazo de los potenciales votantes, y con ello, crear una estrategia y toma de decisiones ligada a las mismas.

Las tecnologías de la Web semántica, y en particular, las ontologías, permiten generar una base de conocimiento homogéneo, compartido y reutilizado, con un formato legible por la máquina y enriquecido para apoyar el descubrimiento, la integración, la representación y la gestión del conocimiento. En los últimos años, las ontologías y los grafos de conocimiento (KG) han demostrado ser eficaces para propiciar una solución de integración de datos y ayudando con la complejidad de encontrar relaciones significativas dentro de los datos heterogéneos.

Este trabajo se enfoca en dos principales contribuciones, la primera, generar un modelo ontológico basado en el dominio de marketing político, y la segunda, en una propuesta de sistema para la población automática de un grafo de conocimiento a partir de datos masivos sociales en el dominio del marketing político.

Nuestra propuesta de modelo ontológico se realizó basado en los criterios y necesidades del marketing político sobre las campañas políticas. La construcción de la ontología fue a través de metodología estándar *Ontology 101*, donde se definió el dominio y alcance de la ontología, consideró la reutilización de ontologías existentes, se enumeraron los términos importantes para la ontología, hasta la definición de clases, propiedades, relaciones e instancias base. La ontología se bautizó como PMont (por sus siglas en inglés como *Political Marketing Ontology*) la cual responde preguntas clave del marketing político como: ¿qué demanda el electorado?, ¿qué dice la opinión pública sobre el candidato (imagen, propuestas)?, ¿qué dice la opinión pública sobre el partido político al que pertenece el candidato?, ¿qué tipo de mensaje y lenguaje político debe establecer el candidato para impactar positivamente en la ciudadanía?, ¿qué tipo de propuestas de campaña debe diseñar?, ¿en qué medios el candidato tiene mayor impacto?, entre otras. La PMont permitió la integración de la información disponible sobre el electorado y candidatos a través de distintas fuentes de datos.

La segunda aportación de este trabajo de investigación es una solución automatizada basada en textos en castellano a través de técnicas de Machine Learning (ML) y Procesamiento del Lenguaje Natural (NLP, del inglés "*Natural Language Processing*"), recolección datos significativos de fuentes de medios digitales semiestructurados y no estructurados, procesamiento datos masivos y finalmente la población de un grafo de conocimiento previamente definido por un modelo ontológico del dominio de marketing político. La propuesta de sistema de automatización cuenta con las siguientes fases o componentes: (i) conectividad a fuentes y almacenamiento de datos, (ii) procedimiento de extracción de información, (iii) proceso de población y validación de la ontología, y finalmente, (iv) el grafo de conocimiento resultante.

Como resultado de esta tesis doctoral se tiene un grafo de conocimiento poblado con información validada, precisa, consistente y confiable. El grafo de conocimiento fue evaluado a través de 18 requisitos de calidad sobre las dimensiones de accesibilidad, contextualidad, intrínseca y representatividad, de los cuales nos brinda una base de conocimiento óptima para la inteligencia de mercado del marketing político.

# Abstract

Modern political campaigns are based on strategies formulated in what has become known as political marketing. Political marketing ranges from the definition of the political product, through a detailed analysis of the needs of the electorate, to the development of campaigns and the management of political communication. Market intelligence is understanding the real demands of the political market, will be the first stage on this complex process.

One of the main problems of political market intelligence is the processing of the diversity of data sources such as surveys, social networking sites, Web pages, databases of political institutions, among others. Social networks sites have become one of the main conversation platforms and channels for sharing experiences and opinions. They encourage public discourse and, in particular, are increasingly used for political issues, such as citizen participation, proselytism or political debate. However, the monitoring, extraction, processing, storage and analysis of social big data in the political sphere is a complicated task due to the heterogeneity of the data source structure. Despite the challenges and complications, these tasks are very important to meet the short, medium and long-term demands of potential voters, and thereby create a strategy and decision-making linked to them.

Semantic Web technologies, and in particular, ontologies, allow the generation of a homogeneous knowledge base, shared and reused, with a machine-readable format and enriched to support the discovery, integration, representation and management of knowledge. In recent years, ontologies and Knowledge Graphs (KGs) have proven to be effective in enabling a data integration solution and helping with the complexity of finding meaningful relationships within heterogeneous data.

This work focuses on two main contributions, the first, to generate an ontological model based on the domain of political marketing, and the second, in a framework proposal for

automatic Knowledge Graph population from social big data in the political marketing domain.

Our ontological model proposal was made based on the criteria and needs of political marketing on political campaigns. The construction of the ontology was through the standard *Ontology 101* methodology, where the domain and scope of the ontology were defined, the reuse of existing ontologies was considered, the important terms for the ontology were listed, up to the definition of classes, properties, relationships and base instances. The ontology was named PMont (for its acronym in English as Political Marketing Ontology) which answers key questions of political marketing such as: what does the electorate demand?, what does public opinion say about the candidate (image, proposals)? What does public opinion say about the political party to which the candidate belongs? What kind of message and political language should the candidate establish in order to have a positive impact on the citizens? What kind of campaign proposals should he design? What media does the candidate have the greatest impact?, among others. The PMont allowed the integration of the information available on the electorate and candidates through different data sources.

The second contribution of this research work is an automated system based on texts in Spanish through Machine Learning (ML) and Natural Language Processing (NLP) techniques, collecting significant data from semi-structured and unstructured digital media sources, processing big data and finally the population of a Knowledge Graph previously defined by an ontological model of the political marketing domain. The automation system proposal has the following phases or components: (i) source connection and data collection, (ii) information extraction, (iii) ontology population and validation process, and finally, (iv) the Knowledge Graph. While the process is carried out automatically, the instances added to the Knowledge Graph are annotated so that human experts can supervise the results of the automated population in an optional later stage.

As a result of this doctoral thesis, we have a Knowledge Graph populated with validated, precise, consistent, and reliable information. The Knowledge Graph was evaluated through 18 quality requirements on the dimensions of accessibility, contextuality, intrinsic and

representativeness, of which it provides us with an optimal knowledge base for political marketing market intelligence.

# Capítulo 1. Introducción

El objetivo general del marketing político es maximizar la victoria electoral con una serie de técnicas de investigación, planificación, gestión y comunicación, diseñadas y ejecutadas para la manipulación y persuasión del votante a través de medios de difusión masiva, ya sean tradicionales (radio, televisión, volanteo) y no tradicionales (o medios digitales, como Internet, aplicaciones móviles, redes sociales y más) (Costa, 1994). Tradicionalmente, el marketing político realiza un estudio detallado del electorado con el objetivo de conocer la demanda política. Para después, formular la estrategia política que permee en el posible votante de manera positiva a través de un plan de acción en medios de difusión.

Desde finales del Siglo XX, al aumentar la demanda de partidos y candidatos, se convirtió de suma importancia enfatizar en el interés social, ya que los actuales ciudadanos piden honestidad, ideas propias, confianza, propuestas creativas en base a la renovación política. Previo y durante las campañas políticas, los responsables de crear las estrategias electorales establecen un sistema de obtención y análisis de información del electorado. Conocer sus gustos, intereses y qué es lo que desea, para con ello establecer una comunicación que impacte y forme seguidores. Esto se vuelve casi imposible al ser miles o millones de posibles votantes. Donde las encuestas de opinión, una herramienta básica del marketing político, se vuelven una técnica obsoleta. Por ello es necesaria la tecnología y los medios masivos, en especial aquellas herramientas omnipresentes y flexibles que apoyen la extracción y almacenamiento de información precisa y objetiva (Antoniades, 2021).

En la actualidad, las campañas políticas modernas requieren cada vez más el acceso y gestión de datos masivos. Se sabe que los datos masivos brindan la posibilidad de predecir el futuro a través de procesos de análisis de datos, ofreciendo una ventaja competitiva y atribuyendo gran parte de su éxito a la rapidez y fiabilidad de procesar la información transformándola en conocimiento electoral. El principal problema de las campañas electorales es la descentralización de la información, pero a su vez, es un punto importante en el nuevo escenario político. Ya que, en el pasado, los medios tradicionales eran los que influenciaban principalmente al electorado. La prensa, centralizada, producía la misma

información del candidato para repercutir rápidamente en la opinión pública. Hoy en día, Internet, y sus cientos de nuevas tecnologías colaborativas e informativas, donde el usuario es el creador de la información y líder de opinión, se genera un ambiente de inteligencia colectiva, pero difícilmente se puede obtener, clasificar, ordenar y almacenar en dominios específicos para ser utilizados con fines electorales (Costa, 1994). Este nuevo ecosistema electoral, de los medios libres y la heterogeneidad de información, brinda la necesidad desarrollar de sistemas que identifiquen tendencias y ofrezcan decisiones automáticas gracias al almacenamiento homogéneo, semántico y ordenado de la información.

Las tecnologías semánticas proveen de características sobresalientes para enfrentar los retos actuales del marketing político dentro del procesamiento de la información. Una de ellas es la implementación y construcción de ontologías de dominio con estructuras de datos estandarizadas y homogéneas disponibles en formatos OWL y RDF, y a su vez, con la posibilidad de vincular una gran variedad de fuentes de datos heterogéneos (no estructurados, semi-estructurados y estructurados) para la población del modelo ontológico, y como resultado, un grafo de conocimiento leíble por humanos y sistemas informáticos, reusable y distribuido (van Atteveldt et al., 2007).

El enfoque de este trabajo es lograr ofrecer una novedosa herramienta y técnica para el marketing político aplicado a un candidato, con el consumo y gestión de datos masivos sociales del electorado, un modelo ontológico con clases, propiedades y relaciones semánticas que permitan la población de nuevas tendencias, y finalmente, la automatización de los procesos para disminuir el error humano y de tiempo invertido (Anduiza et al., 2012).

Lo expuesto anteriormente ha sido una de las principales motivaciones para el desarrollo de este trabajo de investigación. Esta tesis doctoral tiene como objetivo principal el desarrollar un sistema automático para la población de un grafo de conocimiento del dominio de marketing político, el cual permitirá homogenizar, reutilizar y compartir información producida por el electorado proveniente de fuentes de datos no estructurados, semi-estructurados y estructurados desde los medios digitales.

Para cumplir con objetivo principal de esta tesis doctoral, se ha compuesto este documento en seis capítulos principales, descritos a continuación:

El Capítulo 2 está destinado al estado del arte, describe los conceptos y definiciones del marketing político, haciendo un breve recorrido histórico hasta llegar a los actuales retos que se presentan en las campañas políticas modernas. A su vez, se realiza una profunda investigación sobre las tecnologías ligadas a la Web semántica, en específico haciendo hincapié en las ontologías y grafos de conocimiento, sus usos y aplicaciones, beneficios en cara a los sistemas tradicionales de almacenamiento de la información.

El Capítulo 3 está dedicado a la propuesta de ontología de dominio para el marketing político bautizada como PMont (*Political Marketing Ontology*), se describe el desarrollo del modelo ontológico, la elección de conceptos, propiedades y relaciones basados en responder las preguntas de investigación que beneficien las campañas políticas por medio del conocimiento del electorado o la inteligencia de mercados. Por último, la validación de la propuesta a través de una herramienta encargada de detectar malas prácticas de diseño que podrían, potencialmente, provocar errores en el modelado del conocimiento.

El Capítulo 4 describe la propuesta de sistema automático de la población del grafo de conocimiento del marketing político. Se presenta la arquitectura funcional del sistema basada en cinco componentes principales: (i) el módulo de conexión de fuente y recopilación de datos, (ii) extracción de información, (iii) ontología PMont, (iv) población ontológica y proceso de validación, y finalmente, (v) el grafo de conocimiento resultante. Se profundiza en cada uno de los componentes de la arquitectura funcional, por ejemplo, desde las técnicas de extracción de datos, lenguaje de programación, técnicas de NLP y ML, y más.

El Capítulo 5 está destinado a la validación del grafo de conocimiento resultante, se establecen los requerimientos de evaluación, partiendo de la primicia de la población del modelo bajo un entorno simulado, y posteriormente, se analizan sus resultados.

El Capítulo 6 describe las principales características del enfoque de nuestra propuesta, las conclusiones de esta tesis doctoral, las limitaciones y posibles nuevas vías de trabajo futuro.

Finalmente, Capítulo 7 presenta las contribuciones científicas derivadas de esta tesis doctoral, basadas en publicaciones en revistas y congresos.

# Capítulo 2. Estado del Arte

El presente capítulo provee describir los conceptos y definiciones del marketing político, así como también el estado actual de tecnologías involucradas en esta tesis doctoral.

El siguiente apartado 2.1 presenta la definición y un breve resumen histórico del marketing político, para posteriormente describir sus fundamentos y el estado actual de las técnicas aplicadas sobre los procesos electorales, referido como campañas políticas.

El apartado 2.2 describe el estado actual de las tecnologías que forman parte del núcleo central de este trabajo de investigación que son: (i) web semántica, (ii) datos enlazados y (iii) ontologías y grafos de conocimiento.

El apartado 2.3 hace un recorrido por el estado actual del análisis de datos masivos, incluyendo las aplicaciones técnicas y tecnologías como ML y NLP sobre datos sociales y/o semánticos.

Por último, el apartado 2.4 presenta los objetivos, motivaciones y metodologías de esta tesis doctoral.

## 2.1 Marketing Político

Esta sección contiene información focalizada en definiciones de los conceptos del marketing político, haciendo un breve recorrido por sus antecedentes hasta el estado actual de las técnicas empleadas en campañas electorales.

### 2.1.1 Definiciones y antecedentes

Un de las definiciones del marketing político se describe en (Alonso Coto & Adell, 2011) como “la disciplina orientada a la creación y desarrollo de conceptos políticos relacionados con partidos o candidatos específicos para satisfacer a ciertos grupos de electores a cambio de sus votos”. Esta definición incluye tres conceptos fundamentales (i) producto político,

(ii) la identificación y segmentación del mercado de votantes, y (iii) la inteligencia de marketing aplicada a la política (Juárez, 2003).

El producto político (i) se refiere a las ideas a transmitir por parte de los políticos y partidos, el cual debe definirse a partir de la identificación de las necesidades del electorado. Por otro lado, la segmentación del electorado (ii) consiste en dividir la masa electoral heterogénea en secciones más pequeñas que tienen algo en común con el objetivo de detectar grupos suficientemente grandes para los que el producto político puede resultar especialmente atractivo. Para ello se pueden utilizar diferentes técnicas como la geográfica (es decir, el lugar donde vive la gente según regiones y zonas dentro de esas regiones), conductual (es decir, basada en las acciones del individuo), demográfica (es decir, edad, tipo de familia, clase social, ingresos, etc.) o psicográficos (es decir, características de estilo de vida, valores comunes, creencias, actitudes, actividades, intereses y opiniones). Finalmente, la inteligencia de marketing (iii) permite comprender qué quiere el mercado político, es decir, el electorado, de las élites políticas, es decir, los partidos políticos y los candidatos, utilizando técnicas de investigación cuantitativa y cualitativa. El objetivo último de la inteligencia de marketing es situar el producto político en un nicho ideológico inaccesible para los competidores por su ventaja competitiva, que es capaz de atraer el número de votos suficiente para alcanzar el objetivo electoral deseado.

Por lo tanto, el marketing político surge para comprender las nuevas formas de hacer política moderna desde un enfoque transversal que engloba el marketing, el análisis del comportamiento de los partidos políticos y los votantes, la comunicación y el uso efectivo de los diferentes medios de comunicación (tradicionales y digitales) como vías de promoción (Scammell, 1999).

Las campañas electorales se realizan sobre la base de procesos evolutivos con el propósito de construir una estrategia política rentable para el electorado. Los votantes (iii) pueden definirse como consumidores de activos políticos (i); por tanto, un candidato o partido político (ii) tiene la necesidad de satisfacer dicha demanda mediante un estudio detallado del electorado (iii) y comunicar correctamente la oferta política (ii) (Maarek, 2011).

La historia del marketing político tiene sus orígenes con el término "propaganda política" a mediados del siglo XX. Proviene del latín "propagare" que significa "difundir, extender, ser divulgado", donde el objetivo principal era la difusión masiva de ideas arbitrarias que buscan influir en los sistemas de valores y comportamientos de los ciudadanos. La evolución del proceso de influir en el comportamiento de los ciudadanos con la democratización de los procedimientos políticos dio lugar al marketing político como una rama nueva e independiente de la ciencia política cuya actividad no es solo difundir un mensaje, sino también estudiar la psicología de las masas, ética institucional, estadísticas, medios de comunicación, entre otros (Moore, 2010).

El marketing político nace oficialmente en la década de los 50's cuando por primera vez un candidato a la presidencia de los Estados Unidos, Dwight Eisenhower, contrató la agencia de publicidad BBDO (Barton, Durstine, Osborn & Batten Co.) para trabajar en su imagen y encargarse de la mercadotecnia mediática durante la campaña electoral. Esto supuso, por tanto, la incorporación de técnicas de investigación de mercados y publicitarias a la comunicación política. El objetivo principal fue convertir al candidato en un agente político de transformación social (producto político) a través de una sólida estrategia discursiva a través de los medios de comunicación, en ese momento radio y televisión. Dwight Eisenhower pasó de ser el militar que luchó en la Segunda Guerra Mundial a ser visto por los votantes como un administrador eficiente, convirtiéndose en el 34º presidente de los Estados Unidos (1953-1961). Con lo que se manifestaron los aspectos fundamentales que engloban la función del marketing político descritos como: producto político, segmentación del mercado y inteligencia de marketing aplicada a política (Alonso Coto & Adell, 2011). En específico, los conceptos claves son: política, comunicación, mercadotecnia e investigación de mercados.

La comunicación por sí misma se define, según RAE, como la "*transmisión de señales mediante un código común al emisor y al receptor*", y política, como el "*arte o traza con que se conduce un asunto o se emplean los medios para alcanzar un fin determinado*". Ambos conceptos tienen una similitud: el transmitir un mensaje. Entendiendo, en este contexto, que se trata de transmitir un mensaje desde el candidato a la ciudadanía, los gobiernos a la ciudadanía y, por último, los partidos a la ciudadanía. Hablamos en realidad

de un intercambio bidireccional de información, que fomente la participación ciudadana. Por lo tanto, la comunicación política se entiende como la responsable de transmitir ideas, información y actitudes en el ámbito público.

La mercadotecnia, según Philip Kotler (considerado el padre de la mercadotecnia moderna), busca satisfacer las necesidades de un grupo de individuos a través de un intercambio de bienes o servicios, con el apoyo de procesos sociales y administrativos (Kotler & Armstrong, 2013)

Debido a la necesidad de información para el desarrollo de una estrategia de marketing, es fundamental realizar investigación de mercados, por medio de la cual se indagará sobre las necesidades, gustos y preferencias del público objetivo. La American Marketing Association (AMA) define a la investigación de mercados como la "*función que enlaza al consumidor, al cliente y al público con el comercializador a través de información*". En resumen, es una guía donde, a través de una serie de pasos, realiza recolección de datos para mejorar la toma de decisiones (Fernandez, 2004).

El marketing político, entonces, establece los objetivos y programas para satisfacer la demanda social, influenciando el comportamiento de los ciudadanos mediante un discurso político atrayente durante el periodo electoral. A su vez, busca desarrollar una estrategia efectiva que optimice recursos e ideas, con el fin de ser coherente a las expectativas del electorado y el candidato, definiendo un horizonte político certero (Gómez, 2015).

Algunos autores exponen la definición de marketing político con ideas generales, pero con vínculos en común. Por ejemplo, en (Gómez, 2015) se destacan las siguientes definiciones:

- "El proceso mediante el cual los candidatos políticos y las ideas son dirigidas a los votantes en orden de satisfacer sus necesidades políticas y ganar su apoyo para apoyar al candidato y sus ideas", Avraham Shama, profesora de marketing de la Universidad de Baruch, Nueva York.
- "El marketing político es el conjunto de técnicas comunicativas que dispone un partido o un político para intentar modificar la opinión y comportamiento de los

> electores para ser elegido y obtener los máximos votos posibles, es decir, acuerdo entre los gobernantes y los gobernados y cambio de información entre estos, a través de canales de información", Jean-Marie Cotteret, politólogo francés.

En conclusión, el marketing político es un conjunto de técnicas y herramientas utilizadas para transmitir ideas e influenciar a los electores a favor de un candidato.

### 2.1.2 Modelo de marketing político

El modelo de marketing político brinda el conocimiento genérico de los conceptos claves para trazar las estrategias orientadas en dos tipos de enfoques: candidatos y partidos. Representa los lineamientos específicos para desarrollar la teoría política a la práctica, definiendo e identificando elementos que fortalecerán el plan de marketing y permitirán su aplicación durante la campaña electoral.

Bruce I. Newman, consejero de Bill Clinton y autor de "Marketing para el presidente", formuló un modelo genérico basado en electores, el cual fue aplicado con éxito durante la campaña presidencial de Clinton entre 1992 y 1993, donde obtuvo el triunfo. Éste consta de los siguientes puntos (Alonso Coto & Adell, 2011):

- El foco en el candidato: definir el concepto de partido, de producto político y de venta a los votantes.
- La campaña del marketing:
    - Segmentación del mercado de votantes para: identificar necesidades, crear perfiles de votantes y decidir segmentos de potenciales votantes.
    - Posicionamiento del candidato: identificar las fortalezas y debilidades del candidato, buscar diferencias con la competencia, decidir la imagen a mostrar para cada segmento de votantes de interés, decidir la imagen global a mostrar integrando lo común de las anteriores.
    - Formulación e implantación de la estrategia: definición de la campaña y su plataforma (investigación de mercados, producto político, *marketing pull*

destinado a medios masivos, *marketing push* destinado a las bases), desarrollo y control de la organización.

- Fuerzas externas: tecnología (internet, TV, etc.), cambios estructurales (reglas y convenciones, regulaciones financieras, debates, etc.), influencia de los distintos actores (candidatos, consultores, encuestadores, medios, otros partidos, grupos de interés, votantes, etc.)
- Campaña política: etapa preliminar, etapa primaria, etapa de convención y etapa de elección.

El modelo de Newman define una campaña basada en las necesidades del mercado de electores, centrándose en las preocupaciones y deseos, y no en los puntos de vista personales (utilizado sólo como arranque, sin embargo, no como objetivo final), permitiendo maximizar la eficiencia de la comunicación política en la inteligencia de marketing.

Newman diseña un modelo genérico de marketing político, pero si se desea detallar por partido o candidato, se realizan dos modelos diferenciados por su estructura, a saber, el modelo de marketing político aplicado a partidos (MOP por sus siglas en inglés '*Market Oriented Party*') y el modelo de marketing político aplicado a candidatos. El MOP, definido por Lees-Marshment, está pensado para partidos que desean establecer una orientación de mercado global. Esto es, busca la satisfacción de los votantes encontrando las demandas principales para diseñar un producto que satisfaga estas necesidades globales. A su vez, implementa cambios que soporten el producto por parte de la estructura del partido, sin tratar de adoctrinar a los votantes por una ideología absoluta. Por otro lado, el modelo de marketing político aplicado a candidatos, definido por Iordanis Kotzaivazoglou, establece los mismos puntos que el MOP para partidos, con la gran diferencia de que éste, por su falta de estructura, no se centra en un mercado global, sino establece segmentos o *target groups* para hacer uso más eficiente de los recursos y maximizar la satisfacción de los votantes.

Estos modelos se utilizarán para crear un plan de marketing político, el cual se describe en el apartado 2.1.3.

### 2.1.3 Plan de marketing político

El plan de marketing político se refiere a la serie de etapas o fases continuas y subsecuentes el cual construye la estrategia de una campaña política. Es la implementar de la teoría política a la práctica de campaña. Está constituido por las siguientes etapas (Zamudio, 2015):

- Análisis: describe el contexto y entorno social, económico y político, además de la situación y posición de la opinión pública. De estos componentes de análisis se desglosan los siguientes tipos con el objetivo de obtener la información necesaria para implementar la estrategia:

    - Factores demográficos
    - Factores económicos
    - Análisis del mercado político
    - Diagnóstico de los factores de marketing.

- Previsión: resultado del análisis, los dirigentes deben subrayar la tendencia, tanto de forma global del mercado como desde el punto específico sobre el partido. A esto se debe añadir el análisis de las circunstancias aleatorias o imprevistas que inciden en el candidato.
- Objetivos: después de analizar el mercado político, obtener los aspectos favorables y negativos, puntos fuertes y débiles, se debe decidir hacía dónde se quiere ir y qué es lo que se desea conseguir. Estos objetivos deben ser concretos y realistas, cuantificables, alcanzables, desarrollados en plan de acción y, por último, controlables. Algunos ejemplos de objetivos dentro del plan de marketing político son:

    - Investigación de mercados políticos.

- Producto político: candidato, partido y programas.
- Publicidad política.
- Captación de votos.

- Estrategias: una estrategia describe el camino que se va a trazar para conseguir un objetivo. Es la forma de utilizar los recursos de manera correcta para desarrollar una serie de acciones planificadas, principalmente para no cometer errores comunes durante el trayecto. Las estrategias se elaboran con base a dos etapas principales:

  - Elaboración de la estrategia con las ventajas e inconvenientes, valorando posibles riesgos y el grado de posibilidad para conseguir los objetivos.
  - Elección definitiva de una de las alternativas como consecuencia de la discusión y evaluación.

- Planes de acción: se determinan las tácticas a realizar con acciones específicas; esto es sinónimo de programa de marketing político. Un plan de acción debe dar respuesta a las siguientes cuestiones: ¿qué acciones concretas se deberán realizar?, ¿quién o quiénes las realizarán?, ¿en qué momento iniciarán y finalizarán?, ¿qué recursos económicos y humanos deberán asignarse? En resumen, los planes de acción deben reunir los siguientes elementos para su desarrollo:

  - Acciones a realizar.
  - Calendario de actividades.
  - Resultados a obtener.
  - Responsables de las acciones.

- Ejecución y control: una vez definido todo lo anterior, se pondrá en ejecución el plan de marketing. Durante su ejecución se vigilará cada proceso por si existe una desviación producida a lo largo del desarrollo, y poder solucionarla y corregirla de forma inmediata, si es necesario.

### 2.1.4 Aplicación y retos del marketing político en campañas electorales

El marketing político tiene su principal aplicación durante las campañas electorales. Las campañas electorales, también denominadas campañas políticas, son procesos de proselitismo y atracción de electores realizado por los partidos políticos y sus candidatos con el objetivo final de obtener votos del ciudadano. Estas campañas se pueden llevar a cabo a través de múltiples medios o formas de proselitismo, desde la sociedad civil, asociaciones de ciudadanos, organizaciones no gubernamentales, partidos políticos y, por supuesto, medios de comunicación. Las campañas electorales son instrumentos que legitiman los procesos para ascenso al poder en sistemas políticos democráticos (Valdez Zepeda, 2010).

En los últimos tiempos el comportamiento político, desde la acción de acudir a votar hasta la participación ciudadana en la política, se ha transformado considerablemente. La ciudadanía ha cambiado su actitud hacía la elite política. Son cada vez menos predecibles. Lo medios de comunicación se han descentralizado y vuelto más libres. La penetración y auge de las telecomunicaciones y las nuevas tecnologías entre los ciudadanos han sacudido la arena política dando lugar a nuevos retos. Estos retos incluyen transformaciones y adhieren nuevos conocimientos para favorecerse de los sistemas políticos democráticos y sociedades más digitalizadas e informadas (Alonso Coto & Adell, 2011). Algunos de ellos se describen a continuación:

- Disposición de más información: los candidatos disponen de mayor cantidad y calidad de información sobre los electores, la contienda y los adversarios. Del mismo modo, los votantes tienen más información sobre los candidatos y partidos. La información es descentralizada, a través de distintos medios.

- Uso de nuevas tecnologías: los procesos políticos se han reproducido cada vez más en escenarios tecnológicos. Desde el uso de Internet, base de datos, dispositivos móviles y más.

- Nuevas formas de hacer proselitismo: la llegada de las nuevas tecnologías propició nuevas formas de desarrollar proselitismo; ahora se conoce con mayor precisión el perfil del votante, sus gustos, sus patrones de conducta y, por tanto, es posible hacer una segmentación específica del electorado. Toda esta información se transforma en conocimiento para el candidato y los partidos políticos, siendo de gran ayuda como una ventaja competitiva.

- Predicción de resultados: la medición de la intención del voto se realiza a través de distintos estudios cuantitativos y cualitativos tales como las encuestas, herramienta básica para conocer un resultado posible de la campaña. Aunque éstas tienen un margen de error que podría llegar a ser un instrumento meramente intuitivo.

- Predominio de campañas mediáticas: los medios de comunicación se convirtieron en el espacio por excelencia para la interacción social y la difusión. Desde el invento de la radio, la televisión y ahora Internet, los agentes políticos los han utilizado como instrumentos para transmitir y cautivar al electorado.

- Existencia de un nuevo elector: por las características de la sociedad de la información y el conocimiento, el nuevo electorado se volvió más crítico, más selecto, ya que goza de mayor conocimiento y educación.

- Articulación de nuevas estrategias: el nuevo escenario del electorado exige campañas más sofisticadas y creativas. Por tal motivo la interactividad y la segmentación de los mercados electorales se vuelve imprescindible.

- Nuevos partidos y nuevos candidatos: la forma de generar política de los candidatos y partidos ha cambiado. El electorado ya no desea escuchar lo rimbombante y la verborrea de siempre. Desea un político con lenguaje cotidiano, igualitario, de comunicación bidireccional.

Teniendo en cuenta las anteriores afirmaciones, es imprescindible que durante las campañas electorales los candidatos lleven a cabo un conjunto de actividades y acciones con el objetivo principal de transmitir su mensaje político y convencer al electorado que es la opción más viable de acuerdo con sus necesidades como ciudadano.

La evolución de las comunicaciones con la llegada de Internet, la Web y posteriormente múltiples herramientas y aplicaciones digitales como las redes sociales (Twitter, Facebook, y más), han sido pilares para la segunda transformación del marketing político moderno del siglo XXI, definiendo nuevas estrategias de propaganda, imagen, discurso de los líderes políticos. Propiciando un escenario interactivo y social para realizar acciones de proselitismo por parte del candidato o partido político. A su vez, se ha transformado la imagen pública del candidato, su lenguaje y formas de interactuar con el ciudadano. El ciudadano también ha sido agente de cambios. Con el acceso a la información, potencializa su debate y su poder de decisión. Las élites de poder ascienden a través de campañas cada vez más personalizadas y creativas. Un ejemplo de éxito, fue la victoria del expresidente Barack Obama se atribuye a la estrategia de activismo digital en las redes sociales implementada como canal para la propagación del mensaje electoral y la participación efectiva del electorado (Harfoush, 2010).

