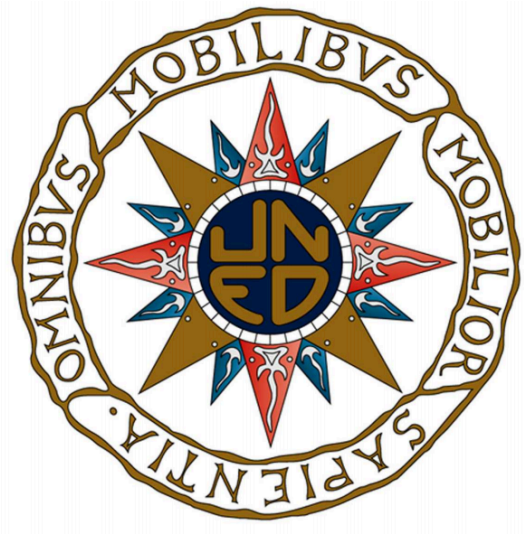


Estrategias para la creación de un sistema con Bases del Conocimiento (Wikidata y Wikipedia) dirigido a la conceptualización y el aprendizaje.

Trabajo de Fin de Máster

UNED

E.T.S. de Ingeniería Informática
Departamento de Inteligencia Artificial
Métodos en Inteligencia Artificial Avanzada: Fundamentos, Métodos y Aplicaciones.

Alumna: Patricia Mayo Tejedor

Director: José Luis Fernández Vindel

Septiembre 2018, Madrid, España

Resumen

Las grandes bases de conocimiento disponibles en la Web Semántica son lo suficientemente maduras y robustas como para usarse de base fundamental única en sistemas de Aprendizaje o de Inteligencia Artificial. Sin embargo, el acceso a este conocimiento necesita de usuarios que comprendan los lenguajes de consulta semántica, por lo que la información queda oculta en tales repositorios. Además, la cantidad de datos no relevantes pero enlazados crea ruido innecesario que dificulta las tareas de la Inteligencia Artificial. **En este trabajo se presenta una propuesta estratégica de etapas y sintonización de parámetros para la creación de un sistema interactivo dirigido al aprendizaje del usuario y al uso de la Web Semántica en aplicaciones de la Inteligencia Artificial.** En este sistema híbrido se han integrando datos estructurados y no estructurados de Wikidata y Wikipedia, se han filtrado los datos irrelevantes, y se han devuelto los resultados como paquete de datos en forma de grafo personalizable. Todo el sistema se ha basado en el diseño y testeo de unas estrategias de parametrización que garantizan el balance entre la cantidad de datos (el potencial enorme de la Web Semántica) y la eficacia (asegurando la relevancia de los datos incluso después del filtrado). Se ha demostrado lo trascendental que resulta la plataforma no sólo para los usuarios que desean hacer una consulta, aprender sobre un concepto, o que necesitan de un auxiliar en sistemas de autorías propios, sino que también para su utilidad en aplicaciones de Inteligencia Artificial como datos fundamentales que alimentan al sistema.

Abstract

The large knowledge bases available in the Semantic Web are mature and robust enough to be used as a fundamental basis in Learning or Artificial Intelligence systems. However, access to this knowledge requires users who understand semantic query languages, so that information is hidden in the repositories of stories. In addition, the amount of data not relevant but linked creates unnecessary creativity that hinders the tasks of Artificial Intelligence. ***This paper presents a strategic proposal of stages and parameters for the creation of an interactive system aimed at user learning and using the Semantic Web in applications of Artificial Intelligence.*** *In this hybrid system, the structured and unclassified data of Wikidata and Wikipedia have been integrated, the irrelevant data has been filtered, and the results have been returned as a data package in the form of a customizable graph. The whole system has been based on the design and the state of the parameterization strategies that guarantee the balance between the amount of data (the enormous potential of the Semantic Web) and the effectiveness of data security. We have found the transcendental nature of the platform not only for users who want to consult, learn about a concept, or who need an assistant in their own authoring systems, but also for its use in Artificial Intelligence applications as fundamental data. that feed the system.*

Agradecimientos

Quisiera dedicar todo el esfuerzo y trabajo a mi madre, mi hermana y a mi novio, por todo su apoyo incondicional y por todo el amor y confianza que me han demostrado.