La constante evolución del marketing político está ligada al desarrollo de valores humanos temporales como la libertad civil y la democracia, pasando de una representación acrítica a una activamente crítica, donde el escrutinio público es cada vez más minucioso y las redes sociales han fomentado la participación activa. participación ciudadana durante los procesos electorales (Ganduri et al., 2020). El marketing político ha demostrado ser una ciencia con un conjunto eficiente de prácticas, elementos y herramientas para la influencia política de los candidatos (Zuiderveen Borgesius et al., 2018). Sin embargo, las técnicas de comunicación y conocimiento del electorado se han vuelto más complejas debido, en parte, a la aparición de Internet y la explosión de los sitios o servicios de redes sociales (SNS) (Alonso Coto & Adell, 2011) (Jain et al., 2015). Asimismo, conocer las necesidades de los votantes a través de la información de distintas fuentes se ha convertido en una tarea ardua pero importante por ser uno de los principales objetivos del marketing político (Antoniades, 2021). Las necesidades de nuevas estrategias de extracción, clasificación y almacenamiento

de información que permitan el análisis de datos masivos son tareas fundamentales, y uno de los principales desafíos a resolver cuando se trata de información proveniente de diversas fuentes es el de la heterogeneidad de los datos.

Las tecnologías de la Web semántica han demostrado su utilidad en la gestión, almacenamiento y análisis de datos a nivel de conocimiento (Hoppe et al., 2018). En particular, las ontologías brindan una comprensión compartida del conocimiento sobre un dominio específico (Pinto et al., 2009) y pueden facilitar la integración de datos de fuentes heterogéneas (Guedea-Noriega & García-Sánchez, 2018), mientras que los grafos de conocimiento (KGs) enfatizan en la comprensión contextual, por ser conjuntos de hechos interrelacionados que describen entidades, eventos o cosas del mundo real y sus interrelaciones en un formato comprensible para humanos y máquinas (Barrasa et al., 2021).

En los siguientes apartados de este capítulo se profundiza sobre las tecnologías mencionadas.

## 2.2 Web Semántica y grafo de conocimiento

En esta sección se describen las tecnologías centrales para el desarrollo de esta la tesis doctoral como lo es: Web semántica (2.2.1), ontologías y grafos de conocimiento (2.2.2), y finalmente, datos enlazados (en inglés '*Linked Data*') (2.2.3).

### 2.2.1 De la Web a la Web Semántica

Desde su nacimiento en 1989 por su creador Tim Berners-Lee, la Web ha crecido de manera explosiva, convirtiéndose en la mayor base de datos del mundo. Las cifras actuales de más de cinco billones de usuarios y casi de ocho billones de páginas Web existentes (Stats, 2021), nos marcan un referente de las grandes dimensiones de información almacenada en Internet.

El mayor crecimiento de Internet se basó gracias al desarrollo y uso de nuevas herramientas para la producción de contenido y la comunicación, favoreciendo a partir del 2004 al auge

de la segunda generación de la Web denominada como Web 2.0 o Web social, donde los usuarios comunes, sin conocimientos de programación, tuvieron la facilidad de generar y publicar información con mayor facilidad en la red (Nafría, 2008).

Sin embargo, esta gran base de datos llamada Internet es entendible solamente por los humanos, por lo que su organización general se vuelve casi imposible, dificultando al propio usuario encontrar la información que necesita y personalizarla. Por ello, surge un nuevo conjunto de reglas para enriquecer a la Web, posibilitando a las máquinas el entendimiento y la automatización de determinadas acciones, definida como Web Semántica (Berners-Lee et al., 2001).

La Web Semántica reúne iniciativas destinadas a promover una futura Web cuyas páginas estén debidamente organizadas, estructuradas y codificadas de tal manera que los ordenadores sean capaces de efectuar inferencias y razonar a partir de sus contenidos. Con este fin añade metadatos específicos, semánticos y ontológicos, evolucionando del lenguaje etiquetado simple utilizado hasta este momento como es el caso del HTML. Con el objetivo de describir el contenido, su significado y la relación de los datos en las páginas Web, se emplean documentos RDF (por sus siglas en inglés '*Resource Description Framework*') (W3C, 2014), en donde se específica el sujeto, verbo y objetos para describir el recurso y afirmar propiedades como "es parte de", "es autor de", con valores de personas, enlaces, etcétera.

La adición de estos metadatos enriquecen a Internet y permiten ampliar la interoperabilidad entre los sistemas informáticos, posibilitando el uso de agentes inteligentes, los cuales buscan información sin operadores humanos, esto es, automatizan determinadas tareas en el acceso al contenido publicado en la Web (García-Sánchez & Guedea-Noriega, 2017).

#### *2.2.1.1 Fundamentos de la Web Semántica*

El término Web Semántica vuelve a ser atribuido al creador de la Web, Tim Berners-Lee, en el 2001 a través de su documento titulado "The Semantic Web" (Berners-Lee et al., 2001). En él, el autor define algunos puntos importantes en el imaginario de la Web futura, hiperconectada y organizada.

La visión de Tim Berners-Lee es de disponer de datos en la Web definidos y enlazados de tal forma que puedan ser utilizados por las máquinas y no solamente para mostrar contenido sino también para automatizar tareas, integrar y reutilizar datos entre aplicaciones.

Por tal motivo, la esencia principal de la Web Semántica es codificar con nuevas estructuras sus páginas y enriquecerla con metadatos, de tal manera que no sea necesario la vista humana para entender la información publicada en la red.

Una de las actividades más frecuentes para un usuario de Internet es la búsqueda de información; acción que realiza a través de los buscadores Web como Google, Bing o Yahoo!, cotidianamente el usuario es muy objetivo en su búsqueda, otras ocasiones no es tan específico y los resultados llegan a ser cientos o miles, sin ni siquiera ser uno de ellos el adecuado, por tal motivo se enfrenta a una problemática de contexto y semántica.

Por ejemplo, un usuario busca la palabra “Banco”, una palabra que tiene dos significados y connotaciones, puede ser banco de sentarse (silla) o banco de datos o banco de dinero, por tal motivo el usuario debe especificar con más detalle. Ejemplos como estos son muchos, afortunadamente, buscadores como Google han trabajado por años la estructura semántica de sus resultados, pero aún así, es trabajo efectuado en su tecnología, y no directamente en los sitios Web.

La estructura de las páginas Web han evolucionado a lo largo de los años, pasando los estándares desde la versión 1 del HTML, XHTML y hasta el actual HTML5, todos bajo la supervisión y publicación de la W3C (Consorcio de la World Wide Web), dirigido por Tim Berners-Lee.

Donde cada versión del lenguaje de etiquetado tiene mayor estructuración entendible para las máquinas, y a su vez, para los propios seres humanos que las programan.

```
<html>
<head>
  <title>Título del documento</title>
</head>
<body>
<table>
<tr>
  <td><h1>Titulo del encabezado</h1></td>
</tr>
<tr><td><p>Texto informativo sobre la pagina </p></td></tr>
<tr><td><p>Mas informacion sobre la pagina</p></td></tr>
</table>
</body>
</html>
```

**Figura 1 Código de página Web con lenguaje HTML versión 1**

En la Figura 1 se muestra una página sin definiciones en su estructura de código, realizada con tablas <table> y etiquetas como <h1> como encabezado y <p> como párrafos, pero no crea una interpretación para las máquinas, pero sí para los humanos (véase Figura 2).

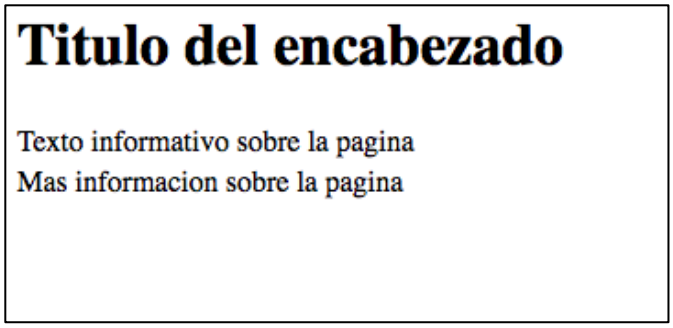



**Figura 2 Vista de página Web con lenguaje HTML versión 1**

La Web Semántica propone las pautas para presentar los contenidos, su estructura y enriquecimiento del formato con mayor descripción para sus elementos utilizando metadatos, como se muestra en la Figura 3.

```
<pagina>
  <nombre>Titulo del encabezado</nombre>
  <descripcion>Texto informativo sobre la pagina </descripcion>
  <descripcion>Mas informacion sobre la pagina</descripcion>
</pagina>
```

**Figura 3 Página Web con metadatos**

Web Semántica aplica elementos de datos enlazados, propiedad que fomenta las relaciones semánticas capaces de construir ontologías como herramientas para la representación del conocimiento, como se muestra en comparativo en la Figura 4 y Tabla 1 (Vaqué, 2014).

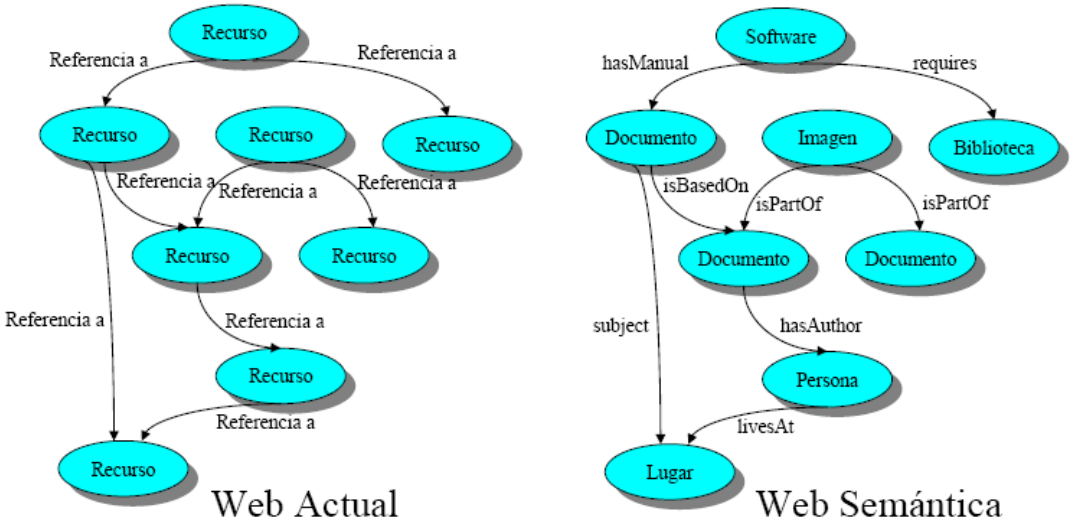



**Figura 4 Comparación entre estructura de Web actual y Web Semántica (Vaqué, 2014)**

Como se aprecia en la Figura 4 los enlaces de la Web actual se basan en hipervínculos `<href>` a recursos, sin contextualización de referencia, mientras que en la Web semántica sus relaciones son con base a estructuras ontológicas.

| **Característica** | **Web actual** | **Web Semántica** |
|---|---|---|
| Lenguaje principal | HTML | XML (RDF, OWL) |
| Forma y estructura | Documentos no estructurados | Documentos estructurados |
| Semántica | Semántica implícita | Etiquetado explícito (metadatos, web semántica) |
| Relación entre contenido y forma | HTML = fusión de forma y contenido | Estructura en capas de forma y contenido: XML + transformación (p.ej., XSL) a HTML, WML, PDF, u otros formatos |
| Editabilidad | Documentos estáticos | Documentos dinámicos |
| Descomponibilidad y recomponibilidad | Sitios web monolíticos, independientes | Bricolaje (agregación), sindicación, reasignación de contenido |
| Interactividad | Medio de difusión unidireccional | Web editable, bidireccional |
| Audiencias | Para consumo humano | Para humanos y ordenadores (p.ej., *servicios web*) |

| | | |
|---|---|---|
| Control de producción | Centralizado | Descentralizado |

**Tabla 1 Comparación entre estructura Web actual y Web Semántica (Vaqué, 2014)**

La RAE define a las ontologías como "parte de la metafísica que trata del ser en general y de sus propiedades trascendentales" (Real Academia Española, 2021), pero dentro del contexto de la Web, definen un lenguaje común para describir varios temas dentro de un dominio de discurso. Una ontología une varios esquemas en una estructura de datos, incluyendo las entidades relevantes y sus relaciones dentro del dominio.

Autores como Studer, Benjamins y Fensel definen a la ontología como "una especificación formal y explícita de una conceptualización compartida" (Studer et al., 1998). Esta abstracción de definición engloba a las ontologías como la representación formal y explícita de "estructuras de conocimiento a través de conceptos, sus propiedades, sus atributos, las relaciones con otros conceptos y los axiomas relacionados" (Paredes Valverde, 2017).

Las ontologías tienen una gran variedad de propósitos, como el razonamiento inductivo, la clasificación, y una variedad de técnicas de resolución de problemas. Uno de los principales es para establecer conceptos y relaciones, pudiendo ser compartidas por todos (Vaqué, 2014).

En resumen, una ontología tiene una lista infinita de términos y sus relaciones entre estos. Los términos son conceptos o clases de objetos los cuales forman parte importante del dominio.

Si ejemplificamos las ontologías en el mundo real, podemos abarcar un dominio de computadora, la cual tiene conceptos como pantalla, teclado, mouse, CPU, etcétera. Estos se relacionan entre sí y tiene una jerarquía de clases. A su vez, incluyen propiedades (teclado conectado al CPU) y restricciones de valor (únicamente el teclado X se puede conectar al CPU Y), enunciados disyuntivos (el microprocesador y el mouse son disjuntos), relaciones lógicas entre objetos (cada computadora debe componerse de un

microprocesador para su funcionamiento), entre más características (García Sánchez, 2007).

Las ontologías tienen un lenguaje formal para ser publicado en la Web, llamado OWL (del inglés '*Web Ontology Language*'). El OWL facilita la publicación y compartición de datos con un modelo de marcado construido sobre RDF y codificado en XML.

#### *2.2.1.2 Arquitectura de la Web Semántica*

La arquitectura de la Web Semántica está definida en capas, donde cada capa tiene una función y una tecnología semántica.

Cada capa tiene un nivel de abstracción, iniciando por la capa base, o inferior, la cual contiene tecnologías encargadas de la identificación y representación de recursos, mientras las capas superiores son las que permiten la inteligencia de la Web Semántica.

A continuación se muestra la Figura 5 la cual describe visualmente la arquitectura de la Web Semántica y posteriormente la descripción de cada una de las capas (Paredes Valverde, 2017).

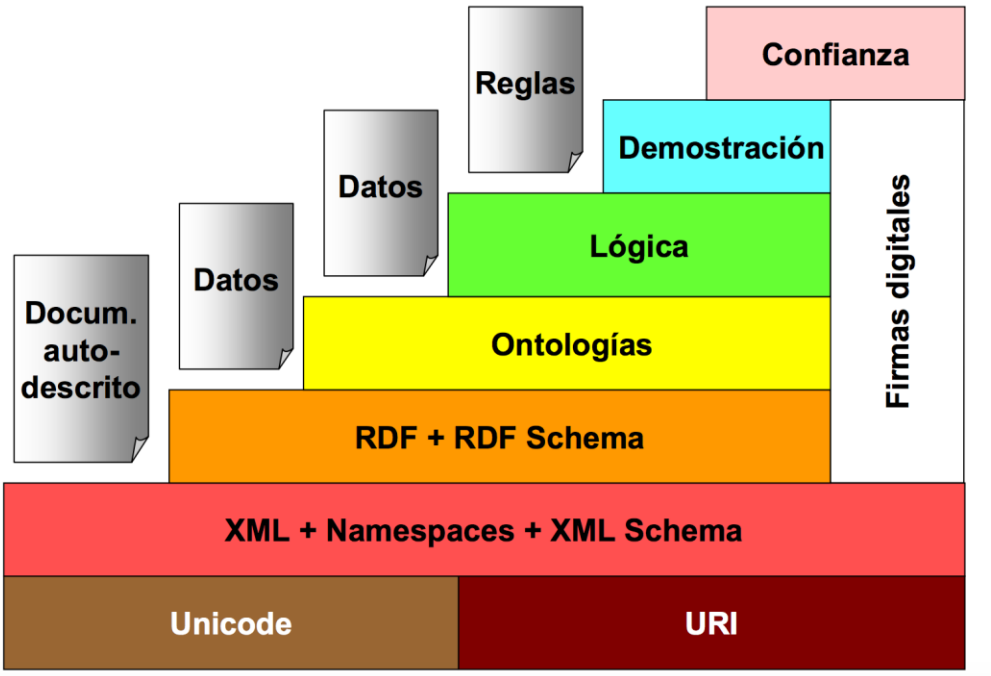



**Figura 5 Arquitectura de la Web Semántica (Universidad de Oviedo, 2005)**

- **Unicode y URI**: En el primer nivel de las capas se encuentran las tecnologías de identificación, por un lado, Unicode, es un estándar de codificación de caracteres permite que todos los lenguajes humanos puedan ser utilizados para la Web, y mientras URI (por sus siglas en inglés como *Uniform Resource Identifier*), provee de una cadena de caracteres que identifican los recursos de una red de forma univoca. En resumen, ambas tecnologías se utilizan para identificar los recursos en la Web de manera inequívoca para ser presentados en cualquier idioma.
- **XML + Namespaces + XML Schema**: El XML (por sus siglas en inglés *eXtensible Markup Language*) es un meta-lenguaje el cual permite definir, almacenar y compartir datos en forma legible en la Web. El XML NS o Namespaces proporcionan un método para evitar conflictos de nombres de elementos. Y por último, el XML Schema es el encargado de describir la estructura (elementos y atributos), tipos de datos y las restricciones de los contenidos de los documentos en lenguaje XML, también conocido como XSD (*XML Schema Definition*).

- **RDF + RDF Schema**: El lenguaje RDF permite la representación de todos los recursos dentro de la Web Semántica. Un recurso es cualquier objeto que pueda tener datos asociados a él, ya sea en el mundo real o en el abstracto. Se basa en la especificación del sujeto-predicado-objeto. Esta especificación define la relación entre dos objetos, integrado por dos sujetos y un objeto. El elemento restante, el predicado, indica la relación entre dichos elementos. Los 3 elementos de la especificación sujeto-predicado-objeto se identifica en la URI. Los documentos RDF se representa en forma de grafos etiquetados, donde sujeto y objeto son los nodos y el predicado es el arco que une a ambos nodos. El RDF Schema es la extensión semántica del lenguaje RDF, el cual proporciona los elementos básicos para la descripción de vocabularios dentro de las ontologías. Por lo que describe los recursos como clases, los organiza de forma jerárquica, define sus relaciones como propiedades, y por último, establece sus dominios y rangos.
- **Ontologías**: esta capa provee de un vocabulario con reglas para mejorar la expresividad y funcionalidad de la capa anterior con la integración de nuevos conceptos, relaciones y propiedades para permitir conceptualizar un dominio preciso.
- **Lógica**: esta capa establece las reglas de inferencia para ayudar a los agentes de software en procesar y relacionar la información, así como para convertirlo en conocimiento. Tiene como característica el proceso automático de información a nivel semántico.
- **Demostración/Prueba**: tiene como principal objetivo el ejecutar y evaluar la capa lógica, para que en conjunto con la capa superior (confianza) determinen la validez y confiabilidad de las fuentes de información.
- **Confianza**: esta capa evalúa el proceso de la capa anterior (prueba/demostración) y con ello comprobar de forma exacta las fuentes de información.
- **Firma digital**: define el ámbito de confianza entre las capas Prueba y Web Semántica. Esta firma digital es conformada por mecanismos criptográficos para permitir a máquinas y agentes de software la verificación de la seguridad y confiabilidad de la fuente.

#### *2.2.1.3 HTML5, previo a la Web Semántica*

HTML5 es la última versión del lenguaje HTML, propuesto por la W3C y publicado de manera oficial el 23 de octubre del 2014, sin embargo, las prácticas del lenguaje estaban siendo utilizado por los desarrolladores con muchos años de anterioridad.

El HTML5 nace para consolidar una Web nativa y semántica, con nuevos elementos propios del lenguaje, de tal forma que no será necesario *plugins* de terceros (como Flash) para mostrar contenidos multimedia y de cualquier tipo, sin desechar las etiquetas propuestas en las versiones anteriores de HTML o XHTML. Sus características de Open Web (o código libre) hace renderizable en cualquier navegador tanto para equipos de escritorio como para dispositivos móviles, haciendo flexible y ligero su interpretación [1].

Las versiones anteriores de HTML tuvieron muchos problemas en la interpretación y manipulación del DOM (por sus siglas en inglés como *Document Object Model*). El DOM es la interfaz de plataforma para representar los objetos en forma de árbol dentro de los documentos HTML/XHTML. Anteriormente, no existía un modelo estándar para acceder al DOM, cada navegador Web lo interpretaba de forma diferente por tal motivo se recurría a código extra de JavaScript para generar páginas dinámicas.

HTML5 además añadió nuevas tecnologías nativas, con características semánticas y mayor facilidad de lectura para los navegadores. A continuación se describe las propiedades principales de la nueva versión de HTML (Mozilla, 2022) :

- Limpieza en el código: con HTML5 se desechó mucho código para la descripción del documento y elemento raíz del HTML, con la simpleza de `<!DOCTYPE html>` la etiqueta de lenguaje `<html lang=”es”>` (o "en" para caso idioma inglés) y el formato de caracteres con el metadato `<meta charset=”utf-8” />`.
- Semántica: permite describir con mayor facilidad la estructura y la semántica de la página Web con etiquetas como `<section>`, `<article>`, `<nav>`, `<header>`, `<footer>`, `<aside>`, `<audio>`, `<video>`.

- Conectividad: con HTML5 la comunicación con el servidor toma formas nuevas e innovadoras con el uso de Web Sockets, WebRTC y eventos del servidor.
- Sin conexión y almacenamiento: permite que las páginas web almacenen datos de manera local en el lado del cliente y operar sin conexión de manera más eficiente utilizando Application Cache, WHATWG, IndexeDB y LocalStorage.
- Multimedia: proporciona un total soporte para utilizar contenido multimedia de forma nativa con el uso de las etiquetas HTML como `<audio>` y `<video>`, incluyendo los controladores por defecto, además provee de un API para manipulación de cámara Web, todo esto sin utilizar *plugins* o herramientas de terceros.
- Gráficos y efectos 2D/3D: HTML5 incluye una amplia gama de nuevas características de gráficos para la Web como lo son canvas 2D, WebGL, SVG, entre otras más.
- Rendimiento e integración: brinda una mayor optimización en velocidad y un mejor uso del hardware para dispositivos móviles y de escritorio.
- Acceso al dispositivo: se suma una característica importante de HTML5, el desarrollo de APIs para el uso de varios componentes internos de entrada y salida de los dispositivos. Con ello se omite uso de herramientas que desgastan la memoria RAM y fomentan la lentitud del dispositivo.
- CSS3: se une la nueva versión del CSS (por sus siglas en inglés *Cascading Style Sheets*), con muchas novedades nativas de la propia hoja de estilo como son esquinas redondeadas, sombras, gradientes, transiciones o animaciones, y nuevos layouts como multi-columnas, cajas flexibles o maquetas de diseño en cuadrícula (*grid layouts*).

### 2.2.2 Ontologías y grafos de conocimiento

#### *2.2.2.1 Definición de ontología*

Una de las tecnologías principales de la Web semántica son las ontologías. Las ontologías se definen como “una especificación explícita y formal de una conceptualización

compartida" (Studer et al., 1998). En el campo de la Inteligencia Artificial (IA) las ontologías tienen como objetivo representar el conocimiento basado en la formalización declarativa y simbólica (Russel & Norvig, 2009), lo que quiere decir que "define los términos y relaciones que conforman el vocabulario de un área temática, así como las reglas para combinar términos y relaciones para definir extensiones del vocabulario" (Neches et al., 1991).

Una de las definiciones de ontología muy citada es la descrita por Gruber en (Gruber, 1993):

"Una ontología es una especificación explícita de una conceptualización"

Gruber considera a la conceptualización como entidades compuestas por objetos, conceptos y relaciones dentro de una determinada área. Por *explícita*, se entiende como las restricciones definidas.

Borst actualiza la definición basada en la de Gruber, de la siguiente manera (Borst, 1997):

"Una ontología es una especificación formal de una conceptualización compartida"

En esta redefinición, Borst agrega *formal*, para hacer énfasis en la necesidad de establecer de ontologías comprensibles por los sistemas informáticos.

Posteriormente, y tomando como base las definiciones previas, en (Studer et al., 1998) se definió a la ontología como "una especificación *formal* y *explícita* de una *conceptualización compartida*", incluyendo cuatro conceptos principales: formal, explícita, conceptualización y compartida. En dicho trabajo se examinó a profundidad la nueva definición llegando a las siguientes conclusiones:

"La *conceptualización* se refiere a un modelo abstracto de algún fenómeno en el mundo al haber identificado los conceptos relevantes de ese fenómeno. *Explícita* significa que el tipo de conceptos utilizados y las restricciones sobre su uso están definidos explícitamente. Por ejemplo, en los dominios médicos, los conceptos son enfermedades y síntomas, las relaciones entre ellos son causales y una restricción es que una enfermedad no puede

causarse a sí misma. *Formal* se refiere al hecho de que la ontología debe ser legible por máquina, lo que excluye el lenguaje natural. *Compartida* refleja la noción de que una ontología captura conocimiento consensuado, es decir, no es privado de algún individuo, sino aceptado por un grupo."

Podemos destacar un proceso evolutivo en la definición de las ontologías basado en las características de formalidad y explicitud. Como resultado, una ontología es un modelo que representa formalmente y explícitamente estructuras de conocimiento a través de conceptos, propiedades, atributos, relaciones y axiomas. Estos elementos de la ontología se describen en el siguiente apartado 2.2.2.2.

#### *2.2.2.2 Elementos de la ontología*

La ontología se formaliza a través de cinco elementos fundamentales (Neches et al., 1991) (Rodríguez García, 2014): clases, atributos, relaciones, axiomas e instancias.

- Una **clase** o también conocido como **concepto**, representa cualquier entidad u objeto dentro de un dominio. Estos brindan mecanismos para la agrupación de recursos con características compartidas. Tienen la posibilidad de tener diferentes atributos y relaciones entre conceptos.
- Los **atributos** representan la estructura interna de los conceptos. Se clasifican en dos tipos de acuerdo con su origen: específicos y heredados. Los específicos son propios del concepto del cual pertenecen, mientras que los heredados, por sus relaciones taxonómicas en las que el concepto desempeña el rol de hijo y por consiguiente, hereda los atributos de la clase padre.
- Las **relaciones** representan las interrelaciones de los conceptos dentro de un dominio. Se definen como cualquier subconjunto de un producto de n conjuntos, esto se describe formalmente como: R: C1 x C2 x … x Cn. Entre los diferentes tipos de relaciones se encuentran las taxonómicas, mereológicas o partonómicas, pero finalmente, las binarias son las más destacadas.
- Los **axiomas** son expresiones que son siempre ciertas. Son incluidas dentro de la ontología con diversos propósitos, como especificar el significado de los

componentes del modelo ontológico, establecimiento de restricciones sobre los atributos, argumentos de relaciones, entre otras más.

- Las **instancias** o también conocidas como individuos, representan elementos concretos dentro de un dominio, los cuales se describen como términos de sus conceptos. En síntesis, son las ocurrencias en el mundo real de los conceptos.

#### *2.2.2.3 Tipos de ontología*

A continuación se describen los tipos de ontología basados en el trabajo de Geri Steve y colaboradores (Steve et al., 1997):

- **Ontologías de dominio**: representan el conocimiento especializado relacionado un área de conocimiento o dominio específico, tales como pueden ser político (Guedea-Noriega & García-Sánchez, 2020), médico (Shah et al., 2015), psicología (Jung et al., 2017), nutrición (Kim et al., 2017), entre otros.
- **Ontologías genéricas**: se representan por conceptos generales y fundacionales del conocimiento tales como las estructuras parte/todo, cuantificación, y tipos de objetos. Son especialmente reutilizables en otros dominios.
- **Ontologías representacionales:** especifican las conceptualizaciones que implican los formalismos de representación del conocimiento, también se les denominan como meta-ontologías (ontologías de alto nivel).

Guarino clasifica las ontologías de acuerdo a su nivel de dependencia con respecto a una tarea en particular o tipo de enfoque (Guarino, 1998) (ver Figura 6):

- **Ontologías de alto nivel**: describen conceptos generales como espacio, tiempo, objetos, eventos, accesiones, las cuales están independientes del dominio en específico.
- **Ontología de dominio y de tarea**: describen el vocabulario relacionado en un dominio en particular y/o actividad/tarea específica.

- **Ontología de aplicación**: describen los conceptos dependientes del dominio en particular y/o actividad/tarea específica.

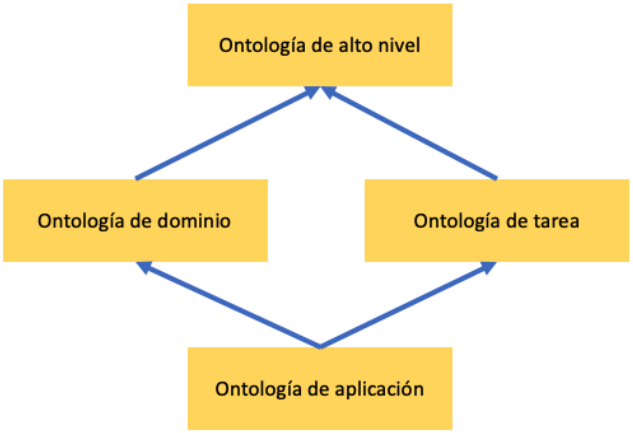



**Figura 6 Clasificación de las ontologías por tipo de enfoques (Guarino, 1998)**

Es importante subrayar que las clasificaciones de las ontologías se realiza bajo criterios de los autores antes mencionados, sin embargo, pudieran existir otras clasificaciones más extendidas o especificas dependiendo el área de desarrollo y aplicación.

#### *2.2.2.4 Lenguajes para el desarrollo de ontologías y grafos de conocimiento*

A lo largo de los primeros años y en la actualidad, han surgido diversos lenguajes para el desarrollo de ontologías. Sin embargo, en este trabajo de investigación nos enfocaremos en dos de las principales y las cuales son mantenidas por el Consorcio de la Word World Wide (W3C) como son RDF y OWL.

##### 2.2.2.4.1 Resource Description Framework (RDF)

RDF es un modelo de datos para representar recursos y sus relaciones. RDF se define sobre la tecnología de representación semántica XML y se basa en tripletas “sujeto–predicado–objeto” (W3C, 2014). Donde dos nodos (sujeto y objeto) se unen a través de un arco (predicado).

Cada tripleta define una conexión entre dos entidades del grafo de conocimiento. El conjunto de relaciones aceptables y tipos de entidades define la ontología o modelo ontológico del KG, que también es su estructura general. Por ejemplo, puede ser un grafo de objetos geográficos, estructuras biomédicas o páginas web. Dada una colección de entradas, KG nos permite realizar inferencias (Fensel et al., 2020).

Estas tripletas expresan afirmaciones tales como: *Héctor* (sujeto) *nació en* (predicado) *Colima* (objeto) (véase Figura 7) y se representan por un grafo RDF a través de identificadores basados en URI (por sus siglas en inglés de *Universal Resource Identifier*), es decir, por una cadena de caracteres estandarizados que permiten identificar los recursos dentro del grafo (véase Figura 8).

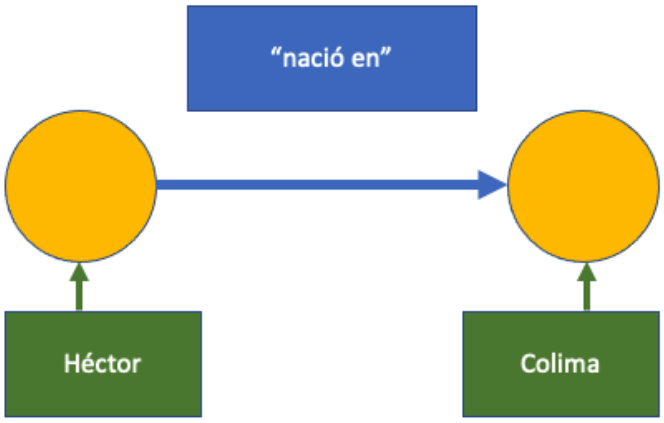



**Figura 7 Tripleta de sujeto "Héctor" - predicado "nació en" - objeto "Colima"**

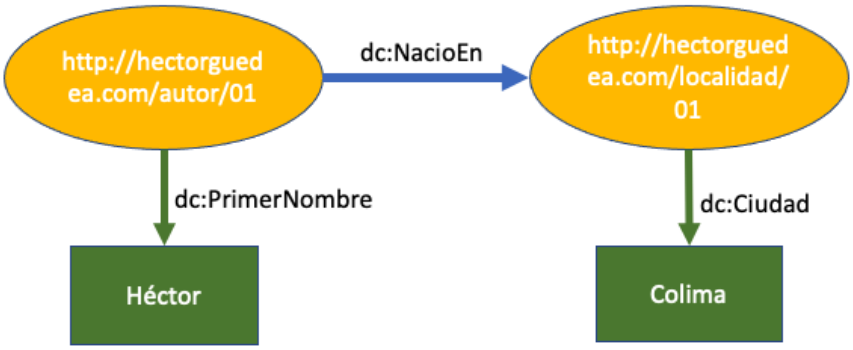



**Figura 8 Ejemplo de Grafo RDF basado en la tripleta de la Figura 7**

RDFS provee una extensión del RDF añadiendo nuevo vocabulario y metadatos tales como `rdfs:Class`, `rdfs:Resource` y `rdf:Property`, los cuales permiten definir clases, recursos y propiedades (Shadbolt et al., 2006).

#### 2.2.2.4.2 Web Ontology Language (OWL)

OWL es un lenguaje estándar para el desarrollo de ontologías, es mantenido por la W3C, presentado el 2004 (McGuinness & van Harmelen, 2004) y actualizado en el 2009 como en la última versión OWL2 (W3C, 2012).

El lenguaje es utilizado para representar de manera explícita el significado de los términos en vocabulario y relaciones entre los mismos. Está diseñado para favorecer a los agentes de software en el procesamiento y exploración la información de forma más significativa, ya que provee de vocabulario adicional que permite describir clases y propiedades a través de una semántica formal.

En OWL, al igual que en RDFS, los elementos básicos son tres: clases, propiedades e individuos. El primer elemento define al grupo de individuos que comparten propiedades similares. Los individuos se definen como instancias de clase. Las propiedades se representan en dos tipos *Data Property* y *Object Property*. La primera propiedad relaciona individuos con un valor literal, por ejemplo, para una instancia de la clase Persona sería *`fechaNacimiento`*, valor fecha y número. La segunda propiedad define y establece relaciones entre los individuos o instancias, por ejemplo, `EsCasadoCon` relacionaría una instancia de Persona con otra.

La reciente versión de OWL2 añade nuevas funcionalidades para la mejora de la expresividad del lenguaje. Entre ellas destaca la posibilidad de definir las claves en las clases, cadenas de propiedades, tipos de datos y rangos más complejos, restricciones de cardinalidad como son las propiedades asimétricas, reflexivas y disjuntas (W3C, 2012).

En la Figura 9 presenta las relaciones existentes entre las dos versiones del lenguaje, sub-lenguajes y perfiles.

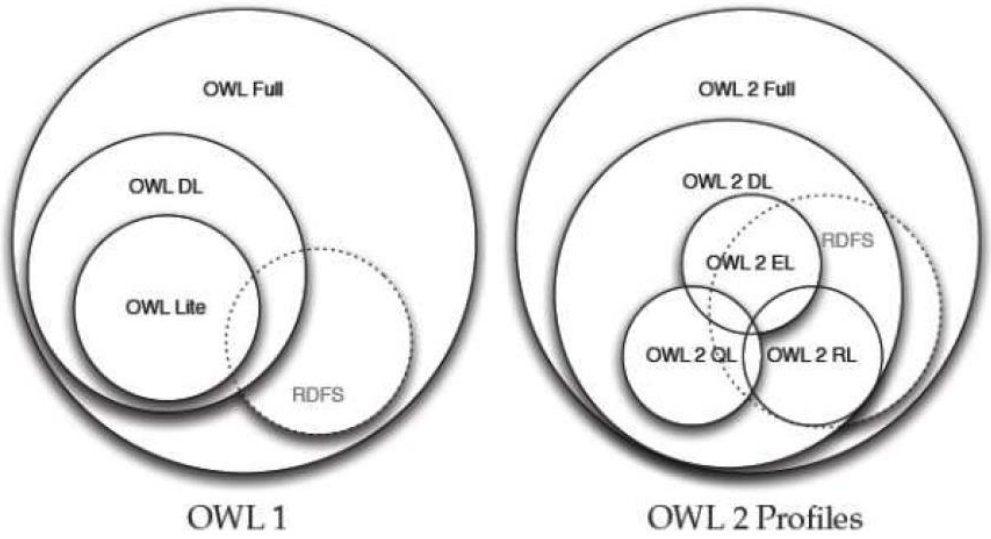



**Figura 9 Representación gráfica de la relación entre los lenguajes y perfiles OWL**

Existen editores de código para ontologías y sistemas de adquisición de conocimiento con el propósito de facilitar la manipulación, gestión y visualización. Protégé (https://protege.stanford.edu/) es un IDE desarrollado por la Universidad de Stanford, en colaboración con la Universidad de Mánchester, brinda también la posibilidad de incorporar nuevos componentes basados en las necesidades y requerimientos de la usuario y ontología (N. F. Noy & Mcguinness, 2001) (ver Figura 10).

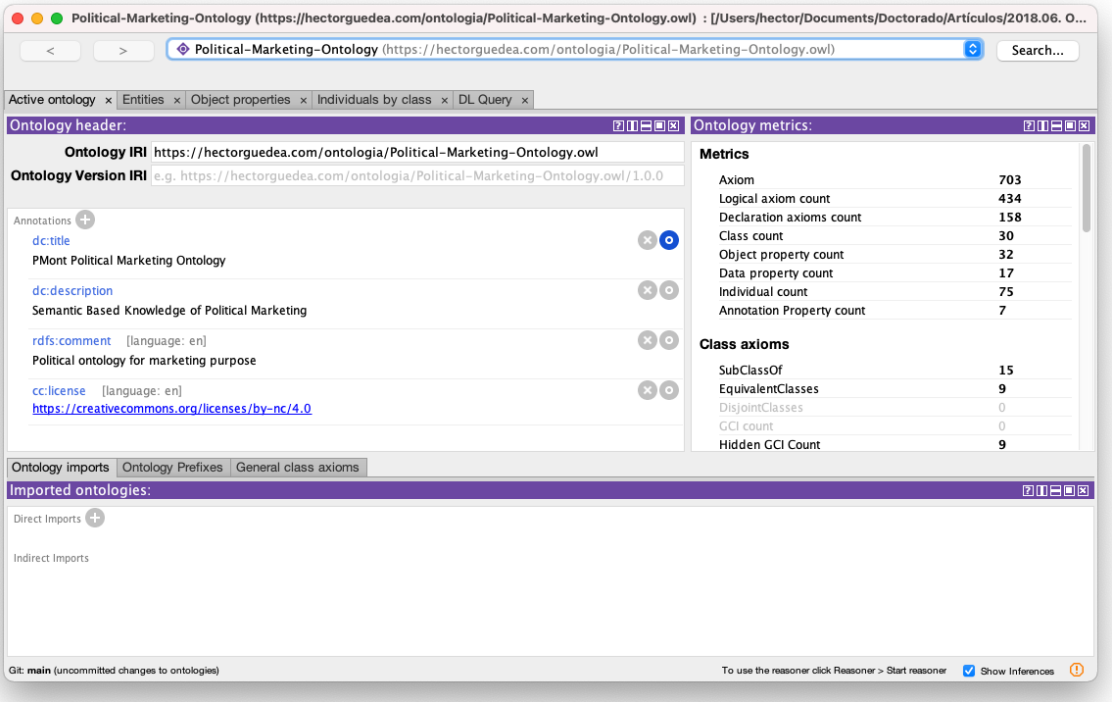



**Figura 10 Pantalla inicial de Protégé al abrir ontología**

### *2.2.2.5 Lógica y razonamiento*

Para que los sistemas de software tengan la posibilidad de realizar procesos de inferencia sobre la información almacenada en las ontologías, se debe fundamentarse a través de un formalismo lógico (García Sánchez, 2007). La lógica brinda la estructura y reglas de inferencia formal a la representación del conocimiento (Paredes Valverde, 2017).

El razonamiento permite inferir nuevo conocimiento representado implícitamente sobre la base de conocimiento (Tessaris, 2009). La lógica descriptiva (DL, por sus siglas en inglés *Description Logic*) es un lenguaje utilizado para la representación del conocimiento terminológico de un dominio de aplicación de una forma estructurada y formalmente comprendida. Un sistema de representación de conocimiento o KR (*Knowledge Representation*) basado en DL provee de beneficios para llevar a cabo las configuraciones

de la base del conocimiento, razonamiento y gestión, en la Figura 11 se muestran los componentes principales. Donde *TBox* define el "componente de terminología" y describe un dominio de aplicación estableciendo las clases, propiedades y restricciones como un vocabulario de dominio. Mientras que *ABox* es el "componente de aserción", contiene declaraciones sobre los individuos o instancias en términos de su vocabulario (Paredes Valverde, 2017).

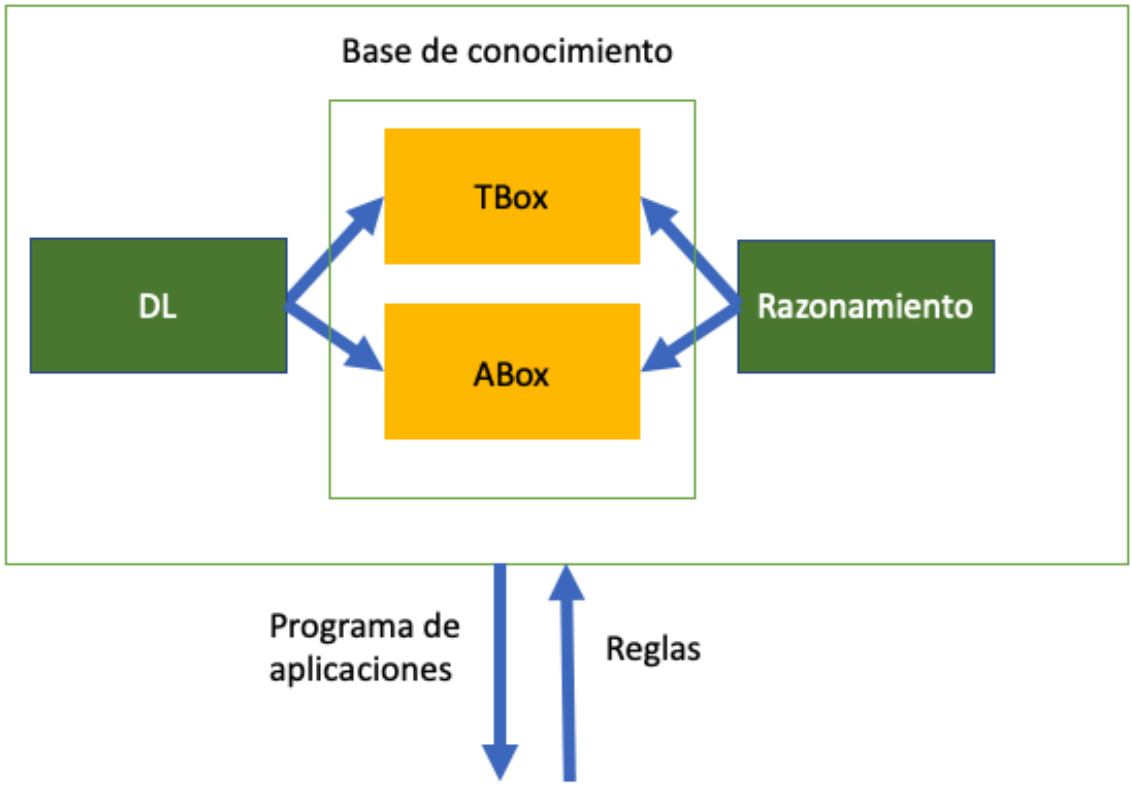



**Figura 11 Arquitectura de un sistema de representación de conocimiento basado en lógica descriptiva (DL)**

#### *2.2.2.6 Definición de grafo de conocimiento*

El término grafo de conocimiento (KG) fue acuñado por Google en 2012, refiriéndose a su uso del conocimiento semántico en la búsqueda web, y a su vez se usa para referirse a bases de conocimiento de la Web Semántica como DBpedia o YAGO. Desde una perspectiva más amplia, cualquier representación basada en grafos de algún conocimiento podría considerarse un grafo de conocimiento (esto incluiría cualquier tipo de conjunto de datos RDF, así como ontologías lógicas de descripción). Sin embargo, no existe una definición

común sobre qué es y qué no es un KG. En lugar de intentar una definición formal de lo que es un KG, nos restringimos a un conjunto mínimo de características de los grafos de conocimiento, que usamos para diferenciar los KG sobre otras colecciones de conocimiento que no se considerarían como grafos de conocimiento (N. Noy & Paulheim, 2016):

- Describe principalmente entidades del mundo real y sus interrelaciones, organizadas en un grafo triple.
- Define posibles clases y relaciones de entidades en un esquema.
- Permite potencialmente interrelacionar entidades arbitrarias entre sí.
- Posibilidad de cubrir varios dominios temáticos.

Desde sus inicios, la Web Semántica ha promovido una representación del conocimiento basada en grafos, por ejemplo, impulsando el lenguaje estándar RDF. El grafo sobre el lenguaje RDF se expresa en triples *sujeto-predicado-objeto* (ver Figura 12), mencionadas a detalle en los apartados anteriores. En tal representación de conocimiento basada en grafos, las entidades, que son los nodos del grafo, están conectadas por relaciones, que son los bordes del grafo, y las entidades pueden tener tipos, expresados por su relación. En muchos casos, los conjuntos de posibles tipos y relaciones se organizan en un esquema u ontología, que define sus interrelaciones y restricciones de uso.


**Figura 12 Triple sujeto-predicado-objeto**

En el ejemplo a continuación (Figura 13), se muestra un grafo basado RDF con diversas afirmaciones, relaciones y propiedades en los triples con respecto a la banda Los Beatles, en donde hace alusión a sus integrantes y canciones.

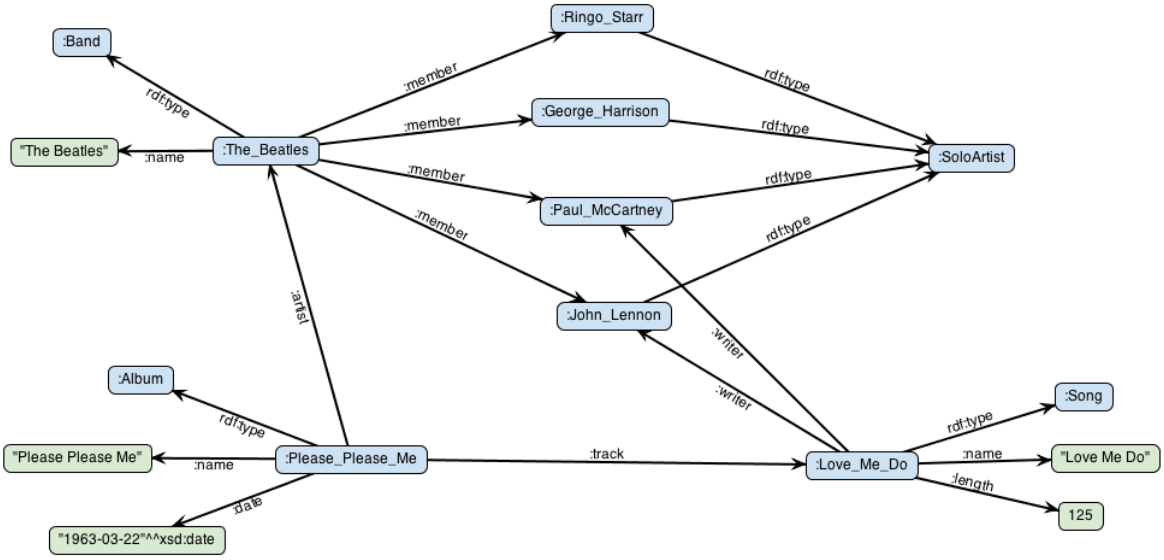



**Figura 13 Ejemplo de RDF Graph sobre Los Beatles** (Stardog, 2022)

Con la llegada de los datos enlazados, se propuso interconectar diferentes conjuntos de datos en la Web Semántica. Mediante la interconexión, la colección resultante podría entenderse como un gran grafo de conocimiento global (aunque de naturaleza muy heterogénea). Hasta la fecha, miles de conjuntos de datos están siendo interconectados en la nube de datos abiertos vinculados, y la mayoría de los enlaces conectan entidades idénticas en dos conjuntos de datos (N. Noy & Paulheim, 2016).

### 2.2.3 Datos enlazados (*Linked Data*)

Una de las características principales de la Web Semántica es enlazar los datos, ya sean distribuidos libremente o de forma privada en la Web, con el propósito de generar vínculos entre estos para favorecer la contextualización y semántica de la información, es por ello que nace el conjunto de buenas prácticas llamado como Datos Enlazados o Linked Data (W3C, 2022).

La combinación de Web Semántica y datos enlazados incrementa su potencial no solamente como publicación de documentos, sino a su vez, la información relacionada que describa al contenido, su significado y la relación de datos (Biblioteca del Congreso Nacional de Chile, 2012).

Las relaciones de la Web actual funcionan por medio de hipertextos e hipermedia, la vinculación de textos, imágenes y videos sobre documentos en HTML; en cambio, la Web Semántica propone el entrelazado de datos a través de documentos RDF, tecnología utilizada para describir los recursos Web, especificar metadatos y representar información, además con la serie de reglas descritas por la semántica.

Por ejemplo, una empresa X desea vincular su localización a través de una fuente externa, por tal motivo, se enlazarán los datos entre ambas, de esta manera tanto la fuente externa como la empresa X enriquecen su información en ámbitos geográficos, y porque no decirlo, el propio usuario de Internet tomará beneficios en la búsqueda y resultado sobre estos contenidos (véase Figura 14).

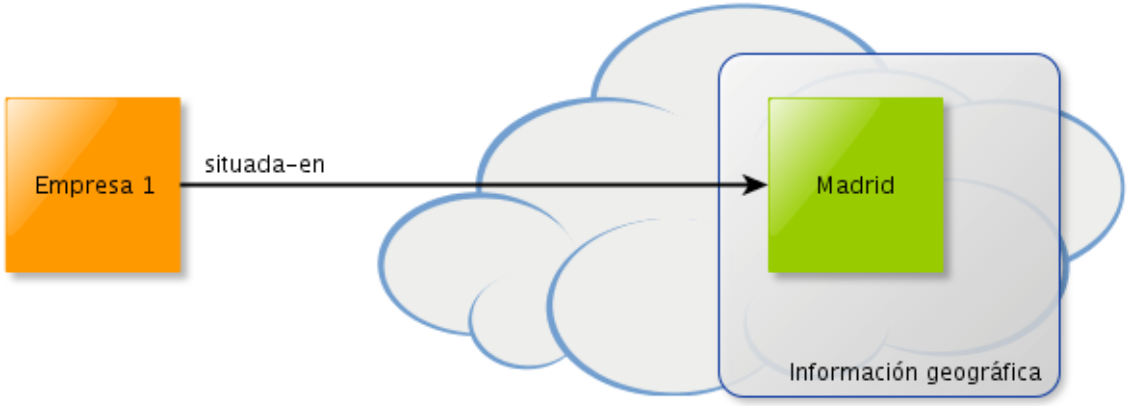



**Figura 14 Ejemplo de aplicación de Linked Data**

En general, Linked Data permite construir la Web de los datos, una gran base de datos interconectada y distribuida por la Web, donde cada dato se vincula con otro, lo que contribuirá a la exploración automática y manual entre los documentos.

El siguiente grafo (ver Figura 15) representa a distintos conjuntos de datos de diversos tipos, organizados mediante colores por dominios. Estos conjuntos de datos están conectados entre sí de forma que componen la “Nube de Linked Data” o “Nube de Datos Enlazados” (W3C, 2022).

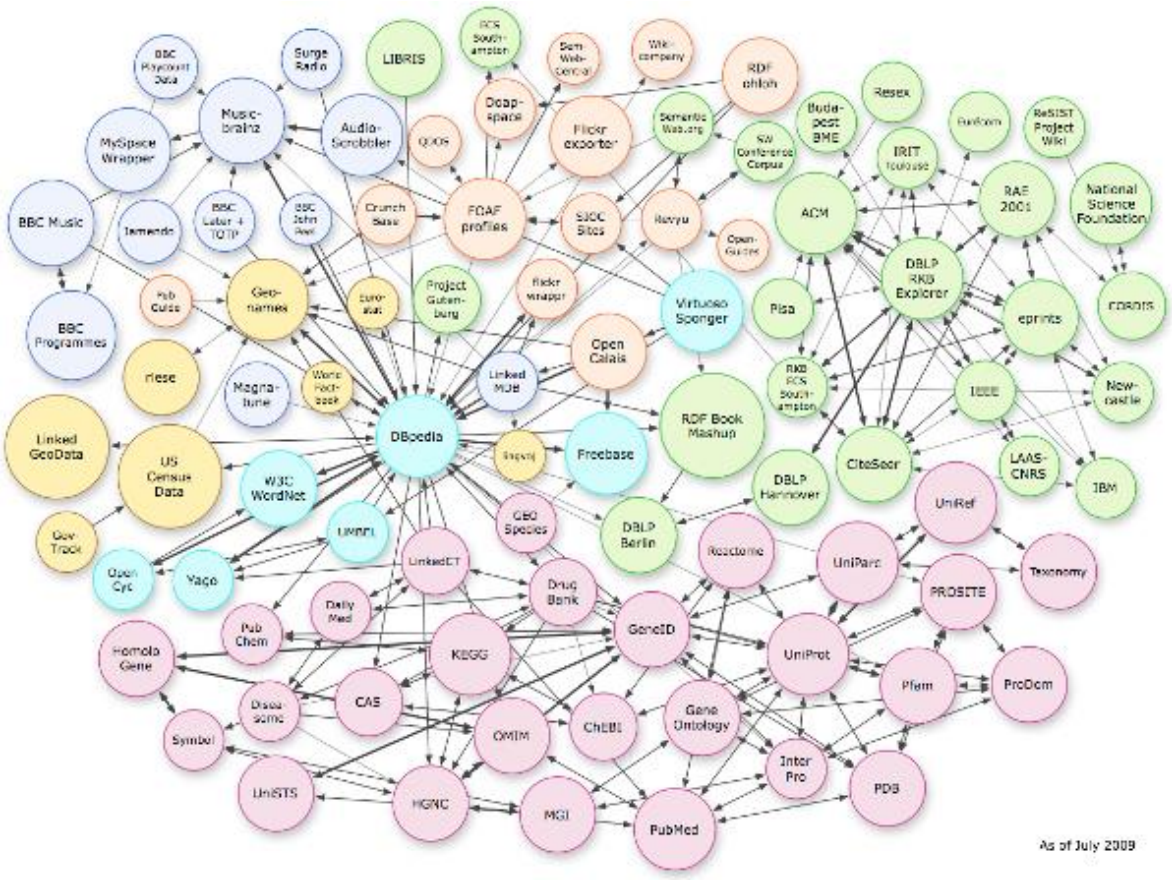



**Figura 15 Grafo de Nube de Datos Enlazados**

El funcionamiento del Linked Data en la Web Semántica funciona con la aplicación de principios básicos tanto para fortalecer el crecimiento de la Web hipertextual, con documentos HTML, como la Web de datos, con documentos descritos en RDF.

Se especificaron cuatro reglas para el desarrollo de Linked Data en Web Semántica:

1. Usar URIs para identificar las cosas: al nombrar los recursos, conceptos o cosas con URIs se ofrece una forma estándar, unívoca y precisa de referencia. Esto es muy usable para la abstracción del lenguaje natural, para conseguir evitar ambigüedades, por ejemplo, los propios nombres de los objetos en distintos idiomas. Un objeto puede ser descrito como “celular” (en español de México), “móvil” (en español de España) o “cellphone” (en inglés), el cual se identificará con una sola referencia URI para todos, sin importar idioma.

2. Usar URIs HTTP: existen muchos esquemas sobre URIs, por lo tanto, se usará las URIs sobre HTTP (por ejemplo, http://dbpedia.org/resource/Cellphone) para afirmar que cualquier recurso pueda ser buscado y accedido en la Web. Se debe tomar en cuenta los URIs no son solo direcciones, sino también identificadores de los recursos.
3. Ofrecer información sobre los recursos usando RDF: después del proceso de búsqueda y acceso a un recurso identificado mediante una URI HTTP, es necesaria la obtención de información útil sobre el recurso, por lo que se utilizará la presentación de la información mediante documentos RDF (véase Figura 16). Así también, si se realizan consultas avanzadas con SPARQL, el resultado será interpretado de forma automática.
4. Incluir enlaces a otros URIs: se enfatiza en conectar datos para compartir información tanto con fuentes externas como internas, con el objetivo de que no queden datos aislados o fuera de un dominio o relación. Esto favorece a enriquecer los recursos en su contextualización y semántica. Por ejemplo, un celular tiene marca o modelo, los cuales pueden ser representados con URIs HTTP y enlazados entre el objeto dentro del documento RDF, como: http://dbpedia.org/resource/iPhone y http://dbpedia.org/resource/Apple (véase Figura 17).

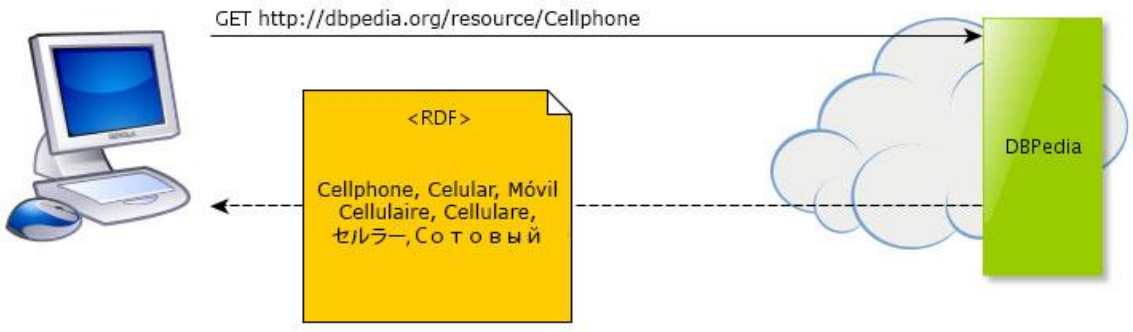



**Figura 16 URIs y RDF**


**Figura 17 Relación entre documentos RDF y URI HTTP**

A través de estas reglas, cualquier recurso es apto para ser enriquecido con cualquier tipo de información especializada, incluso la que no se espera que sea combinable. Con ello se garantiza que al publicar información en RDF y utilizar URIs, cualquiera podría hacer referencia a esos datos. Teniendo así, una Web de datos.

### 2.2.4 Instanciación de ontologías: enriquecimiento de grafos de conocimiento

La Web es un escenario complejo para la extracción de información debido a la enorme cantidad de datos, la variedad tecnológica y la heterogeneidad de las fuentes. Las ontologías se han convertido en la tecnología fundamental de la Web semántica que permite representar el conocimiento de forma legible por máquina y organizar la información en una estructura homogénea (Beydoun et al., 2005). Se han concebido algunas metodologías que guían el desarrollo de ontologías, ofreciendo consideraciones como la determinación del dominio específico y el alcance ontológico, la reutilización de ontologías existentes, el diseño de T-Box (clases, conceptos, propiedades y relaciones) y la instanciación (A-Box) (Guedea-Noriega & García-Sánchez, 2020).

La construcción manual de ontologías se convierte en un grupo riguroso de tareas, que consumen mucho tiempo, exigen esfuerzo humano, son propensas a errores y requieren algunas habilidades técnicas (Somodevilla García et al., 2018). Es posible resolver estos problemas implementando herramientas automáticas o semiautomáticas para la creación de ontologías, también denominadas OL (Asim et al., 2018). OL comprende métodos y técnicas para la extracción, generación o adquisición de ontologías a partir de textos en lenguaje natural, que implican preprocesamiento de datos, extracción de

términos/conceptos (C) y relaciones (R), adquisición y evaluación de axiomas ( (Asim et al., 2018). En síntesis, OL es el enfoque para automatizar el proceso de adquisición de conocimiento.

OP es la tarea de agregar nuevas instancias de conceptos o relaciones a la ontología, incluyendo también procesos para eliminar instancias redundantes (Petasis et al., 2011). OP no cambia la estructura de la ontología; por lo tanto, la jerarquía de conceptos y las relaciones no taxonómicas no se modifican. OP requiere una ontología de dominio existente y específica con estructura definida (Petasis et al., 2011). Los sistemas OP están relacionados con los sistemas de extracción de información basados en ontologías, por las similitudes en los procedimientos utilizados para asociar datos con conceptos e instancias de ontología (Wimalasuriya & Dou, 2010). El proceso OP manual requiere mucho tiempo y es susceptible a errores humanos; por lo tanto, es muy recomendable utilizar procesos automáticos o semiautomáticos que aprovechen diversas técnicas, como NLP y extracción de información, para adquirir y clasificar instancias de ontologías (Faria et al., 2014).

En los últimos años, los grafos de conocimiento (KG) se han vuelto populares debido a la gran demanda de tecnología de grafos para dar sentido a los datos (Ontotext, 2020). Los KG son un tipo específico de grafo con énfasis en la comprensión contextual. Los KG son conjuntos de hechos interrelacionados que describen entidades, eventos o cosas del mundo real y sus interrelaciones en un formato comprensible para humanos y máquinas (Barrasa et al., 2021).

En una definición simple y clara, los KG representan las palabras como nodos y las relaciones como bordes entre palabras (Kertkeidkachorn & Ichise, 2018). Con todo, los KG permiten que las personas y las máquinas mejoren las conexiones en sus conjuntos de datos (Yoo & Jeong, 2020).

El modelo de grafo de propiedades es el método más popular para crear KG, que consiste básicamente en nodos que representan entidades en el dominio y relaciones que representan cómo se interrelacionan las entidades (Barrasa et al., 2021). Las ontologías permiten tipos de relaciones más complejas que los KG basados en grafo de propiedades, por ejemplo, se pueden establecer propiedades como *part_of*, *compatible_with* o *depend_on* entre

categorías. También permiten la definición de relaciones jerárquicas y la caracterización adicional de las relaciones (p. ej., si son transitivas, simétricas, reflexivas, etc.) (Xu et al., 2020)(Barrasa et al., 2021).

Un KG construido sobre un esquema de ontología adquiere e integra información en una ontología y puede aplicar un razonador para obtener nuevos conocimientos (Yoo & Jeong, 2020).

Las fuentes de datos no estructurados son los recursos de conocimiento más dominantes en la Web y, en consecuencia, los más explotados en los sistemas OP. Por lo general, el primer proceso en un OP es la adquisición de datos, incluida la identificación del tipo de fuente (es decir, imagen, video, texto), que influye en las técnicas de extracción de información que se utilizarán (Lubani et al., 2019).

Existen muchas técnicas de extracción de información, como las utilizadas con interfaces de programación de aplicaciones (API), el web scraping y el uso de bibliotecas especializadas para extraer datos heterogéneos de las redes sociales y los sitios web de noticias.

La popularidad y el crecimiento de las API en los últimos años se debió principalmente a la existencia de plataformas abiertas, puntos de acceso a la información en la nube estandarizados y aplicaciones de terceros. Sin embargo, las limitaciones actuales impuestas por las estrictas políticas de privacidad que aplican a empresas como Facebook, Twitter e Instagram en cuanto a exposición de datos, impiden el acceso a información sensible de los usuarios a través de las API disponibles originalmente.

En Europa, el Reglamento General de Protección de Datos (GDPR) sugiere una comprensión más sólida de la protección de datos y la privacidad de los datos de las redes sociales de las personas (Guzmán-Guzmán et al., 2021). Las API más recientes permiten la capacidad de consultar y obtener acceso a porciones de flujos de actividad social a partir de datos de un solo usuario autenticado, pero no para obtener los datos públicos y globales del usuario. Por ejemplo, en 2015, Facebook cerró su API de búsqueda pública, que brindaba acceso de búsqueda a todas las publicaciones públicas. En 2018, Facebook también cerró el

acceso a su API de páginas, que había permitido a los investigadores extraer publicaciones, comentarios y metadatos asociados de las páginas públicas de Facebook. El cierre de la API de Facebook de la página eliminó todos los términos de servicio (TOS) de acceso al contenido de Facebook. Por lo tanto, actualmente no hay forma de extraer contenido de forma independiente de Facebook sin violar sus TOS (Acker & Kreisberg, 2020).

El web scraping es una de las posibles soluciones al problema de las APIs en el acceso a datos mencionado anteriormente. Se refiere a la práctica de extraer automáticamente contenido de páginas web y otros archivos digitales. De hecho, es una tecnología antigua anterior a las API, más complicada en términos de usabilidad, pero muy flexible y potente. Una de las desventajas prácticas de las soluciones de web scraping es que son más difíciles de aprender y aplicar que las herramientas basadas en API.

El web scraping implica examinar y comprender la estructura del modelo de objeto de documento (DOM) de las páginas. Los elementos DOM que contienen los datos deseados deben insertarse en las funciones de la herramienta elegida y la salida capturada en un archivo. No solo es un flujo de trabajo más complejo de lo que normalmente requieren las API, sino que también implica una solución personalizada para cada sitio web o plataforma de redes sociales (Freelon, 2018).

En los últimos años, el uso de técnicas de web scraping ha aumentado debido a la falta de implementaciones de API y las restricciones de TOS de muchos sitios de redes sociales. Las bibliotecas como Facebook Scraper (https://pypi.org/project/facebook-scraper/) y Twitter Scraper (https://pypi.org/project/twitter-scraper/) son muy populares y se usan a menudo como soluciones de scraping de SNS. Es importante destacar que, como se indicó anteriormente, cada sitio de redes sociales tiene reglas muy específicas y estrictas sobre cómo usar los datos del usuario en función de sus TOS. Aunque la mayoría de los sitios de redes sociales presentan su propia API, que incluye métodos para obtener diferentes tipos de datos de los usuarios (detalles del perfil, número de amigos y más), no permiten el uso de datos públicos y, por lo tanto, el uso de técnicas como web scraping podría infringir los TOS (Sapountzi & Psannis, 2018).

Después de la adquisición de datos, los siguientes pasos de un sistema OP típico son el proceso de extracción de conocimiento (es decir, anotación semántica, NLP y extracción semántica), seguido del proceso de población (es decir, eliminación de redundancia, verificación de consistencia e instanciación de ontología) y, finalmente, la etapa de almacenamiento (generalmente en un *TripleStore*) (Somodevilla García et al., 2018) (Petasis et al., 2011)(Lubani et al., 2019) (Vargas-Vera et al., 2007) (Ayadi et al., 2019). La anotación semántica es el proceso de agregar metadatos sobre conceptos, entidades y relaciones sobre documentos y datos no estructurados. Las máquinas suelen utilizar anotaciones semánticas para comprender el contexto. Los documentos etiquetados semánticamente son más fáciles de encontrar, interpretar, combinar y reutilizar (Somodevilla García et al., 2018).

Las tareas de NLP permiten encontrar anotaciones lingüísticas en textos y documentos no estructurados, incluidos límites de oraciones y tokens, partes del discurso (PoS, por sus siglas en inglés de *Parts of Speech*), entidades con nombre, valores numéricos y de tiempo, análisis de dependencia y circunscripción, atribuciones y relaciones. El proceso de PNL también permite obtener elementos semánticos para incorporarlos a estructuras semánticas como ontologías (Pech et al., 2017).

Existe una gran variedad de aplicaciones y herramientas dedicadas a realizar tareas de PNL como SpaCY (https://spacy.io), Stanford CoreNLP (https://stanfordnlp.github.io/CoreNLP/), NLTK (https://www.nltk.org/) y Textrazor (https://www.textrazor.com) (Achichi et al., 2015).

En general, se pueden distinguir tres enfoques principales cuando se trata de OP para extraer términos específicos de dominio (Ayadi et al., 2019): ( i ) sistemas de población de ontología basados en reglas (Lubani et al., 2019) (Bereta et al., 2021), (ii) sistemas de población de ontología que utilizan ML o NLP (Vargas-Vera et al., 2007)(Ait-Mlouk et al., 2020), y (iii) sistemas de población de ontología que utilizan enfoques estadísticos (Achichi et al., 2015).

En los sistemas basados en reglas, el grupo de reglas está diseñado y desarrollado para la ubicación, clasificación y extracción de información en categorías predefinidas, como personas, organizaciones, expresiones de tiempo, lugares, conocidas como entidades con nombre.

Los NER son herramientas y técnicas utilizadas para la automatización de estas tareas. En (Lubani et al., 2019) los autores describen el uso de NER, un conjunto de predicados predefinidos y una base de datos léxica (WordNet) para generar las reglas de clasificación de instancias que se utilizarán para poblar la ontología. Otra contribución se describe en (Bereta et al., 2021), donde los datos heterogéneos se convierten en triples RDF virtuales y esa representación virtual se consulta sobre la marcha sin materializarla como triples RDF. Este paradigma basado en reglas se centró en R2RM (https://www.w3.org/TR/r2rml/), un lenguaje para expresar asignaciones de datos personalizadas, SPARQL como lenguaje de consulta RDF y un sistema de acceso a datos basado en ontologías (OBDA) para admitir el acceso uniforme a los datos almacenados en fuentes heterogéneas.

En contraste, los autores en (Vargas-Vera et al., 2007)proponen otra estrategia para la tarea OP que hace uso de un esquema ontológico predefinido para la comparación de términos, junto con algoritmos de aprendizaje que usan tareas ML y NLP para la extracción de información de datos no estructurados. La principal desventaja es la supervisión en la selección de los conceptos iniciales y la construcción del esquema. En la misma línea, en (Ait-Mlouk et al., 2020) utilizan reglas de asociación, agrupamiento, clasificación basada en asociaciones a través de análisis avanzados que emplean técnicas de NLP y minería de datos. El sistema utiliza análisis de datos federados, descubrimiento de conocimiento, recuperación de información y nuevas técnicas para manejar la web semántica y la representación de gráficos de conocimiento. El proceso de procesamiento integra datos de múltiples fuentes virtualmente mediante la creación de bases de datos virtuales.

Finalmente, la tarea principal en los sistemas basados en estadísticas es la medición de la similitud semántica entre los términos extraídos de los datos no estructurados y las instancias de la ontología. Sin embargo, este enfoque no es apropiado para todas las situaciones. Por ejemplo, no puede tratar términos no incluidos en los diccionarios de

sinónimos. De manera similar, este enfoque no es adecuado para comprender las abreviaturas. En (Achichi et al., 2015), se propone un algoritmo k-mean para la identificación de temas y la clasificación de datos. El algoritmo compara y obtiene entidades nombradas a partir de datos no estructurados para luego transformarlos en Linked Open Data (LOD) enriquecidos con DBpedia URI.

## 2.3 Análisis de datos masivos semántico

El Análisis de Datos (DA, del inglés *Data Analysis*) es la ciencia que se encarga de presentar información precisa, confiable y oportuna que suponga conclusiones y apoye a la toma de decisiones a través de procesos como la inspección, limpieza, transformación y modelado de datos (Bihani & Patil, 2014). Es importante mencionar que previo al análisis de los datos es indispensable el desarrollo de las fases de integración, visualización y difusión de datos. El análisis de datos he tenido un gran auge en los últimos años por sus números usos y áreas de aplicación como como, por ejemplo, negocios (Airinei & Berta, 2012), ciencias sociales (Kim et al., 2017), o educación (Bello-Orgaz et al., 2016), entre otros. Su extensa área de aplicación ha beneficiado e impulsado el desarrollo de técnicas y estrategias específicas de DA entre las que destacan la minería de datos (*Data mining,* por su popular término en inglés), orientada en predecir y descubrir nuevo conocimiento (Witten et al., 2005), y la inteligencia de negocio (BI, por sus siglas en inglés como *Business Intelligence*) (Chaudhuri et al., 2011), enfocada en la gestión empresarial y la integración de valor con la interpretación de datos para la toma de decisiones.

La arquitectura tecnológica de un sistema de análisis de datos está compuesta, tradicionalmente, por al menos tres componentes principales (Gómez Vieites & Suárez Rey, 2011): (i) herramientas para la implementación de procesos de extracción, transformación, y carga de datos, abreviadas como ETL, por sus siglas en inglés como *Extract, Transform, and Load*, (ii) un repositorio de almacenamiento e integración de datos y (iii) aplicaciones de exploración y visualización de datos, de las cuales podemos mencionar los sistemas de generación de informes, las técnicas de análisis multidimensional y procesamiento analítico (OLAP, por sus siglas en inglés de *Online*

*Analytical Processing*), y las herramientas de minería de datos que utilizan diversas técnicas de estadística, inteligencia artificial y simbólica.

El análisis de datos se divide comúnmente en tres etapas o procesos como lo son: (Molina López & García Herrero, 2006): (1) preprocesado de datos (adquisición, organización y almacenamiento de datos), (2) procesado de datos (aplicación de técnicas tradicionales de análisis de datos) y por último, (3) visualización de datos (presentación de los resultados del análisis). Estos procesos implican algunos retos y dificultades propias de las estructuras de datos implementadas en la actualidad, como lo son (Chaudhuri et al., 2011): (i) la heterogeneidad de las fuentes de datos, atribuido a la propiedad variable, para su integración, limpieza y estandarización del ETL, lo que atribuye retos en implementar análisis de datos en tiempo real, (ii) la inviabilidad de integrar datos no estructurados o semiestructurados en repositorios con sistemas de gestión de bases de datos relacionales (DBMS, por sus siglas en inglés de *Data Base Management System*), y (iii) la obligatoriedad de implementación de servidores OLAP para la exposición multidimensional de los datos.

En la actualidad, la creciente demanda de uso de Internet, y en específico de redes sociales y sitios interactivos por parte de los usuarios, ha producido una gran disponibilidad e importancia de los datos masivos provenientes de fuentes heterogéneas, los procesos tradicionales de análisis de datos se enfrentan a enormes retos casi imposibles de sobrellevar con sus típicas implementaciones de sistemas relacionales, exhibiendo una necesidad de ofrecer una nueva arquitectura que brinde soporte a grandes volúmenes de datos (Sivarajah et al., 2017). De esta manera surge una nueva estrategia y colección de tecnologías definida dentro de los datos masivos.

Los datos masivos, comúnmente mencionados por su término en inglés como *Big data*, es un enfoque novedoso para el análisis de datos. Caracterizado principalmente por tres propiedades propias de los datos, a saber, volumen (es decir, tamaño en bytes), velocidad (es decir, tasa de crecimiento de datos) y variedad (es decir, formato de datos y heterogeneidad de la fuente de datos), que se conocen como las 3V de los datos masivos (Laney, 2001). Otras Vs o propiedades típicamente asociadas con los datos masivos son el

valor, la veracidad, la volatilidad, la validez y la viabilidad (Maté Jiménez, 2014). Gartner proporciona una definición más formal de datos masivos y describe como "activos de información caracterizados por su alto volumen, alta velocidad y gran variedad, que exigen soluciones de procesamiento innovadoras y eficientes para la mejora del conocimiento y la toma de decisiones. hacer en las organizaciones" (Beyer & Laney, 2012). Dadas estas propiedades, se deben considerar una serie de desafíos al tratar con los datos masivos en las implementaciones de análisis de datos, desde la extracción y vinculación de datos heterogéneos provenientes de diversas fuentes, hasta el análisis, organización, modelado y visualización del conocimiento (Mayer-Schönberger & Cukier, 2013) (Kale & Dandge, 2014).

Las tecnologías asociadas a la Web Semántica (Shadbolt et al., 2006) y datos enlazados (*Linked Data*) (Bizer et al., 2009) se han mostrado eficaces para el tratamiento automático de la información en diferentes contextos (Di Iorio & Rossi, 2018; Rahoman & Ichise, 2018). Los modelos ontológicos subyacentes, que se basan en formalismos lógicos, permiten que los sistemas informáticos interpreten de alguna manera la información que se gestiona (Studer et al., 1998). También permiten llevar a cabo procesos avanzados de razonamiento e inferencia (Rodríguez-García & Hoehndorf, 2018). La comunidad científica dentro de estos campos de investigación ha desarrollado herramientas que hacen uso de tecnologías semánticas para (i) integrar datos de fuentes de datos heterogéneas (García-Sánchez et al., 2008)(Santipantakis et al., 2017) y (ii) permitir el análisis de grandes cantidades de datos a nivel de conocimiento (Neuböck et al., 2014), con diferentes grados de éxito.

Las principales dificultades asociadas con la gestión de datos masivos están vinculadas a su colección y almacenamiento, búsqueda, intercambio, análisis y visualización (Kale & Dandge, 2014). La web semántica proporciona los medios para superar algunos de estos desafíos. Ciertamente, las tecnologías web semánticas permiten el procesamiento automatizado de datos a través de una inferencia sofisticada y las técnicas de razonamiento. RDF como modelo estándar para el intercambio de datos en la Web y basada en grafos, lo que permite la representación de datos en forma de triples sujeto-predicado-objeto (W3C, 2014). Los triples RDF se pueden usar para crear conjuntos de datos y establecer relaciones

explícitas entre los datos. Esta colección de conjuntos de datos interrelacionados en la Web se conoce como datos enlazados (Bizer et al., 2009). Uno de los objetivos principales de los datos enlazados es agregar una capa semántica sobre los datos, lo que lo hace comprensible por las máquinas para que puedan realizar algunas operaciones de análisis de datos en nombre de los usuarios humanos (Hu et al., 2013). Por lo tanto, la web semántica puede ayudar en el descubrimiento, la integración, la representación y la gestión del conocimiento (Barceló Valenzuela et al., 2006). En particular, las tecnologías semánticas se han aplicado con éxito en una serie de escenarios para la integración de datos heterogéneos (Konys, 2016), análisis de datos a nivel de conocimiento (Barceló Valenzuela et al., 2006) y visualización de datos vinculados (Brunetti et al., 2013).

En los últimos años, una gran cantidad de documentos de investigación publicados han explorado los beneficios en el uso de tecnologías semánticas en el análisis de datos y los datos masivos (Airinei & Berta, 2012; Bikakis & Sellis, 2016; Konys, 2016; Kureychik & Semenova, 2017; Neuböck et al., 2014; Nuzzolese et al., 2017; Wongthontham & Abu-Salih, 2018). El impacto de la semántica en este campo cubre todo el proceso, desde el preprocesamiento (adquisición de datos y la organización) a la visualización a través del procesamiento y análisis de los datos. Los modelos de ontología se pueden usar para armonizar datos heterogéneos de fuentes estructuradas, semiestructuradas y no estructuradas, lo que permite su almacenamiento integrado en los repositorios de ontología (Konys, 2016). El escenario intrínseco de la Web semántica permite que los procesos de razonamiento infieran en nuevos conocimientos que no se indique explícitamente en los datos de origen (Kureychik & Semenova, 2017). Finalmente, se han propuesto muchas herramientas que proporcionen características de recuperación y visualización para datos enlazados y datos masivos enlazados (Bikakis & Sellis, 2016)(Nuzzolese et al., 2017).

Las ontologías facilitan la gestión de la información a nivel de conocimiento y permiten un acceso integrado a fuentes heterogéneas, pudiendo utilizar diversas técnicas de inferencia y razonamiento para analizar los datos (Neuböck et al., 2014). Sin embargo, se ha demostrado que el análisis de grandes volúmenes de datos tiene implicaciones para la gestión del conocimiento (Chan, 2014). Para manejar una cantidad tan grande de datos, se han propuesto diferentes soluciones (Airinei & Berta, 2012; Bao et al., 2016; Bennett &

Baclawski, 2017; Eine et al., 2017; Kupershmidt et al., 2010; Neuböck et al., 2014). Una posibilidad es el uso de los conceptos y métodos de Análisis Formal de Conceptos (FCA, del inglés *Formal Concept Analysis*), que es un método donde los datos se estructuran en unidades que representan abstracciones formales de conceptos del pensamiento humano, lo que permite una interpretación integral y facilita la representación del conocimiento y el manejo de la información. (Airinei & Berta, 2012; Neuböck et al., 2014). Otro enfoque relevante en este contexto es la generación de una capa de ontología multidimensional para consolidar el análisis en diferentes niveles del repositorio, de manera de agregar más significado a los datos, eliminar redundancias e información irrelevante y obtener un análisis mejorado (Neuböck et al., 2014).

Los sistemas OLAP también pueden beneficiarse del uso de ontologías (Kupershmidt et al., 2010). El modelo ontológico, que proporciona una definición inequívoca de los términos de dominio asociados a las necesidades específicas de OLAP, fortalece la correlación y el enriquecimiento de los datos automáticamente con algoritmos de agrupación. También proporciona los medios para realizar tareas de análisis comparativo de datos muy heterogéneos provenientes de diferentes fuentes, plataformas y tecnologías. En (Bao et al., 2016), los autores sugieren el uso de reglas expresadas en SWRL (Lenguaje de Reglas de la Web Semántica) junto con los mecanismos de inferencia intrínsecos de las ontologías para analizar las bases de conocimiento produciendo así nuevo conocimiento relevante.

Las deficiencias potenciales que pueden surgir de la aplicación de la gestión de datos masivos basada en ontologías se han explorado en (Eine et al., 2017). La complejidad del proceso de razonamiento y los problemas de rendimiento relacionados se pueden abordar considerando lenguajes ontológicos que cambian la expresividad por una complejidad de razonamiento reducida. Sin embargo, los beneficios son enormes. Se facilita la integración de nuevas fuentes de datos y se potencia la usabilidad. Tanto la reutilización de los datos como la mantenibilidad de las aplicaciones también se pueden mejorar con el uso de la semántica. La visualización del conocimiento también se ve afectada, ya que es posible representar la información en diferentes formas, mejorando así la comprensión de los usuarios finales de los conocimientos derivados de la base de conocimiento.

Una de las propuesta para la integración de técnicas de análisis de datos masivos con tecnologías de la Web semántica para ayudar en la definición de estrategias de marketing político a través de la recopilación y el análisis de datos electorales disponibles en Internet fue descrita en (Guedea-Noriega & García-Sánchez, 2018). El marco de trabajo o propuesta fue capaz de extraer datos de varias fuentes, organizar los datos recopilados y crear relaciones semánticas entre los elementos de datos. Como resultado de dicha propuesta fue un repositorio de conocimiento basado en ontología el cual permitió un análisis más sofisticado y preciso que puede conducir a una mejor retroalimentación para mejorar el mensaje político y la estrategia de comunicación. Por último, en (Hector H. Guedea Noriega & Garcia Sanchez, 2019) se realizó una extensa revisión de la literatura sobre análisis semántico de datos masivos con resultados puntuales para cada uno de sus procesos.

## 2.4 Objetivos de la tesis doctoral

### 2.4.1 Motivación

Las campañas electorales modernas enfatizan en la viralización de sus candidatos y la participación activa de los votantes a través de las redes sociales y sitios en Internet en general. Sin embargo, pese a la contribución positiva de los nuevos medios sociales para el proselitismo político y el activismo digital (Héctor Hiram Guedea Noriega et al., 2011), los cuales como desventaja producen cientos de datos imposibles de asimilar por humanos, nace la necesidad de homogenizar, reutilizar y compartir el conocimiento colectivo para que sea accesible para las máquinas, y con ello lograr automatizar y mejorar la eficiencia de los procesos de análisis de datos y toma decisiones basada en la inteligencia de mercados.

La motivación de esta tesis doctoral surge, precisamente, para generar soluciones y nuevos recursos para superar los retos actuales que presenta el marketing político.

### 2.4.2 Objetivos

El objetivo general de esta tesis doctoral es construir un modelo ontológico y sistema de población automática para un grafo de conocimiento a partir de datos masivos sociales del

dominio de marketing político. Para lograr lo antes mencionado es necesario cumplir con los siguientes objetivos específicos:

- Diseño y construcción de un modelo ontológico del dominio de marketing político, un modelo conceptual compartible que involucra los elementos más relevantes y principales del conocimiento operativo de las campañas políticas: candidato, electorado y medios.
- Implementación de técnicas de extracción de información a partir de fuentes de datos no estructurados, semiestructurados y estructurados.
- El enriquecimiento y análisis de datos a nivel semántico y lingüístico a través de técnicas de PNL y ML en textos en español
- Población automática de un grafo de conocimiento para homogenizar, recopilar y relacionar los datos de los conceptos de marketing político.
- Validación del grafo de conocimiento resultante basado en la calidad de sus dimensiones.

### 2.4.3 Metodología

Esta tesis doctoral se desarrolló mediante cuatro fases principales: (1) estudio del estado del arte, (2) diseño y construcción de un modelo ontológico, (3) desarrollo de un sistema de población automática de un grafo de conocimiento y, por último, (4) validación del grafo de conocimiento resultante.

1. Estudio de las definiciones y conceptos de marketing político, retos del marketing político en las campañas electorales modernas, estado actual de las tecnologías semánticas y análisis de datos, y finalmente, beneficios de implementar tecnologías de web semántica sobre los actuales retos del marketing político.
2. Diseño y construcción de un modelo ontológico que responde los principales cuestionamientos del marketing político sobre las campañas electorales.

3. Desarrollo de un sistema de población automática de un grafo de conocimiento sobre el modelo ontológico de marketing político.
4. Validación y/o evaluación del grafo de conocimiento resultante con un reporte basado en los requisitos de calidad de sus dimensiones.

# Capítulo 3. Modelo semántico de conocimiento sobre marketing político

En esta sección se describe el proceso para la construcción de una ontología para el dominio del marketing político. Este proceso se basa en la metodología descrita como *Ontology101* (N. F. Noy & Mcguinness, 2001), que incluye las siguientes etapas: (1) determinar dominio y alcance de la ontología, (2) considerar la reutilización de ontologías existentes, (3) enumerar términos importantes para la ontología, (4) definir las clases y la jerarquía de clases, (5) definir las propiedades de las clases (atributos y relaciones), (6) definir las características de los atributos y relaciones, y (7) crear instancias.

El objetivo más importante es recopilar conocimiento relacionado con el marketing político en forma de ontología es poder analizar posteriormente el conocimiento disponible para obtener información útil para la definición adecuada de estrategias políticas. El diseño adecuado de esta ontología brindará las respuestas necesarias para mejorar la toma de decisiones en este escenario, obteniendo cuatro ítems de conocimiento operativo: cómo es el candidato, cómo son los adversarios, cómo son los votantes y cómo es la elección en general.

Nuestra propuesta de ontología para el marketing político la nombramos PMont (por sus siglas en inglés de *Political Marketing Ontology*) (Guedea-Noriega & García-Sánchez, 2020) desarrollada siguiendo la metodología de desarrollo de ontologías antes mencionada, elaborada en OWL2 (https://www.w3.org/TR/owl2-overview/) usando la herramienta Protégé (https://protege.stanford.edu/). PMont fue concebido para servir como modelo ontológico esencial para nuestro sistema basado en el conocimiento y para responder las principales preguntas de marketing político típicas en una campaña electoral.

A continuación, se describe su desarrollo iniciando por establecer el dominio de interés, ontologías a reutilizar, listado de términos relevantes, jerarquía de clases, propiedades de las clases, características de los atributos y relaciones, y por último, la instanciación.

## 3.1 Dominio y alcance de la ontología

En este apartado se define el dominio de la ontología, se establecen los objetivos principales y específicos, así como también el alcance con respecto a los cuestionamientos sobre el área de interés a abordar.

1. ***¿Cuál es nuestro dominio de interés?***

Marketing político.

2. ***¿Cuáles son los objetivos de la ontología?***

La Tabla 2 ofrece las respuestas a este cuestionamiento, la table ha sido dividida en dos columnas, la primera, describe el objetivo principal, y la segunda, define los objetivos específicos.

| ***Objetivo principal*** | ***Objetivos específicos*** |
|---|---|
| Homogenizar, reutilizar y compartir información producida por el electorado proveniente de fuentes de datos no estructurados, semi-estructurados y estructurados desde los medios digitales (redes sociales y páginas Web). | • Diseñar un modelo ontológico para el marketing político.<br>• Facilitar el descubrimiento de nuevo conocimiento del electorado.<br>• Ayudar a inferir tomas de decisiones.<br>• Generar un grafo de conocimiento para la representación semántica del marketing político. |

**Tabla 2 Objetivos de la ontología de dominio de Marketing Político (PMont)**

3. ***¿Para qué tipos de preguntas la información en la ontología debería proveer respuestas?***

En la Tabla 3 se establecen las siguientes preguntas que representan el alcance de la ontología; sus respuestas equivalen al diseño de sus entidades, relaciones y propiedades.

| | | |
|---|---|---|
| P1. ¿Qué demanda el electorado?<br><br>P2. ¿Qué dice la opinión pública sobre el candidato (imagen, propuestas)?<br><br>P3. ¿Qué dice la opinión pública sobre el partido político al que pertenece el candidato?<br><br>P4. ¿Qué tipo de mensaje y lenguaje político debe establecer el candidato para impactar positivamente en la ciudadanía? | P5. ¿Qué tipo de propuestas de campaña debe diseñar?<br><br>P6. ¿Cuáles son los grupos objetivo del electorado?<br><br>P7. ¿En qué grupo tiene mayor aceptación y en cuál no?<br><br>P8. ¿Qué comparativa tiene el candidato con sus adversarios? | P9. ¿En qué medios el candidato tiene mayor impacto?<br><br>P10. ¿Qué medios son parciales (identificar candidato/partido de su gusto) y qué medios son imparciales?<br><br>P11. ¿Qué parte del electorado tiene decidido su voto (identificar votantes individuales y su candidato/partido elegido) y qué parte del electorado está indeciso (identificar individuos)? |

**Tabla 3 Alcance de la ontología PMont**

4. *¿Quién usará y mantendrá la ontología?*

El usuario principal es el consultor político del candidato. El mantenimiento de la ontología correrá a cargo de un equipo de desarrollo especializado.

## 3.2 Ontologías a reutilizar

Basados en el estado del arte, podemos afirmar que en la actualidad no existe una ontología o modelo ontológico para el marketing político, el cual se adapte a las necesidades y objetivos planteados en las campañas electorales. Sin embargo, existen una gran variedad de ontologías en la Web con estructuras generales que podrían fortalecer la propuesta de nuestro modelo ontológico político. Una de ellas es la ontología basada en tiempo llamada *OWL-Time* (https://www.w3.org/TR/owl-time/), contiene un vocabulario extenso sobre relaciones, conceptos, propiedades sobre dicho dominio. Su implementación será para expresar todo lo relacionado con el contexto tiempo, por ejemplo, periodo de las campañas políticas, fechas de publicación de opiniones del electorado, fechas de nacimiento de los individuos, duración de impacto de cierta acción del candidato, entre otras más (véase Figura 18).

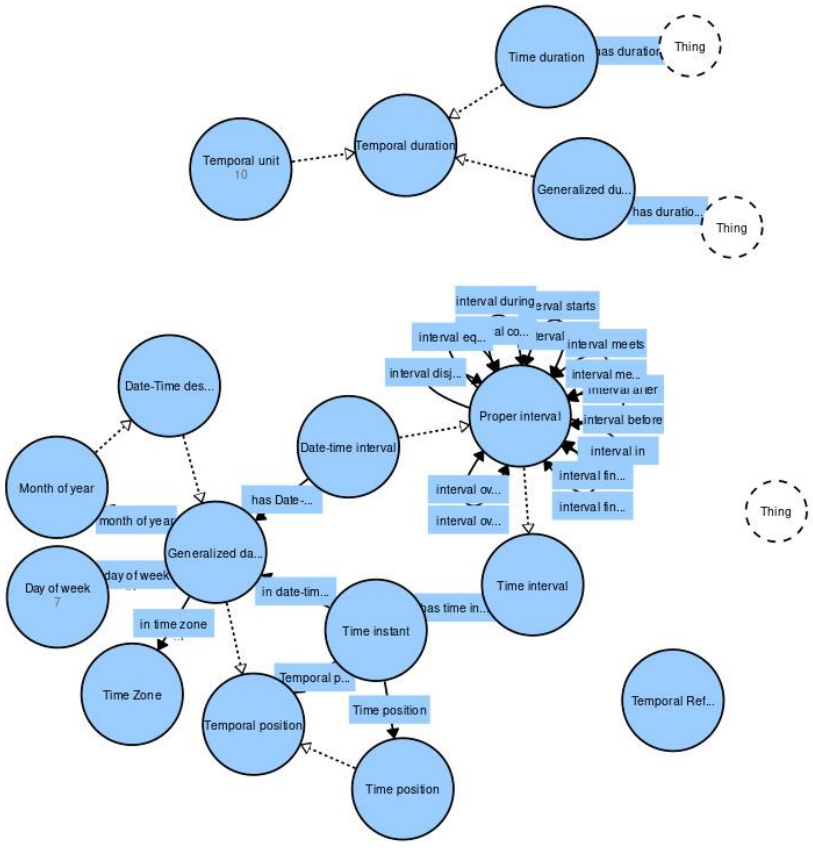



**Figura 18 Clases principales de OWL Time, visualización desde WebVOWL**

Otra ontología de gran utilidad es *Place* (http://dbpedia.org/ontology/Place), la cual describe y expresa conceptos y propiedades del dominio de información geográfica, esto ayudará a contextualizar la geolocalización de las publicaciones, candidatos y áreas de desarrollo de las campañas electorales. Por ejemplo, con los elementos de esta ontología se podrá geolocalizar al electorado y/o ubicación de impacto la propuesta política (véase Figura 19).

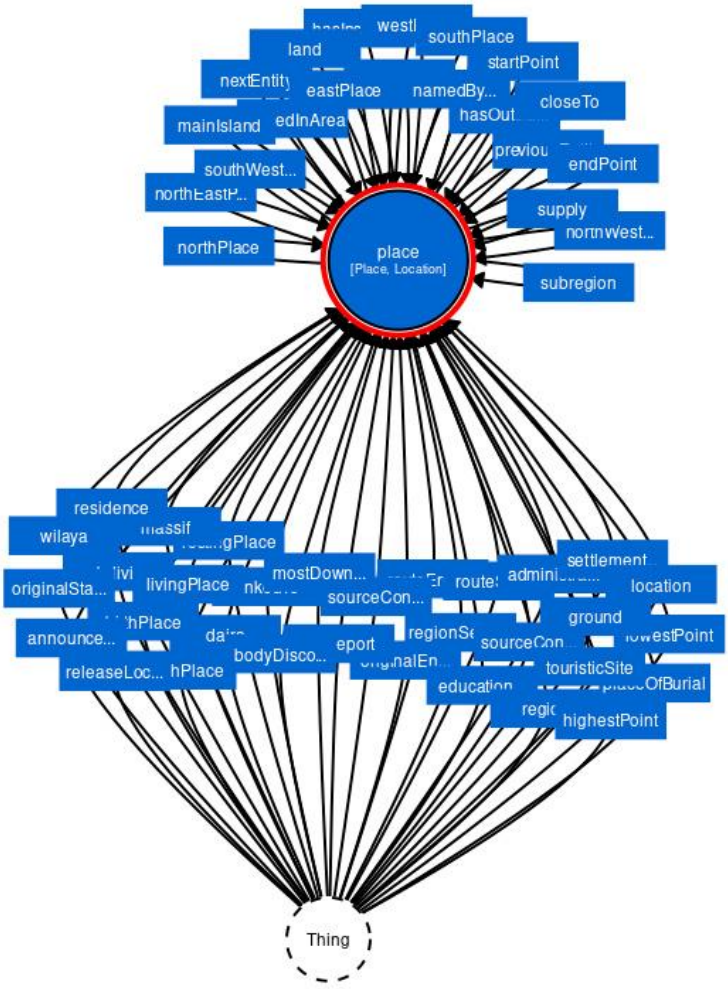



**Figura 19 Grafo de clases principales de Place Ontology**

Por último, la ontología FOAF (http://xmlns.com/foaf/spec/) diseñada para vincular personas e información a través de la Web. Será de utilidad para enriquecer con metadatos y generar un vocabulario adecuado para la definición del electorado, sus intereses y actividades, concibiendo perfiles con identificadores únicos para ser implementados en las diversas relaciones dentro de la ontología (véase Figura 20).

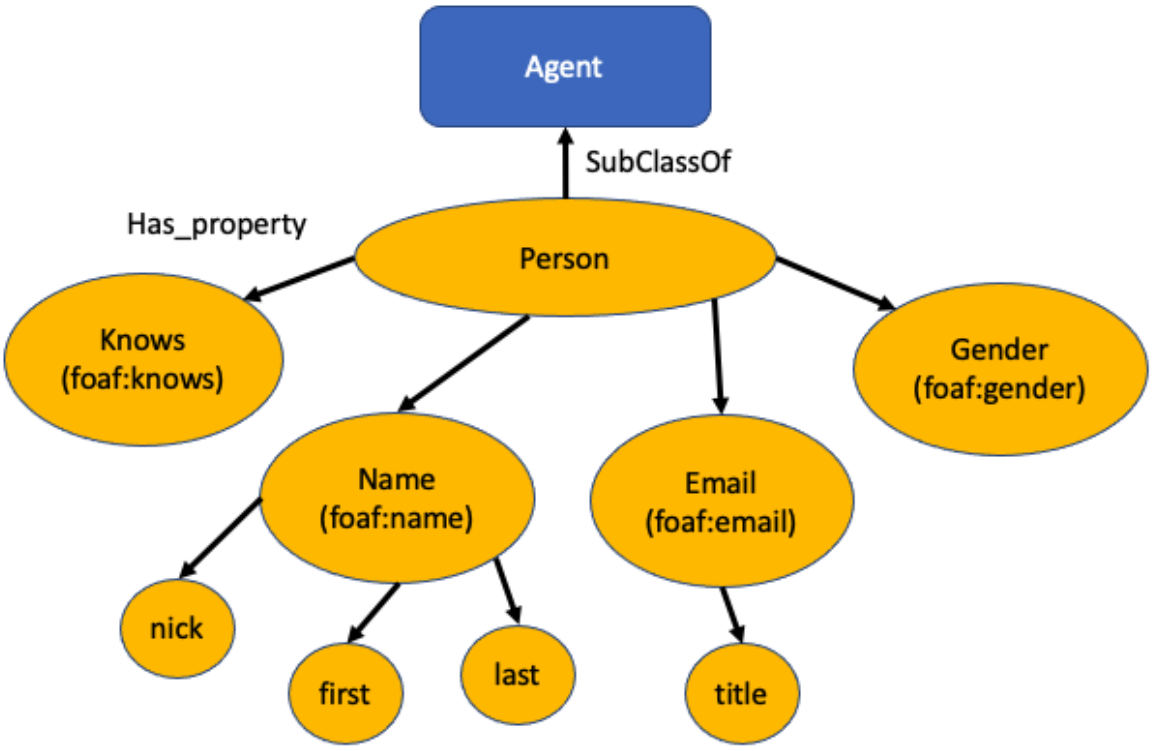



**Figura 20 Fragmento de la ontología FOAF**

## 3.3 Listado de términos relevantes

Los términos relevantes extraídos a partir de las preguntas planteadas en la sección 3.1 son los siguientes: campaña (`campaign`), candidato (`candidate`), partido político (`PoliticalParty`), electorado (`ElegibleVoter`), opinión (`opinion`), publicación (`publication`), propuesta (`proposal`), medios de comunicación (`massmedia`) y fuente de datos (`source`).

## 3.4 Jerarquía de clases

Para el diseño de la jerarquía de clases se estableció el método *top-down* por su enfoque en primero identificar los conceptos genéricos y posteriormente la especialización de sub-clases. En la Tabla 4 se muestran las clases y superclases definidas de acuerdo con las preguntas planteadas en la sección 3.1.

| | | |
|---|---|---|
| • Campaign<br>• CandidateCampaign<br>• Organization<br>  o Political Party<br>  o Mass Media<br>• Person<br>  o Elegible voter<br>  ▪ Candidate<br>  o Not elegible voter | • Publication<br>  o Tipos de publicación (opinion, proposal y news).<br>  o Características de la publicación (polarity, degree of sentiment y topic).<br>• Reputation | • Trustworthiness<br>• Ideology<br>• Geographic impact<br>• Fuentes y formatos<br>  o Structured, semistructured y not structured |

**Tabla 4 Clases y sub-clases de la ontología PMont (parcialmente en inglés)**

De acuerdo con los conceptos principales enumerados anteriormente, el proceso de desarrollo de la ontología se inició con un enfoque de arriba hacia abajo para presentar y definir la jerarquía de clases, subclases y propiedades de los objetos más fácilmente. La ontología ha sido definida en OWL2 (W3C, 2012) usando Protégé. Contiene siete conceptos principales: `Campaign`, `Candidate`, `CandidateCampaign`, `Organization`, `Person`, `Publication` y `Source`. También existen otros conceptos relacionados para cubrir los fines sugeridos en el marketing político. Estos conceptos principales impactan directamente en las preguntas de investigación y se pueden definir de la siguiente manera:

- `Campaign`: esta clase define las fechas de la campaña y elección política. Está relacionado con partidos políticos y candidatos.

- `Candidate`: representa a una persona afiliada a un `Political Party`, con la intención de ser elegida para un cargo público.

- `CandidateCampaign`: esta clase se utiliza para definir relaciones ternarias entre las clases `Campaign`, `Candidate`, y `Political Party`.

- `Organization`: es una superclase que contiene dos subclases, a saber, `Mass Media` y `Political Party`, con el fin de definir estructuras y grupos de personas.

- `Person`: es una superclase que contiene dos subclases, a saber, `EligibleVoter` y `NotEligibleVoter`. `EligibleVoter` es la persona que puede votar por un candidato en una elección política, mientras que `NotEligibleVoter` no puede votar en las circunstancias exigidas por la ley donde se encuentra la elección (por ejemplo, menor de edad).

- `Publication`: representa diferentes tipos de publicaciones. En particular, se consideran tres subclases, a saber, `NewsItem` (noticia creada y compartida por medios masivos), `Proposal` (una propuesta política de un candidato) y `Opinion` (opinión pública de un ciudadano compartida en redes sociales).

- `Source`: indica la fuente de las publicaciones. Es una superclase que contiene `Structured`, `Semistructured` y `Unstructured` como sub-clases.

## 3.5 Propiedades de las clases: atributos y relaciones (slots)

Como se sugirió en la metodología de desarrollo de ontologías mencionada anteriormente, también se definieron las propiedades de las clases o slots, las cuales tienen como objetivo relacionar las diferentes clases y subclases y agregar datos para la inferencia de conocimiento. La Figura 21 y Tabla 5 muestra las clases de alto nivel y sus relaciones.

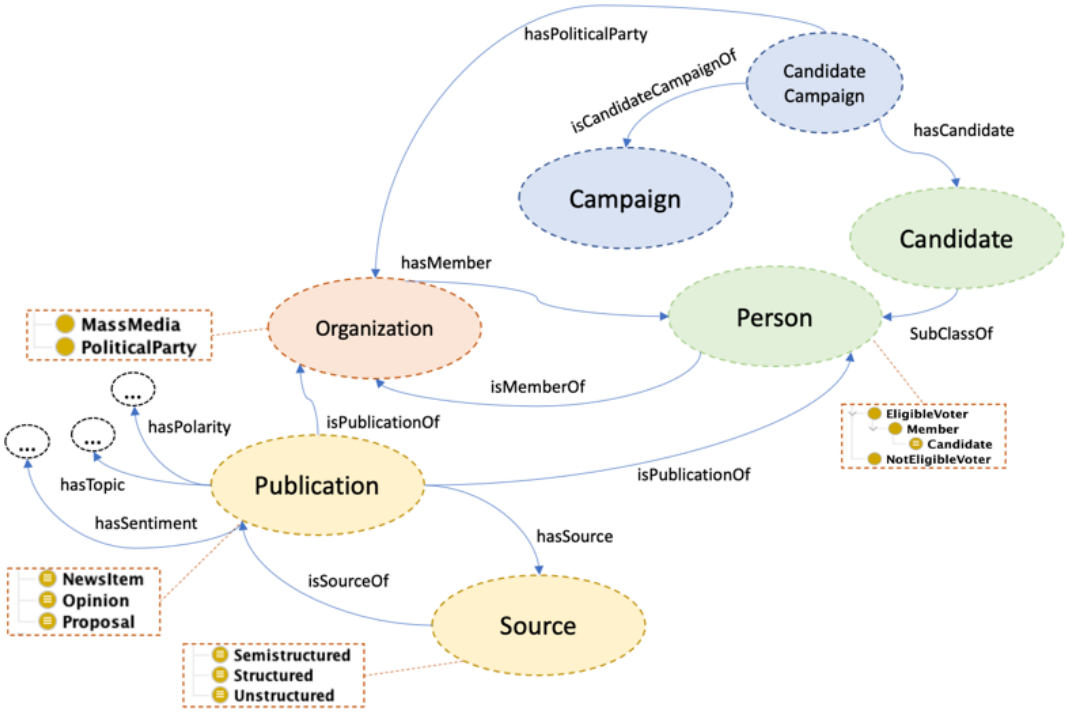



**Figura 21 Las clases de alto nivel de la ontología y sus relaciones**

| **Clase** | **Atributo** | **Relación** |
|---|---|---|
| *Campaign* | *hasDateStartCampaign*, *hasDateEndCampaign*, *hasTitle*, *hasDescription* | *hasCandidateCampaign*, *hasGeographicalImpact*, *hasPlace* |
| *Organization*<br>*Person* | *hasName*, *hasDescription*, *knownAs*, *hasWebUrl*, *hasSocialMediaUrl*, *hasImageUrl* | *hasIdeology*, *hasPublication*, *hasPlace*, *hasReputation*, *hasTrustworthiness* |

| | | |
|---|---|---|
| *Publication* | *hasTitle*, *hasContent*, *hasDatePublication*, *hasImageUrl* | *hasSource*, *hasPolarity*, *hasTopic*, *hasSentiment*, *hasGeographicalImpact* |
| Source | *hasWebUrl*, *hasPriority* | *hasSourceFormat* |

**Tabla 5 Propiedades destacadas de las clases**

## 3.6 Características de los atributos y relaciones

A las propiedades definidas en la ontología se les incluyó una serie de características de valores de cardinalidad (máximos y mínimos), valores por defecto, rango de valores o valores admitidos, con el objetivo de que el modelo ontológico tenga restricciones y limitaciones para evitar duplicidades, ambigüedades y daros incompletos. Por ejemplo, `Organization` y `Person` pueden tener publicaciones de diferentes tipos (`Opinion`, `NewsItem` y `Proposal`) por lo que se establece una cardinalidad de *1* a *n* a la relación `hasPublication`. Cada instancia de `Publication` tiene una y solo una polaridad, sentimiento y temática, por lo que se establece la cardinalidad exacta de *1* a las relaciones `hasPolarity`, `hasSentiment` y `hasTopic`.

## 3.7 Instancias

Para ciertas clases es necesario la previa población de instancias con valores que formen parte de sus atributos de clase, como es el caso de los mencionados en la Tabla 6.

Por ejemplo, `PublicationPolarity` es una clase que almacena tres instancias: *negative*, *neutral* y *positive*, la selección de una de las tres instancias como atributo de una clase de `publication`, `opinion` o `proposal` se efectúa durante el proceso de población del grafo de conocimiento por medio de técnicas de NLP y ML, la cual se describirá en el Capítulo 4 de esta tesis doctoral.

| Clase | Instancia(s) |
|---|---|
| `GeographicalImpact` | *Local*, *National*, *International* |
| `Ideology` | *Left-wing*, *Centre*, *Right-wing* |
| `PublicationPolarity` | *Negative*, *Neutral*, *Positive* |
| `PublicationSentiment` | *Angry*, *Happy*, *Sad*, etc. |
| `PublicationTopic` | *Corruption*, *Culture_and_Sport*, *Economy*, *Education*, *Government*, *Health*, *Security* |
| *Trustworthiness* | *Fake*, *Authentic* |

**Tabla 6 Listado de instancias de base**

## 3.8 PMont (Political Marketing Ontology)

En este apartado se describe a detalle la jerarquía de clases, propiedades de objetos y datos más importantes de la ontología PMont. Se utilizó Protégé como herramienta de visualización de los elementos de la ontología. Por último, se muestran las métricas finales de la ontología.

### 3.8.1 Clases

El número total de clases creadas para PMont fueron 29, de las cuales `Organization`, `Person`, `Publication`, `Source` y `SourceFormat` forman parte de superclases que alojan subclases para favorecer la contextualización y las relaciones a detalle dentro de la ontología (véase Figura 22).

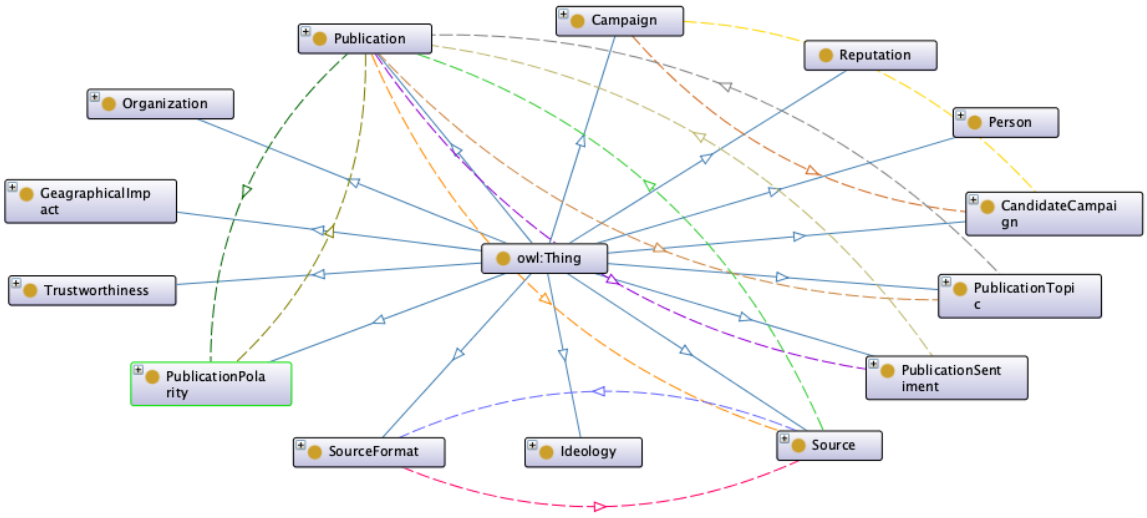



**Figura 22 Jerarquía de clases de PMont**

Por lo tanto, se definieron cinco superclases principales con subclases específicas de la siguiente forma:

- Para `Organization` hemos creado dos sub-clases `MassMedia` y `PoliticalParty`, de las cuales el objetivo es reunir a los medios de comunicación y partidos políticos, respectivamente (véase Figura 23).


**Figura 23 Superclase Organization de PMont**

- Para `Person`, aloja tanto a posibles votantes, candidatos y no elegibles votantes (menores de 18 años), de la siguiente forma (véase Figura 24)


**Figura 24 Superclase Person de PMont**

- Para `Publication`, definimos tres tipos de publicaciones, `NewsItem`, `Opinion` y `Proposal` (véase Figura 25) y, a su vez, se especificaron las reglas de relación entre clases. Por ejemplo, para `Newsitem` es exclusivo de `MassMedia`, `Opinion` es equivalente para la superclase `Person`, y `Proposal` es exclusivo para `Candidate` o `PoliticalParty` (véase Tabla 7).


**Figura 25 Superclase de Publication de PMont**

| | |
|---|---|
| `NewsItem` | `Publication and (isPublicationOf some MassMedia)` |
| `Opinion` | `Publication and (isPublicationOf some Person)` |
| `Proposal` | `(Publication and (isPublicationOf some PoliticalParty)) or (isPublicationOf some Candidate)` |

**Tabla 7 Regla de relación entre subclases de Publication**

- La diferencia entre `Source` y `SourceFormat` es la jerarquía de la fuente de datos, mientras que `Source` establece la fuente de datos en un contexto de aplicación o herramienta (Webpage, Twitter, etcétera), `SourceFormat` describe los tipos de estructuras de datos (JSON, base de datos relacional MySQL, etcétera) (véase Figura 26).

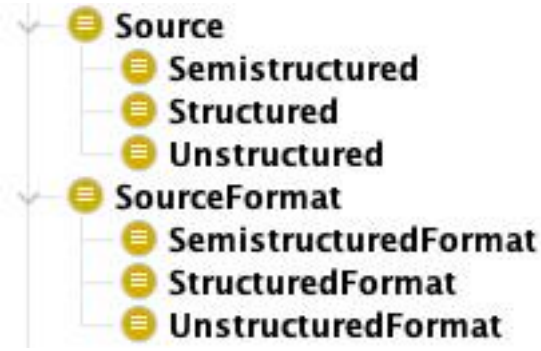



**Figura 26 Superclases de Source y SourceFormat de PMont**

### 3.8.2 Propiedades de objeto (relaciones)

Las propiedades de objeto establecen las relaciones y reglas entre las clases y las instancias, para el caso de PMont definimos un total de 30 (véase Figura 27), de las cuales podemos describir las más importantes:

- `hasPublication`: como se mencionó anteriormente en el apartado 3.6 se utiliza para crear la relación entre `Organization`, Person y `Candidate` con respecto a sus tres tipos de publicaciones `Opinion`, `Proposal` y `NewsItem`.
- `hasPolarity`: más adelante se describirá la implementación de NLP sobre los textos en castellano, básicamente estas propiedades de objeto relacionan el texto con el resultado otorgado por el análisis de sentimiento. La relación se realiza entre instancias, instancia dentro de la superclase `Publication` y `PublicationPolarity`.

- `hasTopic`: al igual que la anterior propiedad, se aplica NLP para descubrir la temática del texto, posterior se asigna a una instancia definida dentro de la clase `PublicationTopic`.
- `hasSource` y `hasSourceFormat`: de acuerdo con la estructura de datos y tipo de fuente se asigna la instancia de `Publication` a una instancia previamente definida dentro de las clases `Source` y `SourceFormat`.
- `hasCandidateCampaign`: asigna una relación entre la campaña política (`Campaign`) y un candidato (`Candidate`), con ello se establece un rango de tiempo y espacio.


**Figura 27 Propiedades de objeto de PMont**

### 3.8.3 Propiedades de datos

En este apartado se muestran las propiedades de datos definidas dentro de PMont con el objetivo de proveer de atributos y enriquecer las instancias.

Las propiedades de datos fueron explícitamente definidas, esto es, los nombres de los atributos contextualizan su valor, como: `hasContent`, `DateStartCampaign`, `hasDateEndCampaign`, `hasSocialMediaUrl`, etcétera (véase Figura 28).


**Figura 28 Propiedades de datos de PMont**

### 3.8.4 Métricas de la ontología

Las métricas obtenidas por la herramienta Protégé después de haber concluido el diseño y desarrollo de la ontología PMont en general, se muestran en la Figura 29.

Estas métricas involucran los axiomas y sus declaraciones, cantidad de clases y subclases, propiedades de objetos y datos, así como también cantidad de instancias definidas actualmente para su funcionamiento.

**Metrics**

| | |
|---|---|
| Axiom | 555 |
| Logical axiom count | 304 |
| Declaration axioms count | 142 |
| Class count | 29 |
| Object property count | 30 |
| Data property count | 17 |
| Individual count | 62 |
| Annotation Property count | 7 |

**Figura 29 Métricas de PMont**

## 3.9 Validación y repositorio de la ontología

Al término del diseño y desarrollo de la ontología PMont, se llevó a cabo la validación y evaluación del modelo ontológico a través de la herramienta OOPS! OntOlogy Pitfall Scanner (http://oops.linkeddata.es).

El resultado fue positivo sin errores mínimos o críticos en todas las evaluaciones divididas en estructura (modelo, relaciones, inferencia), funcionalidad, usabilidad y consistencia (véase Figura 30). El resultado del proceso de validación/evaluación está disponible en https://hectorguedea.github.io/pmont/OnToology/Political-Marketing-Ontology.owl/evaluacion/oops.html.

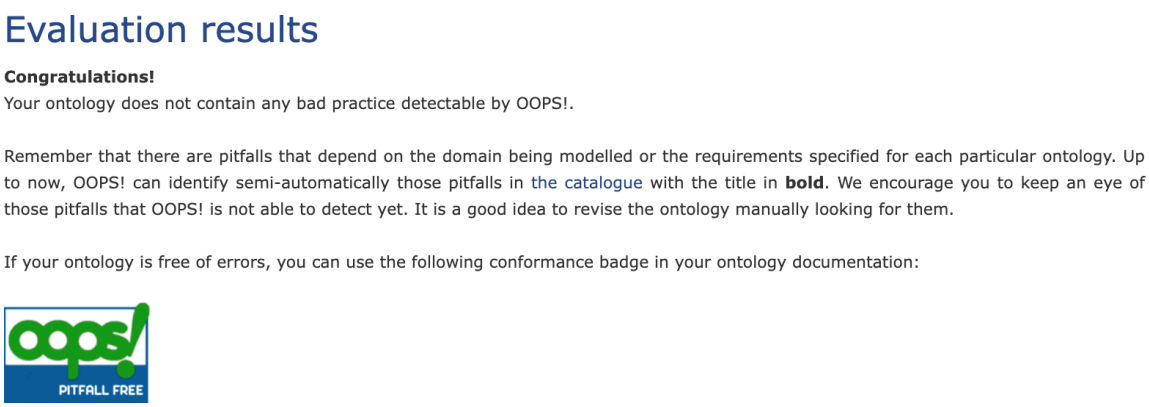



**Figura 30 Resultado de la evaluación de la ontología PMont**

Para finalizar, la ontología PMont se encuentra disponible en un repositorio de Github (https://github.com/hectorguedea/pmont). Además, se generó documentación automática utilizando el sistema OnToology (http://ontoology.linkeddata.es/) y Widoco (https://github.com /dgarijo/Widoco) (véase Figura 31).

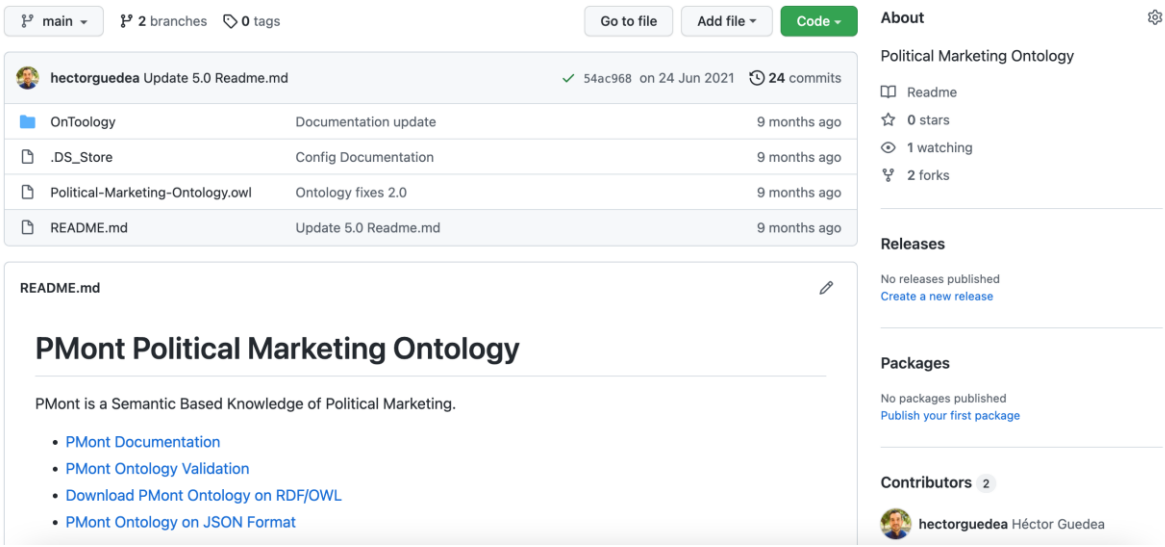



**Figura 31 Repositorio de la ontología PMont en Github**

# Capítulo 4. Entorno semántico para el enriquecimiento de un grafo de conocimiento para marketing político

En este capítulo se describe a detalle un sistema automático para la población de un grafo de conocimiento para marketing político, partiendo del modelo ontológico descrito en el anterior capítulo.

El entorno propuesto tiene como objetivo brindar una solución automatizada basada en textos en castellano sobre los procesos de extracción, colección e inserción de datos masivos relevantes desde fuentes de medios digitales semiestructurados y no estructurados para finalmente enriquecer un grafo de conocimiento previamente definido por un modelo ontológico de contexto de marketing político. La propuesta de entorno semántico contiene los siguientes componentes generales: (i) conexión de fuente y recopilación de datos, (ii) procedimiento de extracción de información, (iii) proceso de población y validación de la ontología, y por último, (iv) el grafo de conocimiento.

A continuación, se presenta la arquitectura funcional del sistema.

## 4.1 Arquitectura del sistema

La arquitectura funcional del sistema se muestra en la Figura 32 y comprende cinco componentes principales mencionados en la introducción de este capítulo. En esta caracterización, tanto la ontología como el grafo de conocimiento se diferencian para distinguir entre el esquema de ontología (es decir, PMont) y la ontología poblada (es decir, grafo de conocimiento). En resumen, el sistema funciona de la siguiente manera. La entrada del sistema son datos heterogéneos de las redes sociales y sitios de noticias. Primero, los datos se recuperan de su fuente con el uso de técnicas de web scraping y API y se almacenan en una base de datos relacional intermedia. Luego, los candidatos a instancias de la ontología PMont se extraen de los datos recopilados con una combinación de métodos y algoritmos NLP, NER y ML. Durante la etapa de población ontológica, se agregan nuevas

instancias de entidades y relaciones a la base de conocimiento, y se lleva a cabo un proceso de validación automática para garantizar la consistencia semántica del grafo con los nuevos axiomas. Finalmente, el grafo de conocimiento aumentado está disponible para su visualización o análisis con el objetivo de ayudar en la definición de campañas de marketing político exitosas.

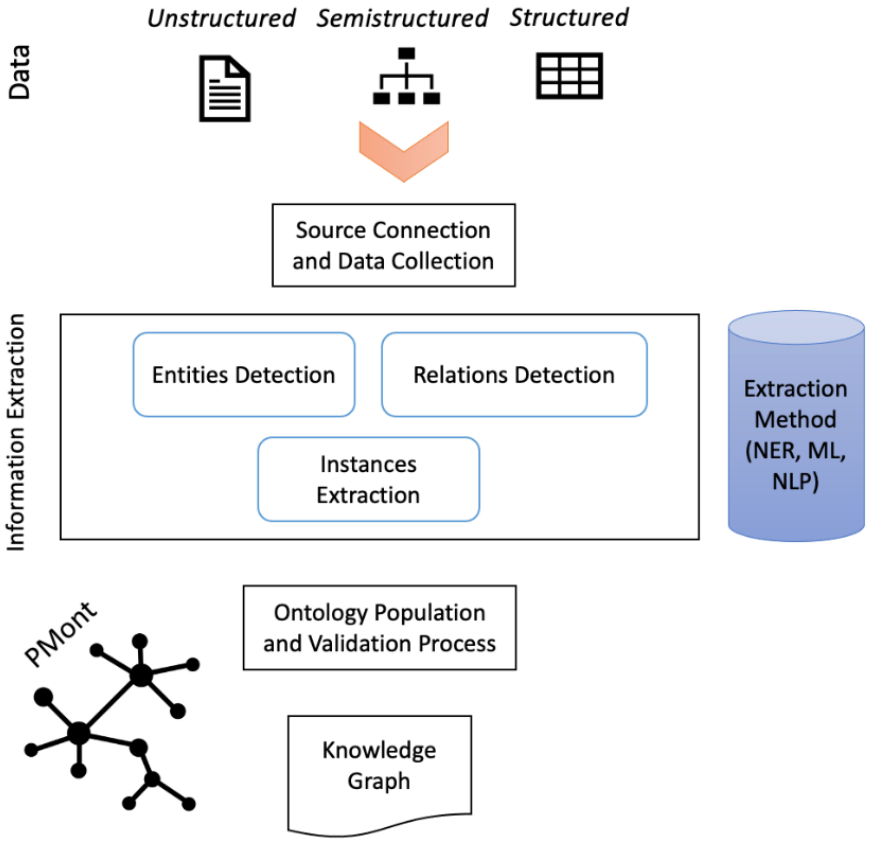



**Figura 32 Arquitectura funcional del sistema**

## 4.2 Arquitectura de software

Se definió la arquitectura de software (véase Figura 33) basado en los requisitos técnicos y operacionales del sistema, se establecieron los siguientes criterios generales:

1. **Servicios de Internet**: agrupa los servicios que se encuentran en la nube como páginas web y redes sociales como Twitter, los métodos de extracción de la información son exclusivamente a través de APIs o web scraping, respectivamente.
2. **Bases de datos locales**: reúne tres servicios de almacenamiento de datos desdela propia ontología desarrollada en formato OWL, un archivo CSV disponible dentro de la carpeta del sistema y una base de datos relacional MySQL dentro del servidor Apache.
3. **Aplicación en Java y modelo MVC**: se seleccionó Java como lenguaje de programación para el desarrollo de este sistema por ser robusto y muy compatible en librerías de conectividad y extracción de datos. El modelo, vista, controlador nos permitió diseñar una lógica para la reutilización de código en distintos componentes del sistema.
4. **Gestor de librerías**: se implementó Apache Maven (https://maven.apache.org) como gestor de herramientas y recursos. Con Maven es muy fácil añadir o quitar dependencias dentro del sistema, por ejemplo, incluir OWLAPI, Standford CoreNLP, Twitter API, Textrazor, entre otros.

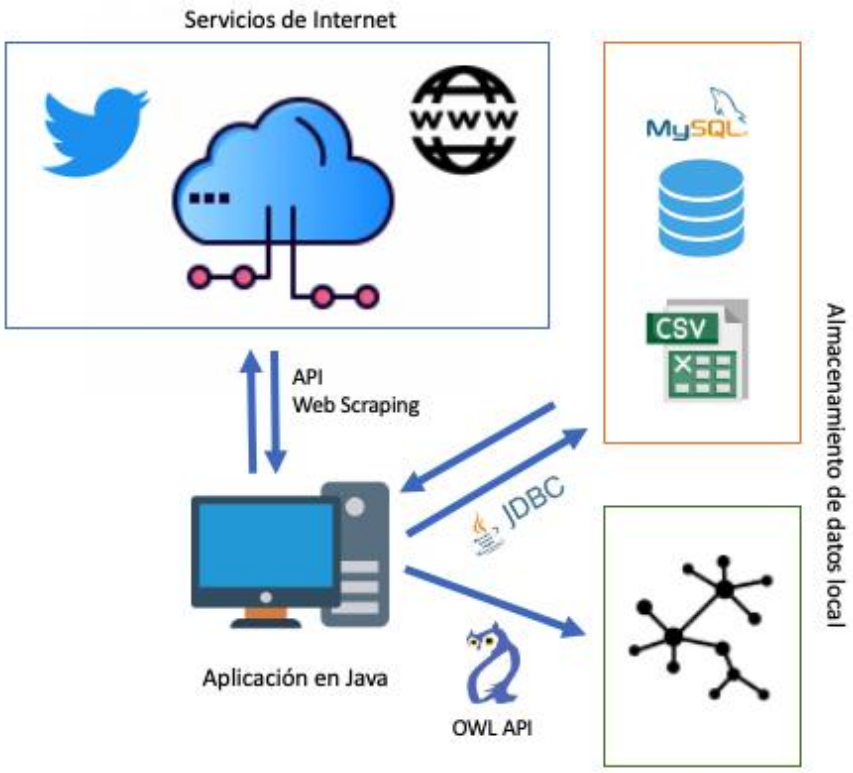



**Figura 33 Arquitectura de Software**

Esta arquitectura de software se nombró como ***Integroly***, el cual reúne el modelo ontológico y los componentes para la población y enriquecimiento del grafo de conocimiento.

## 4.3 Conectividad a fuentes y almacenamiento de datos

La conectividad a fuentes y almacenamiento de datos inicial nuestra propuesta de arquitectura del sistema. Su objetivo principal es recopilar datos de una Su objetivo principal es recopilar datos a partir de una variedad de fuentes seleccionadas (estructuradas, semiestructuradas y no estructuradas) utilizando API y librerías de web scraping.

Se utilizaron tres técnicas o soluciones para conectarse a fuentes de datos elegidas, a saber, (i) API, (ii) web scraping y (iii) bases de datos locales como MySQL.

### 4.3.1 Extracción de datos masivos en redes sociales: Twitter

La extracción de datos masivos desde redes sociales, en específico Twitter, usamos su API para solicitar los datos en tiempo real a través del mapeo de autenticación de OAuth 2.0 y token de acceso. Implementamos Twittered (https://github.com/redouane59/twittered), una biblioteca de Java para consumir la API y adicionalmente, "*Simple Logging Facade for Java*", mejor conocida por su abreviatura como SLF4J (https://www.slf4j.org), dependencia utilizada como una API genérica que hace que el registro sea independiente de la implementación actual. El API de consulta y función de búsqueda de tweets extrae los datos y retorna un objetivo JSON con campos y valores, algunos de ellos son: `tweet id`, `displayed name` (username), `name`, `location`, `text`, `likes-retweets-and-replies count`.

En la Figura 34 se puede apreciar un extracto de código del componen de software encargado de la búsqueda y extracción de datos masivos en Twitter.

```
public void searchTwitter(String searchString) throws Exception {
    MySQLDatabase mysql = new MySQLDatabase();
    TextRazorClass textrazor = new TextRazorClass();

    TwitterClient twitterClient = new TwitterClient(TwitterCredentials.builder()
            .accessToken(GlobalVar.ACCESS_TOKEN)
            .accessTokenSecret(GlobalVar.ACCESS_TOKEN_SECRET)
            .apiKey(GlobalVar.API_KEY_TWITTER)
            .apiSecretKey(GlobalVar.API_KEY_SECRET_TWITTER)
            .build());

     TweetList result = twitterClient.searchAllTweets( query: searchString + " -is:reply -is:retweet", AdditionalParameters.builder()
            .recursiveCall(false).maxResults(100).build());

    try {
        for (Tweet tweet : result.getData()) {

            Tweet tweetin = twitterClient.getTweet(tweet.getId());
            System.out.println(tweetin.getUser().getDisplayedName());
            System.out.println(tweetin.getUser().getName());
            System.out.println(tweetin.getUser().getLocation());

            System.out.println(tweet.getText());
            System.out.println(tweet.getCreatedAt());
            System.out.println(tweet.getLang());
            System.out.println(tweet.getLikeCount());
            System.out.println(tweet.getRetweetCount());
            System.out.println(tweet.getReplyCount());
```

**Figura 34 Fragmento de código de las búsquedas en Twitter**

El resultado de la consulta en formato JSON se almacena de forma temporal en la base de datos relacional MySQL para ser utilizado en los procesos de enriquecimiento de información y listado de instancias candidatas, véase a más a detalle en el apartado 4.3.4.

### 4.3.2 Extracción de datos con Web Scraping

Web scraping se implementa como una técnica para extraer datos de sitios de noticias y otros sitios web de interés para el marketing político. Para ello, utilizamos TextRazor (https://www.textrazor.com), una infraestructura REST auto-hospedada y en la nube completa para la extracción de entidades, frases clave, etiquetas y clasificación. Nuestro sistema permite a los usuarios agregar un solo sitio web o una lista de sitios web para su extracción, así como texto limpio proveniente del contenido principal de la página Web.

En la Figura 35 se muestra un fragmento de código con la función principal para evaluar la URL e iniciar el proceso de web scraping.

```
public void urlAnalize(String apiKey, String url) throws Exception {

    TextRazor client = new TextRazor(apiKey);
    SentimentAnalyzerService analyzerService = new SentimentAnalyzerService();
    String nameURL = this.getHostName(url);

    client.setCleanupMode("cleanHTML");

    client.setExtractors(Arrays.asList("relations, words,entities,topics,senses "));
    client.setClassifiers(Arrays.asList("textrazor_newscodes"));
    client.setCleanupReturnCleaned(true);

    AnalyzedText response = client.analyzeUrl(url);
    System.out.println("Nombre del sitio:" + nameURL);
    System.out.println("Texto extraído del enlace: ");

    String TextContent = response.getResponse().getCleanedText();


    System.out.println(TextContent);
    SentimentType.fromValue(analyzerService.analyse(TextContent));
```

**Figura 35 Web scraping con Textrazor**

Al igual que con la técnica anterior, el resultado de la consulta de web scraping se almacena de forma temporal en la base de datos relacional MySQL, véase a más a detalle en el apartado 4.3.4.

### 4.3.3 Extracción de datos con CSV

Se creó un componente para consulta y lectura de archivos locales en formato CSV para la extracción de la información por medio de filas y columnas utilizando la biblioteca de Java, OpenCSV y CSVReader (http://opencsv.sourceforge.net).

La extracción de datos por medio de CSV se divide en dos técnicas o tipos de utilización:

1. La primera, el CSV actúa únicamente como base de datos con URLs de sitios Web de interés, donde para cada URL o sitio web se aplicará la técnica de web scraping para la extracción de la información relevante (véase 4.3.2).
2. La segunda, el CSV contiene propiamente el contenido relevante, esto es, es una encuesta de opinión con las respuestas de los ciudadanos, es un instrumento de censo del electorado, entre otros.

En la Figura 36 se muestra un fragmento de código principal para la extracción y lectura de valores de archivos locales CVS, sin embargo, es posible que la función reciba una URL como variable *String*, donde se encuentre un archivo de este formato.

```java
public class CSVReaderClass {

    public List<String> CSVReaderMethod(String file) throws IOException, CsvValidationException {

            CSVReader reader = null;
            String[] line = null;
            List<String> outline = new ArrayList<~>();

        try {
                var fileread = new FileReader(file, StandardCharsets.UTF_8);

                reader = new CSVReader(fileread);

                while((line = reader.readNext()) != null) {
                    outline.add(line[0].replaceAll( regex: "[\uFEFF-\uFFFF]", replacement: ""));
                }

            } catch (Exception e) {
                System.out.println("Error " + e.getMessage());
            } finally {
                if (null != reader) {
                    reader.close();
                }
            }
        return outline;
       }

    }
```

**Figura 36 Lectura de valores de archivo CSV**

Al igual que con las técnicas anteriores, el resultado de la consulta de información se almacenará de forma temporal en la base de datos relacional MySQL, véase a más a detalle en el apartado 4.3.4.

### 4.3.4 Esquema de almacenamiento en base de datos relacional

Todos los datos brutos obtenidos a través de las fuentes descritas anteriormente se insertan como filas una base de datos relacional intermediaria basada en una estructura de datos paralela a las clases de ontología PMont. La Figura 37 muestra las tablas creadas en la base de datos MySQL, las cuales son `tweets`, `pages` y `texts` como tipos de fuentes, y una tabla índice utilizada como diccionario de IDs. Por lo tanto, los datos sin procesar se separan de los datos procesados producidos en etapas posteriores del proceso. Es decir, los resultados del proceso de análisis de datos utilizando NLP, NER, técnicas y herramientas

de extracción o clasificación de relaciones, como se describe en la siguiente sección, no se incluyen en la base de datos, pero se administran en la memoria para permitir un procesamiento posterior .

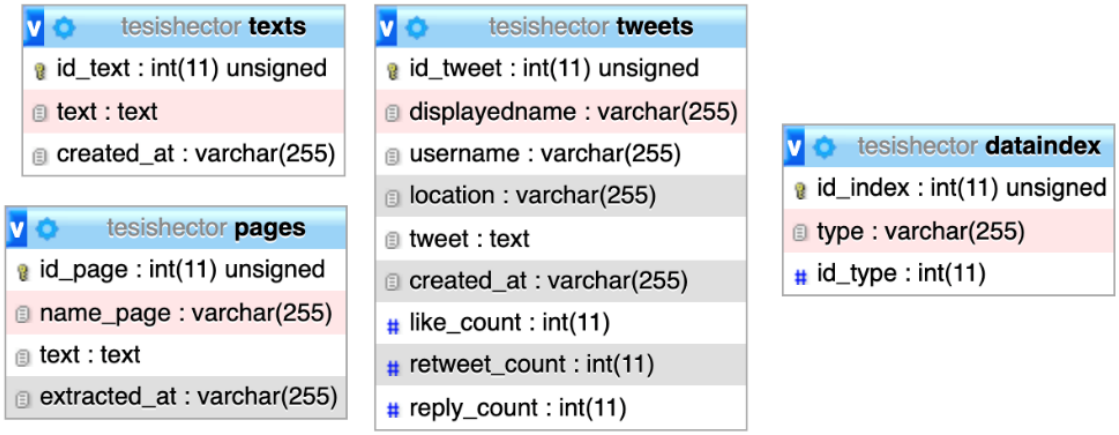



**Figura 37 Esquema de la base de datos relacional**

## 4.4 Extracción de la información y NLP

Para recopilar conocimiento semántico de los datos extraídos y almacenados de manera temporal en la base de datos relacional, utilizamos una serie de técnicas específicas como lo son: NER, ML y NLP. La combinación de estas técnicas facilita la detección y extracción de las entidades, relaciones y atributos requeridos para el proceso de instanciación en la base de conocimiento. TextRazor REST API fue la herramienta elegida para el análisis de texto basado en un enfoque NER y ML, mientras que Stanford CoreNLP, una biblioteca Java con un enfoque de NLP, fue seleccionado para realizar análisis de sentimiento y determinar la polaridad del texto (ya sea positivo, negativo o neutral).

A continuación, se describen los métodos individuales para la extracción de conocimiento según el modelo ontológico de PMont:

### 4.4.1 Análisis de sentimiento

Como se mencionó anteriormente, la extracción de opiniones y polaridades se realiza mediante el kit de herramientas Stanford CoreNLP, un enfoque con una buena comprensión

de los metadatos y textos en castellano, que implementa léxicos de opiniones existentes (incluidas las faltas de ortografía, las variantes morfológicas, la jerga y el marcado de las redes sociales) y la extracción de características básicas (*tokenization*, *stemming*, *POS-tagging*). La combinación completa de funciones nos proporcionó una herramienta robusta para el análisis de sentimiento y polaridad.

### 4.4.2 Reconocimiento de Entidades Nombradas

Para NER (o Reconocimiento de Entidades Nombradas), se utilizó nuevamente la suite de APIs de TextRazor. Esta API funciona aprovechando una enorme base de conocimiento de entidades extraídas de varias fuentes web, incluidas Wikipedia, DBPedia y Wikidata. Un motor de coincidencias se encarga de encontrar relaciones entre el contenido del texto y las posibles entidades agrupadas como diccionarios. Además, se utiliza un etiquetador basado en estadísticas para identificar personas, lugares y empresas, entre otros conceptos, y se aplican expresiones regulares para detectar los elementos menos ambiguos, como direcciones de correo electrónico y sitios web.

### 4.4.3 Relaciones

En cuanto a la extracción de relaciones, generamos datos vinculados utilizando la API de TextRazor, primero, eliminando la ambigüedad de los términos y, segundo, vinculando las entidades a las ID canónicas en la web vinculada. En este proceso, convertimos texto sin procesar en un conjunto de expresiones triple "*sujeto-predicado-objeto*" bien estructuradas que relacionaban entidades, frases y conceptos. El enfoque de vincular entidades a identificaciones canónicas ofrece una gran precisión y recuperación, brindando la flexibilidad para extraer una cantidad ilimitada de tipos de relaciones.

### 4.4.4 Casos de Estudio

Con el objetivo de demostrar la funcionalidad de las diferentes técnicas utilizadas en la sección 4.4 (es decir, NER, extracción de relaciones y NLP) y posteriormente la instanciación en la ontología, este apartado será dedicado a describir todo el proceso.

Actualmente la comunicación con el sistema es a través de línea de comandos, las opciones se presentan en pantalla y el usuario selecciona el tipo de dato o funcionalidad que desea activar. Las opciones actuales son Figura 38:

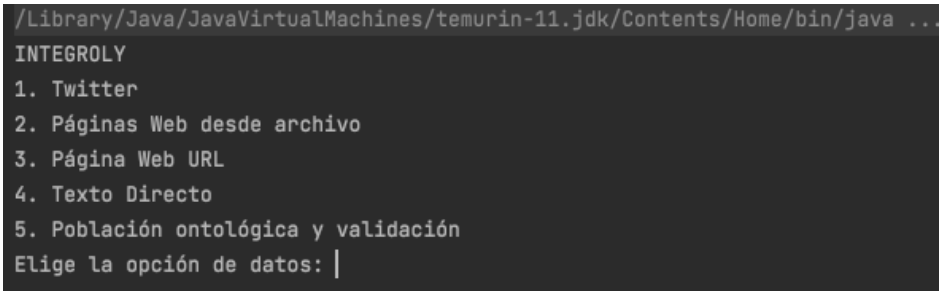



**Figura 38 Opciones de Integroly**

A continuación, se presentan tres tipos de casos con diferencias en la extracción y estructura de datos, el primer caso 4.4.4.1 sobre un texto libre, el segundo caso 4.4.4.2, involucra la extracción de datos en tiempo real desde Twitter para su posterior procesamiento, y por último, extracción de datos con web scraping sobre un sitio web de noticias.

#### *4.4.4.1 Caso #1: Texto libre*

Por motivos de ejemplo seleccionaremos la opción “4. Texto directo” del sistema *Integroly*, indica texto libre de datos no estructurados. El usuario tendrá la posibilidad de agregar el texto que desea. El siguiente texto es un registro real de la tabla de `texts` en la base de datos (un extracto del texto en lenguaje natural escrito en español extraído de un sitio web de noticias):

> “De acuerdo a la evaluación promedio registrada por MITOFSKY en diciembre de 2021 para El Economista, el presidente López Obrador obtuvo una aprobación promedio de 66% durante el último mes del 2021, la más alta desde febrero de 2019 cuando alcanzó el 67%; en diciembre pasado la desaprobación presidencial promedio fue de 34 por ciento. Se identificó que por tipo de ocupación, los campesinos fueron quienes tuvieron el mayor porcentaje a favor del tabasqueño con 83.7 por ciento. Respecto de la 'percepción de seguridad', el 41% de los encuestados opinó que está 'igual', el 28,3% que está 'peor', el 24,7% que está 'mejor', y el 6% 'no requirió'”.

La primera etapa es implementar NER y análisis de extracción de relaciones con el uso de la API TextRazor. Los resultados representados en la Figura 39 muestran las entidades y las relaciones semánticas. En la Figura 40 se muestra la categorización y clasificación del texto a partir de la identificación semántica de frases clave y conceptos destacados relacionados con el dominio. El número al lado de cada categoría y tema es el puntaje coincidente.

```
Entidad:  Tabasco -- [Place, PopulatedPlace, Region, AdministrativeRegion]
Entidad:  2021 -- [Number]
Entidad:  Andrés Manuel López Obrador -- [Agent, Person]
Entidad:  2019 -- [Number]
Entidad:  67 -- [Number]
Entidad:  34 -- [Number]
Entidad:  83.6999999999999999903 -- [Number]
Entidad:  2021 -- [Number]
Entidad:  66 -- [Number]
```

**Figura 39 NER y resultados de extracción de relaciones (parcialmente en español)**

```
Categoría:20000593 politics>government 0.8575
Categoría:11000000 politics 0.8497
Categoría:20000610 politics>government>heads of state 0.5618
Categoría:20000654 politics>political process>political system>democracy 0.4533
Categoría:20000651 politics>political process>political parties and movements 0.4067
Categoría:20000606 politics>government>executive (government) 0.3895
Categoría:20000745 science and technology>social sciences>economics 0.3534
Categoría:20000574 politics>election 0.3384
Categoría:20000592 politics>fundamental rights>human rights 0.3225
Categoría:20000614 politics>government>national government 0.3221

Tema: Politics WikiID: Q7163 Score: 1.0
Tema: Government WikiID: Q7188 Score: 1.0
Tema: Andrés Manuel López Obrador WikiID: Q318508 Score: 0.7553
Tema: Presidencies WikiID: Q11708087 Score: 0.7128
Tema: Latin America WikiID: Q12585 Score: 0.6534
Tema: Mexico WikiID: Q96 Score: 0.644
Tema: South America WikiID: Q18 Score: 0.6053
Tema: Middle America (Americas) WikiID: Q29876 Score: 0.6019
Tema: Human activities WikiID: Q61788060 Score: 0.5981
Tema: Government of Mexico WikiID: Q2182191 Score: 0.591
Tema: Presidents WikiID: Q30461 Score: 0.574
Tema: Politics of Mexico WikiID: Q1155509 Score: 0.5691
```

**Figura 40 Resultados de clasificación**

Finalmente, se aplica el análisis de sentimiento con la biblioteca de Stanford CoreNLP obteniendo tres posibles valores: positivo, neutro o negativo. Para el texto presentado anteriormente, el resultado del análisis de contenido es "neutral" (véase Figura 41).

```
[main] INFO edu.stanford.nlp.pipeline.StanfordCoreNLP - Adding annotator tokenize
[main] INFO edu.stanford.nlp.pipeline.StanfordCoreNLP - Adding annotator ssplit
[main] INFO edu.stanford.nlp.pipeline.StanfordCoreNLP - Adding annotator pos
[main] INFO edu.stanford.nlp.tagger.maxent.MaxentTagger - Loading POS tagger from e
[main] INFO edu.stanford.nlp.pipeline.StanfordCoreNLP - Adding annotator parse
[main] INFO edu.stanford.nlp.parser.common.ParserGrammar - Loading parser from seri
[main] INFO edu.stanford.nlp.pipeline.StanfordCoreNLP - Adding annotator sentiment
[main] INFO edu.stanford.nlp.sentiment.SentimentModel - Loading sentiment model edu
NEUTRAL
```

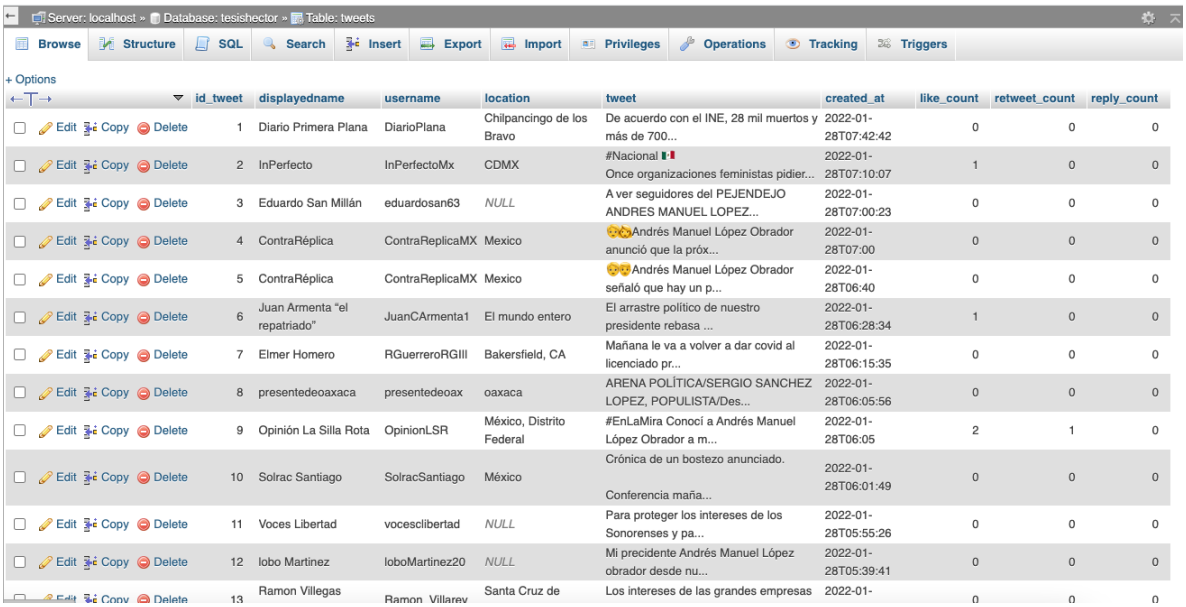
**Figura 41 Resultado del análisis de opinión**

#### *4.4.4.2 Caso #2: Twitter*

En este caso se presenta la búsqueda libre sobre el *stream public tweets* de Twitter, con la obtención de los resultados en tiempo real para mostrarse en pantalla (Figura 42) e insertados en la tabla `tweets` de la base de datos relacional (Figura 43).

```
Elige la opción de datos: 1
Búsqueda en Twitter:
Andrés Manuel López Obrador
Pedro Mellado R.
PedroMelladoR
Guadalajara, Jalisco.
Celebra el presidente Andrés Manuel López Obrador, la aprobación y publicación del decreto de interpretación de la
2022-03-18T17:38:56
es
0
0
0
AND-VOX
ANDVOX1
```

**Figura 42 Búsqueda y resultados de Twitter**


Server: localhost » Database: tesishector » Table: tweets

Browse | Structure | SQL | Search | Insert | Export | Import | Privileges | Operations | Tracking | Triggers

+ Options

| | id_tweet | displayedname | username | location | tweet | created_at | like_count | retweet_count | reply_count |
|---|---|---|---|---|---|---|---|---|---|
| Edit Copy Delete | 1 | Diario Primera Plana | DiarioPlana | Chilpancingo de los Bravo | De acuerdo con el INE, 28 mil muertos y más de 700... | 2022-01-28T07:42:42 | 0 | 0 | 0 |
| Edit Copy Delete | 2 | InPerfecto | InPerfectoMx | CDMX | #Nacional 🇲🇽 Once organizaciones feministas pidier... | 2022-01-28T07:10:07 | 1 | 0 | 0 |
| Edit Copy Delete | 3 | Eduardo San Millán | eduardosan63 | NULL | A ver seguidores del PEJENDEJO ANDRES MANUEL LOPEZ... | 2022-01-28T07:00:23 | 0 | 0 | 0 |
| Edit Copy Delete | 4 | ContraRéplica | ContraReplicaMX | Mexico | 🤓🤓Andrés Manuel López Obrador anunció que la próx... | 2022-01-28T07:00 | 0 | 0 | 0 |
| Edit Copy Delete | 5 | ContraRéplica | ContraReplicaMX | Mexico | 🤓🤓Andrés Manuel López Obrador señaló que hay un p... | 2022-01-28T06:40 | 0 | 0 | 0 |
| Edit Copy Delete | 6 | Juan Armenta "el repatriado" | JuanCArmenta1 | El mundo entero | El arrastre político de nuestro presidente rebasa ... | 2022-01-28T06:28:34 | 1 | 0 | 0 |
| Edit Copy Delete | 7 | Elmer Homero | RGuerreroRGIII | Bakersfield, CA | Mañana le va a volver a dar covid al licenciado pr... | 2022-01-28T06:15:35 | 0 | 0 | 0 |
| Edit Copy Delete | 8 | presentedeoaxaca | presentedeoax | oaxaca | ARENA POLÍTICA/SERGIO SANCHEZ LOPEZ, POPULISTA/Des... | 2022-01-28T06:05:56 | 0 | 0 | 0 |
| Edit Copy Delete | 9 | Opinión La Silla Rota | OpinionLSR | México, Distrito Federal | #EnLaMira Conocí a Andrés Manuel López Obrador a m... | 2022-01-28T06:05 | 2 | 1 | 0 |
| Edit Copy Delete | 10 | Solrac Santiago | SolracSantiago | México | Crónica de un bostezo anunciado. Conferencia maña... | 2022-01-28T06:01:49 | 0 | 0 | 0 |
| Edit Copy Delete | 11 | Voces Libertad | vocesclibertad | NULL | Para proteger los intereses de los Sonorenses y pa... | 2022-01-28T05:55:26 | 0 | 0 | 0 |
| Edit Copy Delete | 12 | lobo Martinez | loboMartinez20 | NULL | Mi precidente Andrés Manuel López obrador desde nu... | 2022-01-28T05:39:41 | 0 | 0 | 0 |
| Edit Copy Delete | 13 | Ramon Villegas | Ramon_Villarey | Santa Cruz de | Los intereses de las grandes empresas | 2022-01- | 0 | 0 | 0 |

**Figura 43 Inserción de resultados de Twitter en BD Relacional**

En las Figuras Figura 44,Figura 45, Figura 46 se muestra la implementación de técnicas de NLP, NER y análisis de sentimiento sobre los textos en castellano de Twitter.

```
Entidad:  Revocatoria del mandato -- null
Entidad:  Andrés Manuel López Obrador -- [Agent, Person]
Entidad:  https://t.co/d6zwTk0SAM -- [URL]
Categoría:11000000 politics 0.8497
Categoría:20000593 politics>government 0.8201
Categoría:20000121 crime, law and justice>law 0.6281
Categoría:20000124 crime, law and justice>law>civil law>regulation 0.4989
Categoría:20000574 politics>election 0.4735
Categoría:20000122 crime, law and justice>law>civil law 0.4447
Categoría:02000000 crime, law and justice 0.4228
Categoría:20000610 politics>government>heads of state 0.4121
Categoría:20000654 politics>political process>political system>democracy 0.4093
Categoría:20000106 crime, law and justice>judiciary 0.394
```

**Figura 44 NER y resultados de extracción de relaciones para texto de Twitter**

```
Tema: Government WikiID: Q7188 Score: 1.0
Tema: Politics WikiID: Q7163 Score: 1.0
Tema: Law WikiID: Q7748 Score: 1.0
Tema: Governance WikiID: Q1553864 Score: 1.0
Tema: Justice WikiID: Q5167661 Score: 0.8734
Tema: Human activities WikiID: Q61788060 Score: 0.78
Tema: Accountability WikiID: Q2798912 Score: 0.7795
Tema: Presidencies WikiID: Q11708087 Score: 0.7252
Tema: Andrés Manuel López Obrador WikiID: Q318508 Score: 0.7092
Tema: Latin America WikiID: Q12585 Score: 0.6867
Tema: Public law WikiID: Q207892 Score: 0.6763
Tema: Middle America (Americas) WikiID: Q29876 Score: 0.67
Tema: South America WikiID: Q18 Score: 0.6605
Tema: Social institutions WikiID: Q178706 Score: 0.6417
Tema: Recall election WikiID: Q1196663 Score: 0.5904
Tema: Presidents WikiID: Q30461 Score: 0.5736
Tema: Issues in ethics WikiID: Q9465 Score: 0.5729
Tema: Democracy WikiID: Q7174 Score: 0.5508
```

**Figura 45 Resultados de clasificación de texto de Twitter**

```
celebro el presidente Andrés Manuel López Obrador, la aprobación y publicación del decreto de interpretación de las leyes vigentes en materia electoral que permitirán a juna
[main] INFO edu.stanford.nlp.pipeline.StanfordCoreNLP - Adding annotator tokenize
[main] INFO edu.stanford.nlp.pipeline.StanfordCoreNLP - Adding annotator ssplit
[main] INFO edu.stanford.nlp.pipeline.StanfordCoreNLP - Adding annotator pos
[main] INFO edu.stanford.nlp.tagger.maxent.MaxentTagger - Loading POS tagger from edu/stanford/nlp/models/pos-tagger/spanish-ud.tagger ... done [1.1 sec].
[main] INFO edu.stanford.nlp.pipeline.StanfordCoreNLP - Adding annotator parse
[main] INFO edu.stanford.nlp.parser.common.ParserGrammar - Loading parser from serialized file edu/stanford/nlp/models/srparser/spanishSR.beam.ser.gz ... done [11.2 sec].
[main] INFO edu.stanford.nlp.pipeline.StanfordCoreNLP - Adding annotator sentiment
[main] INFO edu.stanford.nlp.sentiment.SentimentModel - Loading sentiment model edu/stanford/nlp/models/sentiment/sentiment.ser.gz ... done [0.5 sec].
NEGATIVE
```

**Figura 46 Resultado del análisis de polaridad en texto de Twitter**

#### *4.4.4.3 Caso #3: Web scraping*

Para este caso, el usuario elige la opción 3 del sistema *Integroly*, agrega un enlace o URL de cualquier sitio web de noticias o general, para este ejemplo fue https://www.jornada.com.mx/notas/2022/03/18/politica/consulta-de-revocacion-para-dar-legitimidad-senala-lopez-obrador, el sistema aplicará las técnicas de web scraping, obteniendo nombre del sitio y la información relevante, y limpiando los elementos

innecesarios (Figura 47) para posteriormente insertar los nuevos valores como una nueva fila en la tabla `pages` de la base de datos relacional (Figura 48).

```
Elige la opción de datos: 3
URL de página Web:
https://www.jornada.com.mx/notas/2022/03/18/politica/consulta-de-revocacion-para-dar-legitimidad-senala-lopez-obrador/
Nombre del sitio:jornada.com.mx
Texto extraído del enlace:
21°C - muy nublado
Ciudad de México, CDMX
21°C - muy nublado
Consulta de revocación, para legitimidad del presidente: AMLO
El presidente debe “tener legitimidad”, por ello la consulta de revocación de mandato es un asunto moral y político, sostuvo
Minatitlán, Ver. El presidente “tiene que tener legitimidad” por ello la consulta de revocación de mandato es un asunto moral
Al cuestionarle si el gobierno aceptaría que la Suprema Corte de Justicia de la Nación (SCJN) aceptara la impugnación de decr
Y lacónico, el secretario de Gobernación, Adán Augusto López, le secundó a su modo: “nosotros no vemos que se pueda impugnar,
En este enclave petrolero -donde encabezará la ceremonia del 84 aniversario de la expropiación petrolera- el tabasqueño urgió
```

**Figura 47 Resultado de Web Scraping**


**Figura 48 Inserción del resultado del Web scraping en BD relacional**

En las Figuras Figura 49, Figura 50 y Figura 51 se presentan los análisis del texto resultado del web scraping a través de las técnicas de NER, NLP y análisis de sentimiento.

```
Entidad:  Político -- null
Entidad:  Andrés Manuel López Obrador -- [Agent, Person]
Entidad:  Revocatoria del mandato -- null
Entidad:  Político -- null
Entidad:  Andrés Manuel López Obrador -- [Agent, Person]
Entidad:  Revocatoria del mandato -- null
Entidad:  Andrés Manuel López Obrador -- [Agent, Person]
Categoría:20000593 politics>government 0.8667
Categoría:11000000 politics 0.8497
Categoría:20000654 politics>political process>political system>democracy 0.7171
Categoría:20000651 politics>political process>political parties and movements 0.6895
Categoría:20000610 politics>government>heads of state 0.5987
Categoría:20000574 politics>election 0.5224
Categoría:20000068 conflicts, war and peace>civil unrest>revolution 0.5149
Categoría:16000000 conflicts, war and peace 0.4744
Categoría:20000814 society>values>ethics 0.4236
Categoría:20000808 society>values 0.4047
```

**Figura 49 NER y resultados de extracción de relaciones para texto de página Web**

```
Tema: Andrés Manuel López Obrador WikiID: Q318508 Score: 1.0
Tema: Presidents of Mexico WikiID: Q628004 Score: 1.0
Tema: National Regeneration Movement politicians WikiID: Q18404111 Score: 1.0
Tema: National Regeneration Movement WikiID: Q15717618 Score: 1.0
Tema: Presidents of the Party of the Democratic Revolution WikiID: Q767010 Score: 1.0
Tema: Politics of Mexico City WikiID: Q9081179 Score: 1.0
Tema: Heads of Government of Mexico City WikiID: Q2962015 Score: 1.0
Tema: Mexican Christian socialists WikiID: Q25032899 Score: 1.0
Tema: Politics WikiID: Q7163 Score: 1.0
Tema: Mexico WikiID: Q96 Score: 1.0
Tema: Government of Mexico WikiID: Q2182191 Score: 1.0
Tema: Politics of Mexico WikiID: Q1155509 Score: 1.0
Tema: Government WikiID: Q7188 Score: 1.0
Tema: Politicians from Tabasco WikiID: Q6301155 Score: 0.9878
Tema: Middle America (Americas) WikiID: Q29876 Score: 0.9768
Tema: Institutional Revolutionary Party breakaway groups WikiID: Q105305911 Score: 0.967
Tema: Presidents of Mexican political parties WikiID: Q15296041 Score: 0.955
Tema: Presidencies WikiID: Q11708087 Score: 0.9418
```

**Figura 50 Resultados de clasificación de texto de página Web**

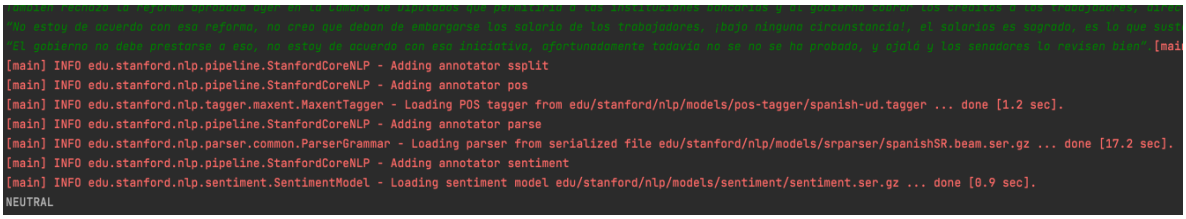



**Figura 51 Resultado del análisis de polaridad en texto de página Web**

Para los casos 4.4.4.2 y 4.4.4.3, posterior a la inserción a la base de datos relacional, y al ser llamados para el proceso de 4.5 desde el sistema de *Integroly*, se aplican las técnicas de reconocimiento de entidades, clasificación, relaciones y análisis de sentimiento, añadiendo capas semánticas para favorecer el proceso de la población ontológica.

## 4.5 Población ontológica y proceso de validación

Al termino de los procesos de extracción de información, los datos se mantienen en la base de datos relacional hasta ser llamados por el usuario del sistema de *Integroly* a través de la opción de *"población ontológica y validación"*, a partir de ese momento los datos son aumentados con anotaciones semánticas con el fin de relacionarlos con el modelo ontológico de PMont (véase Figura 52).

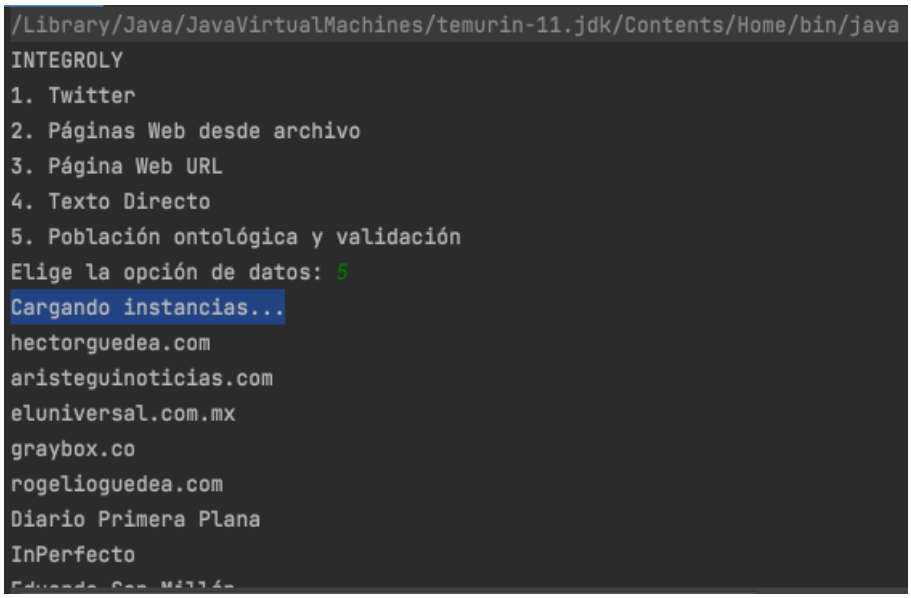



**Figura 52 Población ontológica y validación**

Los registros de la base de datos que no han sido procesados para la instanciación se etiquetan como `candidatos a instancia`, cada entrada se enriquece con los metadatos semánticos por medio de las técnicas de extracción (4.4).

Para la creación de instancias o la población de ontologías, utilizamos la biblioteca llamada OWLAPI (https://github.com/owlcs/owlapi) para crear, manipular y serializar la ontología del dominio o modelo ontológico. La API permite conectarse, leer y escribir sobre el archivo RDF/XML y OWL/XML, también incluye interfaces de razonamiento como FaCT ++ y HermiT, para ayudar a derivar hechos implícitos de un conjunto de hechos explícitos dados, en resumen, la capacidad de inferir consecuencias lógicas.

OWLAPI inicia el proceso de instanciación de los registros denominados `candidatos a instancia`, añade para cada una las anotaciones semánticas antes mencionadas, y además, agrega una referencia de la fuente que es el ID de la entrada de la base de datos (`hasValue`), esto podría ser útil para la validación de la instancia formulada por parte de un experto, llevando así un proceso semi-supervisado. El resultado es un modelo ontológico poblado con nuevas instancias.

### 4.5.1 Casos de Estudio

A continuación, se presentan los tres de los casos que se llevaron a cabo en el proceso de extracción de información, ahora como caso de estudio en el proceso de población ontológica.

#### *4.5.1.1 Caso #1: Texto libre*

En la Figura 53 se muestra la nueva instancia del caso de Texto libre (4.4.4.1) después del proceso efectuado por OWLAPI en *Integroly*, visualizada en Protégé y validada por el razonador FACT++.

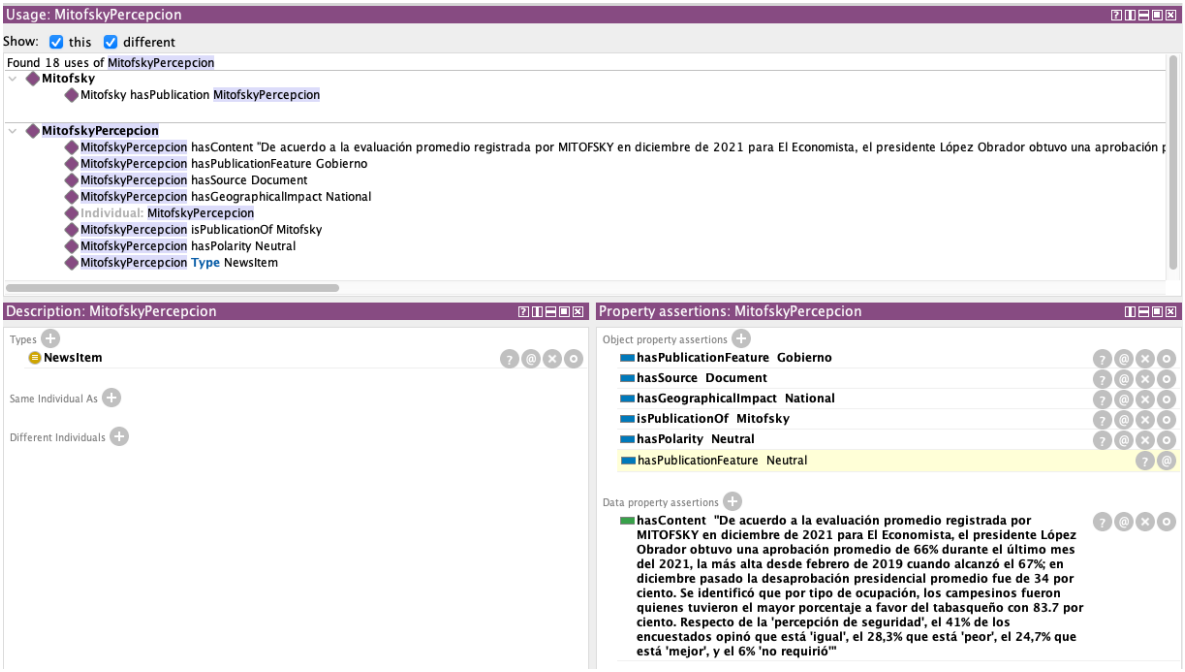



**Figura 53 Nueva instancia en PMont con publicación de texto libre, resultado con el razonador FACT++**

#### *4.5.1.2 Caso #2: Twitter*

En la Figura 54 se muestra la nueva instancia del caso de Twitter (4.4.4.2) después del proceso efectuado por OWLAPI en *Integroly*, visualizada en Protégé y validada por el razonador FACT++.

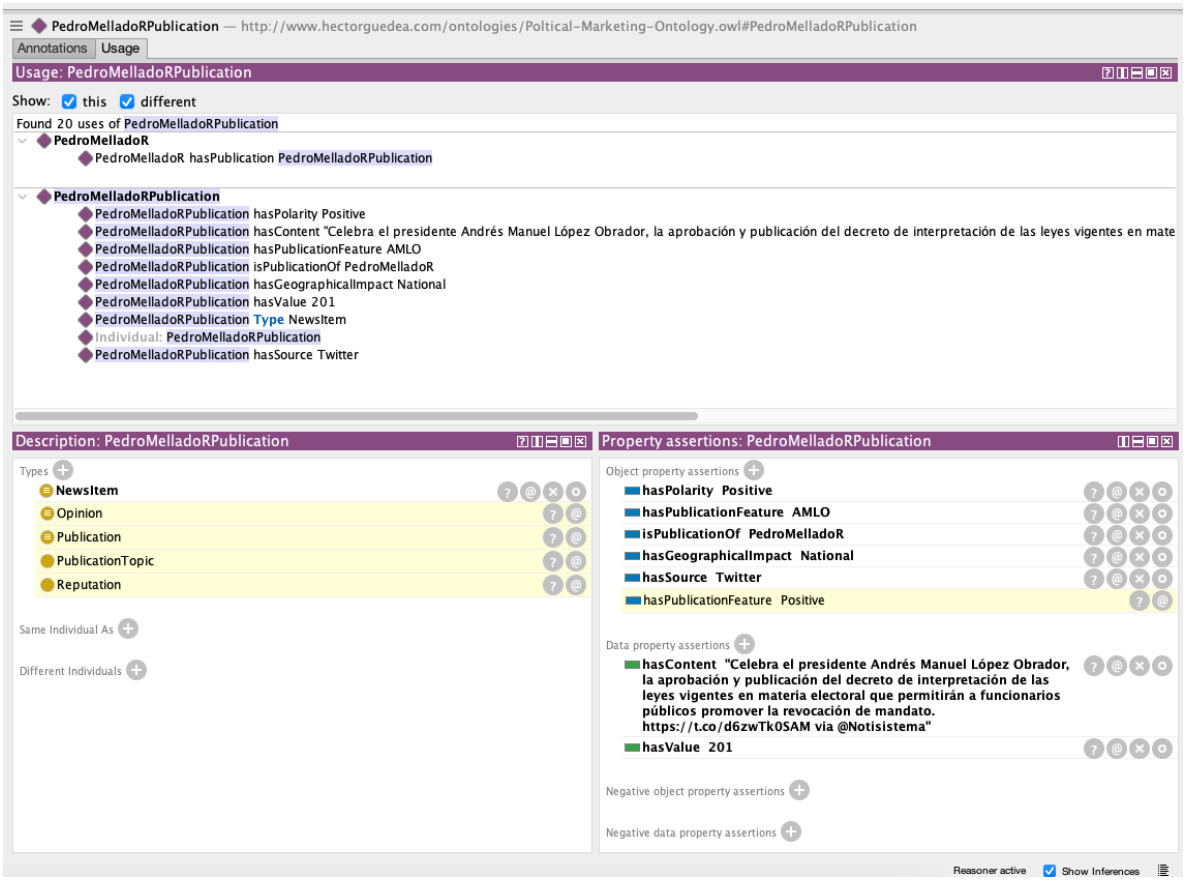



**Figura 54 Nueva instancia en PMont con publicación de Twitter, resultado con el razonador FACT++**

#### *4.5.1.3 Caso #3: Web scraping*

En la se muestra la nueva instancia del caso de web scraping (4.4.4.3) después del proceso efectuado por OWLAPI en *Integroly*, visualizada en Protégé y validada por el razonador FACT++.

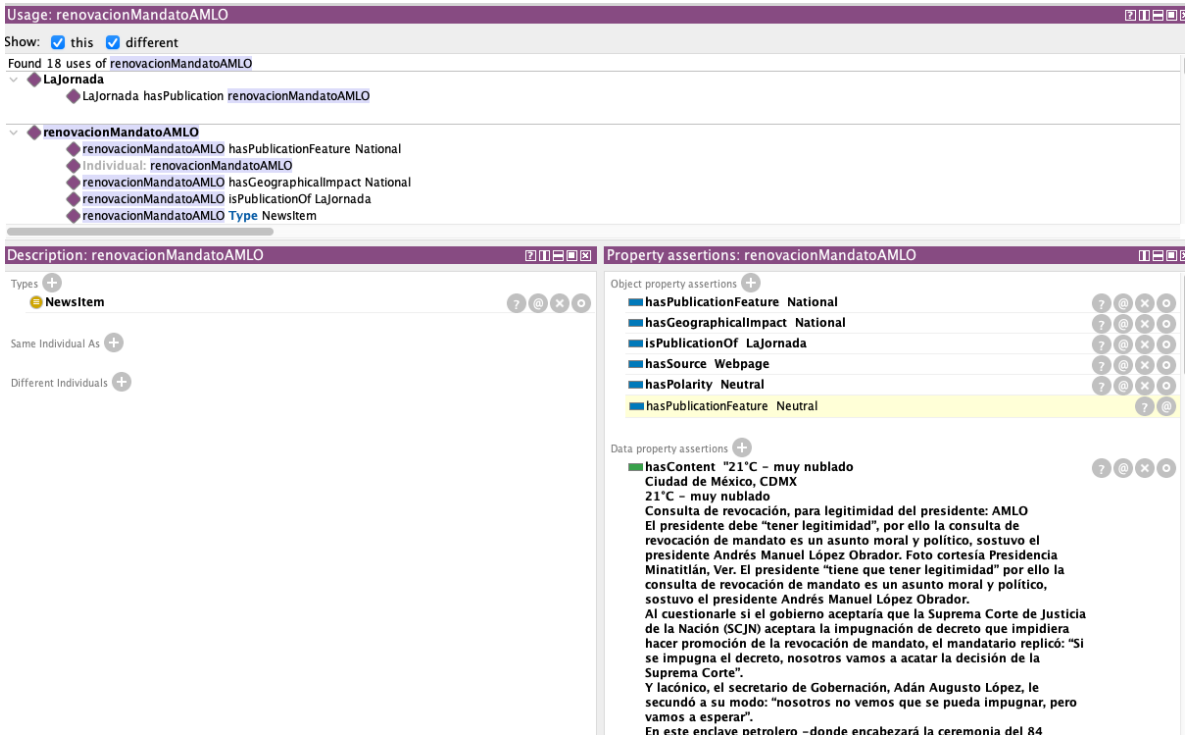



**Figura 55 Nueva instancia en PMont con publicación de Twitter, resultado con el razonador FACT++**

Nuestro método de validación establece las reglas para limpiar los datos procesados previos a la población de la ontología, el objetivo es buscar datos duplicados, ambiguos e incompletos, para mantener la consistencia de la ontología cuando los datos se insertan como una nueva instancia. Para eso, usamos HermIT (http://www.hermit-reasoner.com) como razonador para determinar la consistencia de la ontología e identificar relaciones de subsunción entre clases. HermIT se basa en el cálculo HyperTableau que proporciona un razonamiento mucho más eficiente que cualquier algoritmo conocido anteriormente y, por último, cumple totalmente con OWL 2 Direct Semantics según lo estandarizado por el World Wide Web Consortium (W3C).

## 4.6 Grafo de conocimiento

Finalmente, el grafo de conocimiento se presenta como una representación de la ontología recientemente poblada (modelo ontológico). Visualmente, el grafo de conocimiento

generado podría presentarse utilizando herramientas como OntoGraf (https://protegewiki.stanford.edu/wiki/OntoGraf), WebVOWL (https://service.tib.eu/webvowl/) y OVisualizer (https://github.com/yWorks/ontology-visualizer). Un ejemplo se muestra en la Figura 56.

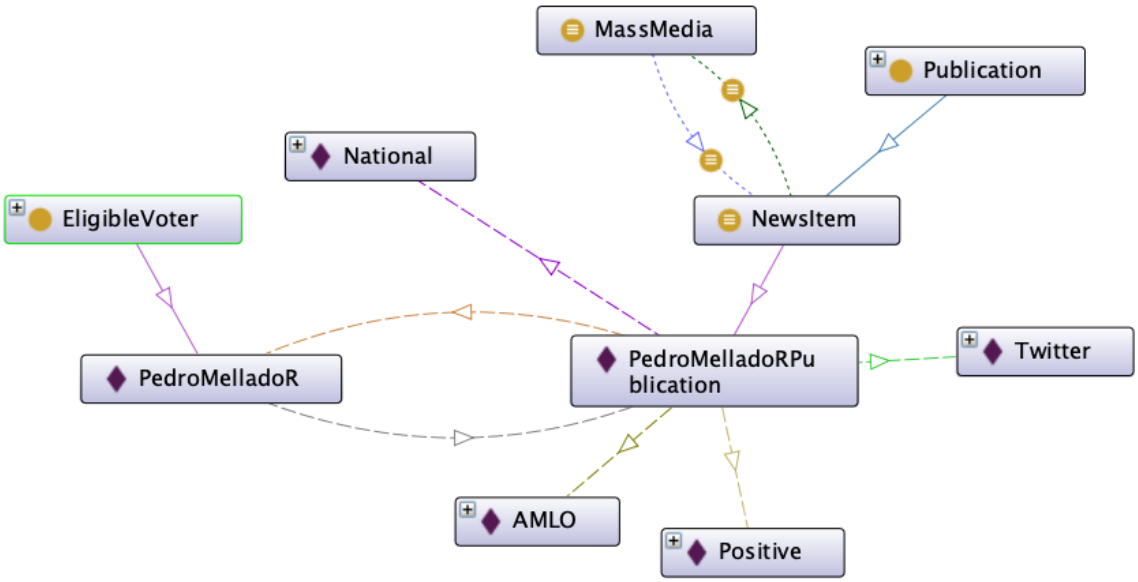



**Figura 56 Nueva instancia visualizada en OntoGraf**

# Capítulo 5. Validación del grafo de conocimiento

En los capítulos Capítulo 3 y Capítulo 4 incluimos procesos de validación y evaluación, desde la recuperación de información hasta la población ontológica. Implementamos razonadores como FACT++ y HermIT para evaluar la ocurrencia de las instancias, pero también, validamos la estructura del modelo ontológico a través de la herramienta OntOlogy Pitfall Scanner, todos con resultados favorables.

Es ahora necesario evaluar el grafo resultante por medio de su sintaxis y semántica. Un grafo de conocimiento de baja calidad produce aplicaciones de baja calidad, por lo tanto, la evaluación se convierte en una tarea necesaria para construir KG de alta calidad (Chen et al., 2019)(Gao et al., n.d.)(Wang et al., 2021).

Basado en la investigación (Chen et al., 2019), se evaluó el grafo de conocimiento a través de los 18 requisitos sobre la calidad relacionada a sus dimensiones (véase Tabla 8 y Figura 57):

| | | |
|---|---|---|
| 1. Los triples deben ser concisos.<br>2. La información contextual de las entidades debe ser capturada.<br>3. El grafo de conocimiento no contiene triples redundantes.<br>4. El grafo de conocimiento se puede actualizar dinámicamente.<br>5. Las entidades deben estar densamente | 8. Los datos para construir un grafo de conocimiento deben ser de diferentes tipos y de diferentes recursos.<br>9. Los sinónimos deben mapearse y las ambigüedades deben eliminarse para garantizar expresiones reconciliables.<br>10. El grafo de conocimiento debe organizarse en triples estructurados para que la máquina lo procese | 13. El grafo de conocimiento debe estar disponible públicamente y ser propietario.<br>14. El grafo de conocimiento debe ser autoridad.<br>15. El grafo de conocimiento debe estar concentrado.<br>16. Los triples no deben contradecirse entre sí.<br>17. Para tareas específicas de dominio, el grafo de |

| | | |
|---|---|---|
| conectadas.<br><br>6. Deben incluirse las relaciones entre los diferentes tipos de entidades.<br><br>7. La fuente de datos debe ser multicampo. | fácilmente.<br><br>11. La escalabilidad con respecto al tamaño de KG.<br><br>12. Los atributos de las entidades no deben perderse. | conocimiento debe estar relacionado con ese campo<br><br>18. El grafo de conocimiento debe contener los últimos recursos para garantizar la frescura. |

**Tabla 8 Requisitos de calidad KG**

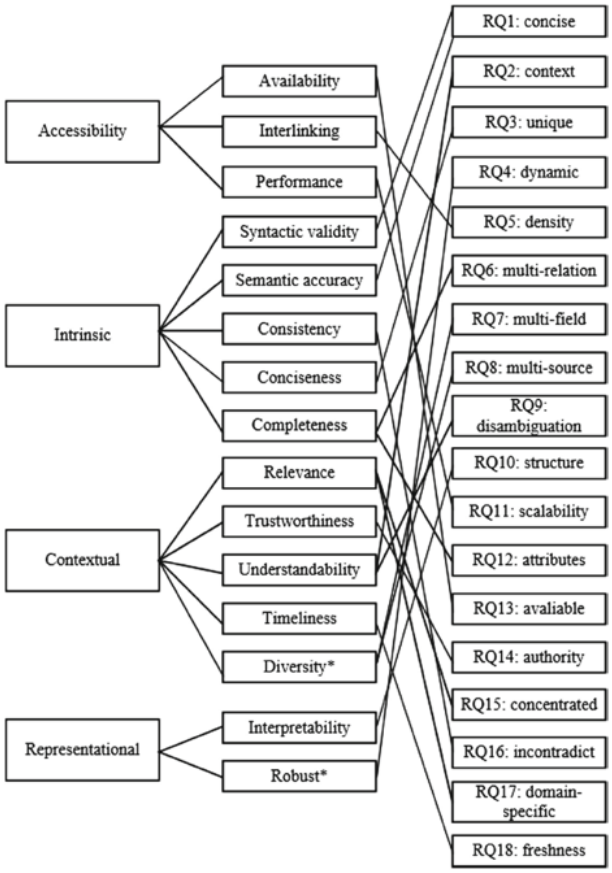



**Figura 57 Requisitos para las dimensiones de calidad**

La Tabla 9 reúne los números de requisito para ser evaluado en el grafo de conocimiento, cada celda que es rellenada con verde indica una respuesta positiva ante su cumplimiento, contrario a amarillo en parcial y rojo a incumplimiento:

| | | | | |
|---|---|---|---|---|
| 1 | 2 | 3 | 4 | 5 |
| 6 | 7 | 8 | 9 | 10 |
| 11 | 12 | 13 | 14 | 15 |
| 16 | 17 | 18 | | |

**Tabla 9 Respuestas a la evaluación del grafo**

En conclusión, la evaluación demostró que el grafo resultante es preciso, consistente, puntual, integro, confiable y disponible como dimensión de evaluación de la calidad del KG. Sin embargo, los puntos a mejorar son los relacionados a la densidad y el *deep learning* del grafo de conocimiento:

- (5) Las entidades deben estar densamente conectadas.
- (9) los sinónimos deben mapearse y las ambigüedades deben eliminarse para garantizar expresiones reconciliables.

# Capítulo 6. Conclusiones y trabajo a futuro

## 6.1 Conclusiones

En los últimos años, uno de los mayores retos principales del marketing político es el acceso y manejo de los datos masivos sociales y la heterogeneidad de las fuentes de datos, produciendo una tardía toma de decisiones, mayor tiempo de consumo en los procesos de gestión y extracción de la información, lo que también simboliza desconocimiento del electorado, costo humano y costo económico.

Para contribuir a la solución de los retos antes planteado, el trabajo de investigación descrito en esta tesis presenta un modelo ontológico y un sistema automático de la población del grafo de conocimiento.

La propuesta de modelo ontológico se construyó a través de la metodología estándar *Ontology 101* conforme a los criterios y necesidades del marketing político sobre las campañas políticas, por lo cual, la ontología se bautizó como PMont (*Political Marketing Ontology*), respondió de manera específica cada una de las preguntas clave del marketing político, iniciando con la más importante: ¿qué demanda el electorado?. La PMont permitió el almacenamiento semántico de la información disponible sobre el electorado y candidatos a través de la posibilidad de consumir los datos desde distintas fuentes de datos.

Para lograr tener una puerta de acceso a integrar nuevos datos al modelo ontológico, se creó una solución automatizada basada en textos en castellano a través de técnicas de ML y NLP, la cual brinda la posibilidad de extraer y coleccionar datos significativos de fuentes de medios digitales semiestructurados y no estructurados, procesamiento datos masivos y finalmente la población de un grafo de conocimiento previamente definido por un modelo ontológico del dominio de marketing político. La propuesta de sistema de automatización prevé los siguientes componentes para su optima aplicación: (i) conexión de fuente y recopilación de datos, (ii) procedimiento de extracción de información, (iii) proceso de población y validación de la ontología, y finalmente, (iv) el grafo de conocimiento resultante.

Como resultado final, se tiene un grafo de conocimiento poblado con información validada, precisa, consistente y confiable. El KG fue evaluado a través de 18 requisitos de calidad sobre las dimensiones de accesibilidad, contextualidad, intrínseca y representatividad, de los cuales aseguran la integración optima de conocimiento para ser utilizado como inteligencia de mercado del marketing político.

## 6.2 Aportaciones

Para puntualizar en las contribuciones y/o aportaciones de esta tesis doctoral y de acuerdo con lo discutido en la sección anterior, las principales aportaciones son las siguientes:

- **Modelo ontológico del dominio para marketing político**: la ontología resultante tiene una estructura única para brindar apoyo durante las campañas políticas en obtener y enriquecer con nuevo conocimiento accesible y reutilizable para sistemas informáticos. Por sus características y basado en el arduo estudio de la literatura, es propiamente el único modelo ontológico de su tipo.
- **Sistema automático de población del grafo de conocimiento**: un sistema basado en código abierto que provee de un conjunto de herramientas y técnicas para la población de un grafo de conocimiento de dominio político.
- **Grafo de conocimiento evaluado y accesible para sistemas informáticos**: grafo de conocimiento con triples evaluadas que pueden ser consumidas por otros sistemas informáticos, lo que simboliza la posibilidad de continuar el desarrollo de nuevos proyectos dentro de este ámbito.

## 6.3 Trabajo a futuro

En este trabajo hemos presentado desde el modelo ontológico hasta la población automática de un grafo de conocimiento para el marketing político, sin embargo, hay posibilidades de continuar el desarrollo y la gestión de este grafo de conocimiento resultante.

Podemos deducir las siguientes propuestas para el trabajo futuro a partir de esta tesis doctoral:

- La aplicación de técnicas de *Machine* y *Deep Learning* para el análisis semántico de la información para inferir nuevo conocimiento y tomar mejores decisiones durante las campañas políticas.
- Actualmente, el sistema funciona a través de una capa de base de datos previa a la inserción en la ontología, lo que evita que el procedimiento sea en tiempo real desde las redes sociales, sin duda, una de los retos y mejoras a enfrentar como trabajo futuro.
- GUI, diseñar e implementar una interfaz de usuario para el sistema, más allá de la línea de comandos.
- Desarrollo de una herramienta de visualización del grafo de conocimiento resultante específica para el marketing político. Esto es, ¿qué elementos e información podemos establecer como importantes para la incorporación y visualización de un *dashboard* de toma de decisiones?. Esta herramienta pudiera beneficiar al encargado de la campaña política en la rápida respuesta ante la demanda del electorado o en el manejo de crisis.

# Capítulo 7. Contribuciones científicas derivadas de la tesis doctoral

## 7.1 Publicaciones en revistas

1. H. H. Guedea Noriega and F. Garcia Sanchez, "Semantic (Big) Data Analysis: an Extensive Literature Review," *IEEE Lat. Am. Trans.*, vol. 17, no. 05, pp. 796–806, May 2019, doi: 10.1109/TLA.2019.8891948 (impact factor 2019: 0,782, Q4).
2. H. H. Guedea Noriega and F. Garcia Sanchez, "Integroly: Automatic knowledge graph population from Social Big Data in the Political Marketing Domain". Under review *Applied Sciences* (impact factor 2020: 2,679, Q2).

## 7.2 Publicaciones en congresos

1. H. H. Guedea-Noriega and F. García-Sánchez, "Construcción de una Ontología para Marketing Político," Tecnol. Educ., vol. 7, no. 1, pp. 38–44, 2020, Accessed: Sep. 07, 2019. [Online]. Available: https://terc.mx/index.php/terc/article/view/14.
2. C. Hernández-Castillo, H. H. Guedea-Noriega, M. Á. Rodríguez-García, and F. García-Sánchez, "Pest Recognition Using Natural Language Processing," *Commun. Comput. Inf. Sci.*, vol. 1124 CCIS, pp. 3–16, Dec. 2019, doi: 10.1007/978-3-030-34989-9_1.
3. H. H. Guedea-Noriega and F. García-Sánchez, "SePoMa: Semantic-Based Data Analysis for Political Marketing," *Commun. Comput. Inf. Sci.*, vol. 883, pp. 199–213, Nov. 2018, doi: 10.1007/978-3-030-00940-3_15.
4. F. García-Sánchez and H. H. Guedea-Noriega, "Intelligent Agents and Semantic Web Services: Friends or Foes?," *Commun. Comput. Inf. Sci.*, vol. 749, pp. 29–43, 2017, doi: 10.1007/978-3-319-67283-0_3.

# Reconocimiento

Esta tesis es parte del proyecto de I+D+i 20963/PI/18, financiado por la Comunidad Autónoma de la Región de Murcia a través de la convocatoria de Ayudas a proyectos para el desarrollo de investigación científica y técnica por grupos competitivos, incluida en el Programa Regional de Fomento de la Investigación (Plan de Actuación 2019) de la Fundación Séneca, Agencia de Ciencia y Tecnología de la Región de Murcia.

# Bibliografía

Achichi, M., Bellahsene, Z., Ienco, D., & Konstantin Todorov. (2015). Towards Linked Data Extraction From Tweets. *EGC: Extraction et Gestion Des Connaissances*, 383–388.

Acker, A., & Kreisberg, A. (2020). Social media data archives in an API-driven world. *Archival Science*, *20*(2), 105–123. https://doi.org/10.1007/s10502-019-09325-9

Airinei, D., & Berta, D. (2012). Semantic Business Intelligence - a New Generation of Business Intelligence. *Informatica Economica*, *16*(2), 72–80. http://search.proquest.com/docview/1030278611?accountid=15299%5Cnhttp://sfx.cbuc.cat/uoc?url_ver=Z39.88-2004&rft_val_fmt=info:ofi/fmt:kev:mtx:journal&genre=article&sid=ProQ:ProQ:abiglobal&atitle=Semantic+Business+Intelligence+-+a+New+Generation+of+Busines

Ait-Mlouk, A., Vu, X. S., & Jiang, L. (2020). Winfra: A web-based platform for semantic data retrieval and data analytics. *Mathematics*, *8*(11), 1–15. https://doi.org/10.3390/math8112090

Alonso Coto, M. A., & Adell, Á. (2011). *Marketing político 2.0: lo que todo candidato necesita saber para ganar las elecciones*. Gestión 2000. http://marketingpolitico20.planetadelibros.com/img/MK_politico.pdf

Anduiza, E., Cristancho, C., & Cantijoch, M. (2012). La exposición a información política a través de internet. *Arbor*, *188*(756), 673–688. https://doi.org/10.3989/arbor.2012.756n4004

Antoniades, N. (2021). Political Marketing Communications in Today's Era: Putting People at the Center. *Society 2020 57:6*, *57*(6), 646–656. https://doi.org/10.1007/S12115-020-00556-6

Asim, M. N., Wasim, M., Khan, M. U. G., Mahmood, W., & Abbasi, H. M. (2018). A survey of ontology learning techniques and applications. *Database*, *2018*(2018). https://doi.org/10.1093/database/bay101

Ayadi, A., Samet, A., De Beuvron, F. D. B., & Zanni-Merk, C. (2019). Ontology population with deep learning-based NLP: A case study on the Biomolecular Network Ontology. *Procedia Computer Science*, *159*, 572–581. https://doi.org/10.1016/j.procs.2019.09.212

Bao, Q., Wang, J., & Cheng, J. (2016). Research on Ontology Modeling of Steel Manufacturing Process Based on Big Data Analysis. *MATEC Web of Conferences*, *45*, 04005. https://doi.org/10.1051/matecconf/20164504005

Barceló Valenzuela, M., Sánchez Schmitz, G. G. A., & Perez-Soltero, A. (2006). La web semántica como apoyo a la gestión del conocimiento y al modelo organizacional. *Revista Ingeniería Informática*, *12*(abril), 1–14. https://www.researchgate.net/profile/Alonso_Perez-Soltero/publication/28109720_La_web_semantica_como_apoyo_a_la_gestion_del_conocimiento_y_al_modelo_organizacional/links/09e415074c143ad0a4000000.pdf

Barrasa, J., Hodler, A. E., & Webber, J. (2021). *Knowledge Graphs: Data in Context for Responsive Businesses* (First Edit). O'Reilly Media. https://neo4j.com/knowledge-graphs-data-in-context-for-responsive-businesses/

Bello-Orgaz, G., Jung, J. J., & Camacho, D. (2016). Social big data: Recent achievements and new challenges. *Information Fusion*, *28*(000), 45–59. https://doi.org/10.1016/j.inffus.2015.08.005

Bennett, M., & Baclawski, K. (2017). The role of ontologies in Linked Data, Big Data and Semantic Web applications. *Applied Ontology*, *12*(3–4), 189–194. https://doi.org/10.3233/AO-170185

Bereta, K., Papadakis, G., & Koubarakis, M. (2021). Ontop4theWeb: SPARQLing the Web On-the-fly. *Proceedings - 2021 IEEE 15th International Conference on Semantic Computing, ICSC 2021*, *February*, 268–271. https://doi.org/10.1109/ICSC50631.2021.00053

Berners-Lee, T., Hendler, J., & Lassila, O. (2001). The Semantic Web. *Scientific American*, *284*(5), 34–43. https://doi.org/10.1038/scientificamerican0501-34

Beydoun, G., Hoffmann, A., Breis, J. T. F., Béjar, R. M., Valencia-Garcia, R., & Aurum, A. (2005). Cooperative modelling evaluated. *International Journal of Cooperative Information Systems*, *14*(1), 45–71. https://doi.org/10.1142/S0218843005001080

Beyer, M., & Laney, D. (2012). The Importance of "Big Data": A Definition. *Gartner Publications*, *June*, 1–7. https://www.gartner.com/doc/2057415/importance-big-data-definition

Biblioteca del Congreso Nacional de Chile. (2012). *Datos Abiertos Enlazados*. http://datos.bcn.cl/es/informacion/que-es

Bihani, P., & Patil, S. (2014). A Comparative Study of Data Analysis Techniques. *International Journal of Emerging Trends & Technology in Computer Science (IJETTCS)*, *3*(2), 95–101. https://doi.org/10.13140/2.1.4255.0722

Bikakis, N., & Sellis, T. (2016). Exploration and Visualization in the Web of Big Linked Data: A Survey of the State of the Art. *6th International Workshop on Linked Web Data Management (LWDM 2016)*, 1–8. http://ceur-ws.org/Vol-1558/paper28.pdf

Bizer, C., Heath, T., & Berners-Lee, T. (2009). Linked Data - The Story So Far.

*International Journal on Semantic Web and Information Systems*, *5*(3), 1–22. https://doi.org/10.4018/jswis.2009081901

Borst, W. N. (1997). *Construction of Engineering Ontologies for Knowledge Sharing and Reuse* [University of Twente]. https://research.utwente.nl/en/publications/construction-of-engineering-ontologies-for-knowledge-sharing-and-

Brunetti, J. M., Auer, S., García, R., Klímek, J., & Nečaský, M. (2013). Formal Linked Data Visualization Model. *Proceedings of International Conference on Information Integration and Web-Based Applications & Services - IIWAS '13*, 309–318. https://doi.org/10.1145/2539150.2539162

Chan, J. O. (2014). Big Data Customer Knowledge Management. *Communications of the IIMA*, *14*(3), 45–56. http://scholarworks.lib.csusb.edu/ciima/vol14/iss3/5/?utm_source=scholarworks.lib.csusb.edu%2Fciima%2Fvol14%2Fiss3%2F5&utm_medium=PDF&utm_campaign=PDFCoverPages

Chaudhuri, S., Dayal, U., & Narasayya, V. (2011). An overview of business intelligence technology. *Communications of the ACM*, *54*(8), 88. https://doi.org/10.1145/1978542.1978562

Chen, H., Cao, G., Chen, J., & Ding, J. (2019). A Practical Framework for Evaluating the Quality of Knowledge Graph. In *Communications in Computer and Information Science: Vol. 1134 CCIS* (Issue September 2020). Springer Singapore. https://doi.org/10.1007/978-981-15-1956-7_10

Costa, L. (1994). *Manual de Marketing Político*. 1–78.

Di Iorio, A., & Rossi, D. (2018). Capturing and managing knowledge using social software and semantic web technologies. *Information Sciences*, *432*, 1–21. https://doi.org/10.1016/j.ins.2017.12.009

Eine, B., Jurisch, M., & Quint, W. (2017). Ontology-Based Big Data Management. *Systems*, *5*(3), 45. https://doi.org/10.3390/systems5030045

Faria, C., Serra, I., & Girardi, R. (2014). A domain-independent process for automatic ontology population from text. *Science of Computer Programming*, *95*(P1), 26–43. https://doi.org/10.1016/j.scico.2013.12.005

Fensel, D., Şimşek, U., Angele, K., Huaman, E., Kärle, E., Panasiuk, O., Toma, I., Umbrich, J., & Wahler, A. (2020). Introduction: What Is a Knowledge Graph? *Knowledge Graphs*, 1–10. https://doi.org/10.1007/978-3-030-37439-6_1

Fernandez, A. (2004). Investigación y tecnicas de mercado. In *Google Libros*. https://books.google.com.mx/books/about/Investigación_y_tecnicas_de_mercado.html

?id=6D8yCgAAQBAJ&printsec=frontcover&source=kp_read_button&hl=es-419&redir_esc=y#v=onepage&q&f=false

Freelon, D. (2018). Computational Research in the Post-API Age. *Political Communication*, *35*(4), 665–668. https://doi.org/10.1080/10584609.2018.1477506

Ganduri, R. N., Reddy, E. L., & Reddy, T. N. (2020). Social Media as a Marketing Tool for Political Purpose and Its Implications on Political Knowledge, Participation, and Interest. *International Journal of Online Marketing*, *10*(3), 21–33. https://doi.org/10.4018/ijom.2020070102

Gao, J., Li, X., Ethan Xu, Y., Sisman, B., Luna Dong, X., & Yang, J. (n.d.). *Efficient Knowledge Graph Accuracy Evaluation*.

García-Sánchez, F., Fernández-Breis, J. T., Valencia-García, R., Gómez, J. M., & Martínez-Béjar, R. (2008). Combining Semantic Web technologies with Multi-Agent Systems for integrated access to biological resources. *Journal of Biomedical Informatics*, *41*(5), 848–859. https://doi.org/10.1016/j.jbi.2008.05.007

García-Sánchez, F., & Guedea-Noriega, H. H. (2017). Intelligent Agents and Semantic Web Services: Friends or Foes? *Communications in Computer and Information Science*, *749*, 29–43. https://doi.org/10.1007/978-3-319-67283-0_3

García Sánchez, F. (2007). *Sistema basado en tecnologías del conocimiento para entornos de servicios web semánticos* [Universidad de Murcia]. https://www.tesisenred.net/handle/10803/10924;jsessionid=94066507D5B5423405CDE3889B3C6EDB#page=1

Gómez, J. G. (2015). *Marketing Político* [Universidad de Valladalid]. file:///Users/hector/Downloads/Marketing Político - Gómez, Javier García.pdf

Gómez Vieites, A., & Suárez Rey, C. (2011). *Sistemas de información : herramientas prácticas para la gestión empresarial*. RA-MA. http://www.ra-ma.es/libros/SISTEMAS-DE-INFORMACION-HERRAMIENTAS-PRACTICAS-PARA-LA-GESTION-EMPRESARIAL-4-EDICION-AMPLIADA/66891/978-84-9964-122-5

Gruber, T. R. (1993). A translation approach to portable ontology specifications. *Knowledge Acquisition*, *5*(2), 199–220. https://doi.org/10.1006/KNAC.1993.1008

Guarino, N. (1998). Formal Ontology and Information Systems. *Formal Ontology in Information Systems: Proceedings of the 1st International Conference*, *46*(June), 3–15. https://doi.org/10.1.1.29.1776

Guedea-Noriega, H. H., & García-Sánchez, F. (2018). SePoMa: Semantic-Based Data Analysis for Political Marketing. *Communications in Computer and Information Science*, *883*, 199–213. https://doi.org/10.1007/978-3-030-00940-3_15

Guedea-Noriega, H. H., & García-Sánchez, F. (2020). Construcción de una Ontología para Marketing Político. *Tecnología Educativa*, *7*(1), 38–44. https://terc.mx/index.php/terc/article/view/14

Guedea Noriega, Hector H., & Garcia Sanchez, F. (2019). Semantic (Big) Data Analysis: an Extensive Literature Review. *IEEE Latin America Transactions*, *17*(05), 796–806. https://doi.org/10.1109/TLA.2019.8891948

Guedea Noriega, Héctor Hiram, Santana, P. C., & Flores Cortes, C. (2011). La contribución del social media en campañas políticas. *Conference: 4to Congreso Internacional En Ciencias Computacionales, CiComp 2011*. https://www.researchgate.net/publication/329192112_La_contribucion_del_social_media_en_campanas_politicas

Guzmán-Guzmán, X., Núñez-Valdez, E. R., Vásquez-Reynoso, R., Asencio, A., & García-Díaz, V. (2021). SWQL: A new domain-specific language for mining the social Web. *Science of Computer Programming*, *207*, 102642. https://doi.org/10.1016/j.scico.2021.102642

Harfoush, R. (2010). *Yes We Did: Cómo construimos la marca Obama a través de las redes sociales - Rahaf Harfoush - Google Libros*. Ediciones Gestión 2000. https://books.google.com.mx/books?id=-s_dqYG21PQC&printsec=copyright&redir_esc=y#v=onepage&q&f=false

Hoppe, T., Humm, B., & Reibold, A. (2018). *Semantic Applications* (T. Hoppe, B. Humm, & A. Reibold (eds.)). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-662-55433-3

Hu, B., Carvalho, N., & Matsutsuka, T. (2013). Towards Big Linked Data: A Large-Scale, Distributed Semantic Data Storage. *International Journal of Data Warehousing and Mining*, *9*(4), 19–43. https://doi.org/10.4018/ijdwm.2013100102

Jain, A., Kumar, A., & Dash, M. K. (2015). Information technology revolution and transition marketing strategies of political parties: analysis through AHP. *International Journal of Business Information Systems*, *20*(1). https://doi.org/10.1504/IJBIS.2015.070903

Juárez, J. (2003). Hacia un estudio del marketing político: limitaciones teóricas y metodológicas. *Espiral*, *IX*(27), 60–95. http://www.redalyc.org/pdf/138/13802703.pdf

Jung, H., Park, H.-A., & Song, T.-M. (2017). Ontology-Based Approach to Social Data Sentiment Analysis: Detection of Adolescent Depression Signals. *Journal of Medical Internet Research*, *19*(7). https://doi.org/10.2196/jmir.7452

Kale, S. A., & Dandge, S. S. (2014). Understanding the Big Data Problems and Their Solutions Using Hadoop and Map-Reduce. *International Journal of Application or Innovation in Engineering & Management (IJAIEM)*, *3*(3), 439–445.

Kertkeidkachorn, N., & Ichise, R. (2018). T2KG: A demonstration of knowledge graph population from text and its challenges. *CEUR Workshop Proceedings*, *2293*, 110–113.

Kim, A. R., Park, H.-A., & Song, T.-M. (2017). Development and Evaluation of an Obesity Ontology for Social Big Data Analysis. *Healthcare Informatics Research*, *23*(3), 159. https://doi.org/10.4258/hir.2017.23.3.159

Konys, A. (2016). A Framework for Analysis of Ontology-Based Data Access. In N. T. Nguyen, G. A. Papadopoulos, P. Jędrzejowicz, B. Trawiński, & G. Vossen (Eds.), *International Conference on Computational Collective Intelligence (ICCCI 2016)* (Vol. 9876, pp. 397–408). Springer, Cham. https://doi.org/10.1007/978-3-319-45246-3_38

Kotler, P., & Armstrong, G. (2013). Fundamentos de Marketing. In *Entelequia: revista interdisciplinar* (11 edición, Vol. 4, Issue 3). Pearson Educación. https://frrq.cvg.utn.edu.ar/pluginfile.php/14584/mod_resource/content/1/Fundamentos del Marketing-Kotler.pdf

Kupershmidt, I., Su, Q. J., Grewal, A., Sundaresh, S., Halperin, I., Flynn, J., Shekar, M., Wang, H., Park, J., Cui, W., Wall, G. D., Wisotzkey, R., Alag, S., Akhtari, S., & Ronaghi, M. (2010). Ontology-Based Meta-Analysis of Global Collections of High-Throughput Public Data. *PLoS ONE*, *5*(9), e13066. https://doi.org/10.1371/journal.pone.0013066

Kureychik, V., & Semenova, A. (2017). Combined Method for Integration of Heterogeneous Ontology Models for Big Data Processing and Analysis. In R. Silhavy, R. Senkerik, Z. Kominkova Oplatkova, Z. Prokopova, & P. Silhavy (Eds.), *Artificial Intelligence Trends in Intelligent Systems* (Vol. 573, pp. 302–311). Springer, Cham. https://doi.org/10.1007/978-3-319-57261-1_30

Laney, D. (2001). *3D Data Management: Controlling Data Volume, Velocity, and Variety*. Application Delivery Strategies. https://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf

Lubani, M., Noah, S. A. M., & Mahmud, R. (2019). Ontology population: Approaches and design aspects. *Journal of Information Science*, *45*(4), 502–515. https://doi.org/10.1177/0165551518801819

Maarek, P. J. (2011). *Campaign Communication and Political Marketing*. Wiley-Blackwell. https://books.google.com.mx/books/about/Campaign_Communication_and_Political_Mar.html?id=S3K1RiqDs2kC&redir_esc=y

Maté Jiménez, C. (2014). Big data. Un nuevo paradigma de análisis de datos. *Revista: Anales de Mecánica y Electricidad*, *41*(6), 10–16.

Mayer-Schönberger, V., & Cukier, K. (2013). *Big data : a revolution that will transform how we live, work, and think*. Houghton Mifflin Harcourt, Boston, Massachusetts. https://books.google.com.mx/books/about/Big_Data.html?id=HpHcGAkFEjkC&source=kp_cover&redir_esc=y

McGuinness, D. L., & van Harmelen, F. (2004). *OWL Web Ontology Language - Overview*. W3C. https://www.w3.org/TR/2004/REC-owl-features-20040210/

Molina López, J. M., & García Herrero, J. (2006). *Técnicas de análisis de datos: Aplicaciones prácticas utilizando Microsoft Excel y Weka*. http://ucua.ujaen.es/jnavas/web_recursos/archivos/weka master recursos naturales/apuntesAD.pdf

Moore, C. (2010). *Propaganda Prints: A history of art in the service of social and political change*. A&C Black. https://books.google.com.mx/books?hl=es&lr=&id=l0KAlsQ2Yj4C&oi=fnd&pg=PP1&dq=propaganda+%2B+history+%2B+political+&ots=QVG-1q2TEq&sig=FuO3CZL0FkNMLv5qOMv80fwgXys&redir_esc=y#v=onepage&q&f=false

Mozilla. (2022). *HTML5 - Glosario | MDN*. https://developer.mozilla.org/es/docs/Glossary/HTML5

Nafría, I. (2008). *Web 2.0: el usuario, el nuevo rey de Internet* . Ediciones Gestión 2000. https://books.google.com.mx/books/about/Web_2_0.html?id=1fZi_ndyc-wC&printsec=frontcover&source=kp_read_button&hl=es-419&redir_esc=y#v=onepage&q&f=false

Neches, R., Fikes, R., Finin, T., Gruber, T., Patil, R., Senator, T., & Swartout, W. R. (1991). Enabling technology for knowledge sharing. *AI Magazine*, *12*(3), 36–56.

Neuböck, T., Neumayr, B., Schrefl, M., & Schütz, C. (2014). Ontology-Driven Business Intelligence for Comparative Data Analysis. In E. Zimányi (Ed.), *Business Intelligence: Third European Summer School, eBISS 2013, Dagstuhl Castle, Germany, July 7-12, 2013, Tutorial Lectures* (pp. 77–120). Springer, Cham. https://doi.org/10.1007/978-3-319-05461-2_3

Noy, N. F., & Mcguinness, D. L. (2001). *Ontology development 101: A guide to creating your first ontology*. https://protege.stanford.edu/publications/ontology_development/ontology101.pdf

Noy, N., & Paulheim, H. (2016). Knowledge Graph Refinement: A Survey of Approaches and Evaluation Methods. *Semantic Web*, *0*, 1–0.

Nuzzolese, A. G., Presutti, V., Gangemi, A., Peroni, S., & Ciancarini, P. (2017). Aemoo: Linked Data exploration based on Knowledge Patterns. *Semantic Web*, *8*(1), 87–112. https://doi.org/10.3233/SW-160222

Ontotext. (2020). *What is a Knowledge Graph? | Ontotext Fundamentals*. https://www.ontotext.com/knowledgehub/fundamentals/what-is-a-knowledge-graph/

Paredes Valverde, M. A. (2017). *Interfaces del lenguaje natural para la consulta y recuperación de información de bases de conocimiento basadas en ontologías* [Universidad de Murcia]. https://www.tdx.cat/handle/10803/405396#page=1

Pech, F., Martinez, A., Estrada, H., & Hernandez, Y. (2017). Semantic Annotation of Unstructured Documents Using Concepts Similarity. *Scientific Programming*, *2017*. https://doi.org/10.1155/2017/7831897

Petasis, G., Karkaletsis, V., Paliouras, G., Krithara, A., & Zavitsanos, E. (2011). Ontology population and enrichment: State of the art. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *6050*, 134–166. https://doi.org/10.1007/978-3-642-20795-2_6

Pinto, F. M., Marques, A., & Santos, M. F. (2009). Ontology-supported database marketing. *Journal of Database Marketing and Customer Strategy Management*, *16*(2), 76–91. https://doi.org/10.1057/dbm.2009.9

Rahoman, M.-M., & Ichise, R. (2018). A proposal of a temporal semantics aware linked data information retrieval framework. *Journal of Intelligent Information Systems*, *50*(3), 573–595. https://doi.org/10.1007/s10844-017-0483-2

Real Academia Española. (2021). *Ontología*. https://dle.rae.es/ontología?m=30_2

Rodríguez-García, M. Á., & Hoehndorf, R. (2018). Inferring ontology graph structures using OWL reasoning. *BMC Bioinformatics*, *19*(1), 7. https://doi.org/10.1186/s12859-017-1999-8

Rodríguez García, M. Á. (2014). Extracción semántica de información basada en evolución de ontologías. *Proyecto de Investigación:* https://digitum.um.es/digitum/handle/10201/41246

Russel, S., & Norvig, P. (2009). *Artificial Intelligence: A Modern Approach* (3rd.). Prentice Hall Press. https://dl.acm.org/citation.cfm?id=1671238

Santipantakis, G., Kotis, K., & Vouros, G. A. (2017). OBDAIR: Ontology-Based Distributed framework for Accessing, Integrating and Reasoning with data in disparate data sources. *Expert Systems with Applications*, *90*, 464–483. https://doi.org/10.1016/J.ESWA.2017.08.031

Sapountzi, A., & Psannis, K. E. (2018). Social networking data analysis tools & challenges. *Future Generation Computer Systems*, *86*, 893–913. https://doi.org/10.1016/j.future.2016.10.019

Scammell, M. (1999). Political Marketing: Lessons for Political Science. *Political Studies*,

*47*(4), 718–739. https://doi.org/10.1111/1467-9248.00228

Shadbolt, N., Berners-Lee, T., & Hall, W. (2006). The Semantic Web Revisited. *IEEE Intelligent Systems*, *21*(3), 96–101. https://doi.org/10.1109/MIS.2006.62

Shah, T., Rabhi, F., & Ray, P. (2015). Investigating an ontology-based approach for Big Data analysis of inter-dependent medical and oral health conditions. *Cluster Computing*, *18*(1), 351–367. https://doi.org/10.1007/s10586-014-0406-8

Sivarajah, U., Kamal, M. M., Irani, Z., & Weerakkody, V. (2017). Critical analysis of Big Data challenges and analytical methods. *Journal of Business Research*, *70*, 263–286. https://doi.org/10.1016/j.jbusres.2016.08.001

Somodevilla García, M., Vilariño Ayala, D., & Pineda, I. (2018). An overview of ontology learning tasks. In *Computacion y Sistemas* (Vol. 22, Issue 1, pp. 137–146). Instituto Politecnico Nacional. https://doi.org/10.13053/CyS-22-1-2790

Stardog. (2022). *RDF Graph Data Model*. https://docs.stardog.com/tutorials/rdf-graph-data-model

Stats, I. W. (2021). *World Internet Users Statistics*. https://www.internetworldstats.com/stats.htm

Steve, G., Gangemi, A., & Pisanelli, D. M. (1997). Integrating medical terminologies with ONIONS methodology. *INFORMATION MODELLING AND KNOWLEDGE BASES VIII (IOS*. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.106.7521

Studer, R., Benjamins, V. R., & Fensel, D. (1998). Knowledge engineering: Principles and methods. *Data & Knowledge Engineering*, *25*(1–2), 161–197. https://doi.org/10.1016/S0169-023X(97)00056-6

Tessaris, S. (2009). *Reasoning web : semantic technologies for information systems : 5th International Summer School 2009, Brixen-Bressanone, Italy, August 30-September 4, 2009 : tutorial lectures*. Springer. https://books.google.com.mx/books?id=JdyeU7zs4-AC&dq=%22DIG+QuOnto%22&source=gbs_navlinks_s

Valdez Zepeda, A. (2010). Las campañas electorales en la nueva sociedad de la información y el conocimiento. *Estudios Políticos*, *20*, 155–165. http://www.scielo.org.mx/pdf/ep/n20/n20a9.pdf

van Atteveldt, W., Schlobach, S., & van Harmelen, F. (2007). Media, Politics and the Semantic Web: An experience report in advanced RDF usage. *Proceedings of the Third European Semantic Web Conference*, *4519*(4519), 205–219. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.112.9564

Vaqué, A. (2014). *Datos en formato Grafo (Parte II - Web Semántica o Linked Data)*. https://www.cloudadmins.org/datos-en-formato-grafo-parte-ii-web-semantica-o-

linked-data/

Vargas-Vera, M., Moreale, E., Stutt, A., Motta, E., & Ciravegna, F. (2007). MnM: Semi-Automatic Ontology Population from Text. In *Ontologies* (pp. 373–402). Springer US. https://doi.org/10.1007/978-0-387-37022-4_13

W3C. (2012). *OWL 2 Web Ontology Language Document Overview (Second Edition)*. https://www.w3.org/TR/owl2-overview/

W3C. (2014). *RDF - Semantic Web Standards*. https://www.w3.org/RDF/

W3C. (2022). *Linked Data*. https://www.w3.org/standards/semanticweb/data

Wang, X., Chen, L., Ban, T., Usman, M., Guan, Y., Liu, S., Wu, T., & Chen, H. (2021). Knowledge graph quality control: A survey. *Fundamental Research*, *1*(5), 607–626. https://doi.org/10.1016/j.fmre.2021.09.003

Wimalasuriya, D. C., & Dou, D. (2010). Ontology-based information extraction: An introduction and a survey of current approaches. *Journal of Information Science*, *36*(3), 306–323. https://doi.org/10.1177/0165551509360123

Witten, I. H., Frank, E., & Hall, M. a. (2005). Data Mining: Practical Machine Learning Tools and Techniques (Second Edition). In *Elsevier* (Second). Elsevier. https://books.google.com.mx/books/about/Data_Mining.html?id=QTnOcZJzlUoC&redir_esc=y

Wongthontham, P., & Abu-Salih, B. (2018, January 4). *Ontology-based Approach for Identifying the Credibility Domain in Social Big Data*. Cornell University Library (ArXiv). http://arxiv.org/abs/1801.01624

Xu, J., Kim, S., Song, M., Jeong, M., Kim, D., Kang, J., Rousseau, J. F., Li, X., Xu, W., Torvik, V. I., Bu, Y., Chen, C., Ebeid, I. A., Li, D., & Ding, Y. (2020). Building a PubMed knowledge graph. *Scientific Data 2020 7:1*, *7*(1), 1–15. https://doi.org/10.1038/s41597-020-0543-2

Yoo, S., & Jeong, O. (2020). Automating the expansion of a knowledge graph. *Expert Systems with Applications*, *141*, 112965. https://doi.org/10.1016/j.eswa.2019.112965

Zamudio, Y. (2015). *Etapas del plan de marketing político*. http://roa.uveg.edu.mx/repositorio/postgrado2015/37/Etapasdelplandemarketing.pdf

Zuiderveen Borgesius, F. J., Möller, J., Kruikemeier, S., Ó Fathaigh, R., Irion, K., Dobber, T., Bodo, B., & De Vreese, C. (2018). Online Political Microtargeting: Promises and Threats for Democracy. *Utrecht Law Review*, *14*(1), 82. https://doi.org/10.18352/ulr.4202