A mis compañeros de piso y de trabajo, por todos los debates, discusiones y distintos puntos de vista que me han hecho abordar la investigación con entusiasmo y abierto el camino a nuevas ideas.

A Carlos Arancón del Valle, que ya ha pasado por este proceso, acabando su trabajo el año pasado, y que me ha ayudado sin tan siquiera conocerme.

A la UNED y profesores, por permitirme realizar este máster a distancia, compaginando así mi trabajo desde IBM Ámsterdam y mis ganas por seguir aprendiendo.

Y por último, y no por ello menos importante, al que estaré siempre agradecida, es a mi tutor y director de este trabajo, José Luis Fernández Vindel, que me ha sabido transmitir la pasión por esta investigación, y que con su paciencia infinita, me ha guiado y animado durante este camino, en el que no han faltado momentos de estrés y de incertidumbre. Espero que estés orgulloso ¡Muchas gracias José Luis!

Índice de Contenidos

Capítulo 1:
Motivación y Objetivos

En la actualidad existen ingentes cantidades de datos guardados y vinculados en bases de conocimiento. Este conocimiento es accesible a través de la Web Semántica, que proporciona un marco común que permite que los datos se compartan y reutilicen a través de los límites de la aplicación, la empresa y la comunidad. A fecha de Julio de 2018, estas grandes bases de conocimiento son suficientemente maduras y robustas como para haber parado de crecer añadiendo contenido nuevo, centrándose más en expandir y mejorar el ya existente[1]. El acceso directo a tal cantidad de datos requiere una comprensión de los lenguajes semánticos de consulta y los conjuntos de datos específicos. El desafío clave en el área de la Web Semántica es proporcionar un fácil acceso a los usuarios a tales datos 'ocultos' en los repositorios de bases del conocimiento. Este desafío se complica además por la cantidad ingente de datos existentes y quizás no tan relevantes para su uso práctico (identificadores, referencias a otras bases, desambiguaciones…). Tal cantidad de datos enlazados, pero irrelevantes, crea ruido innecesario que dificulta no sólo las tareas de la Inteligencia Artificial, sino el acceso a los usuarios, que se ven abrumados. En este trabajo se propone un **sistema mediador entre un agente humano o la aplicación inteligente y dichas bases de conocimiento**. Para el desarrollo de tal sistema se presenta además una **propuesta estratégica de etapas y sintonización de parámetros** que garantizan el balance entre la cantidad de datos (el potencial enorme de la Web Semántica) y la eficacia (asegurando la relevancia de los datos incluso después del filtrado). Se pretende solventar el desafío de la Web Semántica, teniendo en cuenta la madurez del estado de la tecnología de datos enlazados que asienta las bases y posibilita la creación de este tipo de asistente personal para la búsqueda de información y el aprendizaje.

Este sistema mediador debe dejar escoger un concepto o semilla al usuario que desea hacer la consulta, y acompañarle a través del proceso de conceptualización, descubrimiento y contextualización, para que en poco tiempo se haya producido un desarrollo cognitivo de aprendizaje en el usuarios y recopilado un dossier o portafolio de conceptos relacionados entre sí. Para que esto sea posible se va a basar el proyecto en unas bases fundamentales: (i) La teoría de los campos conceptuales de Vergnaud, que supone que el amago del desarrollo cognitivo es la conceptualización y que realza la importancia de la creación de una red de conceptos para tal desarrollo, (ii) La visualización y la representación como componentes vitales del proceso analítico y de aprendizaje y (iii) Las Bases de Conocimiento abiertas como fuentes fiables del que alimentar un sistema. Para proporcionar mayor valor informativo al usuario, se pretende además integrar dos fuentes distintas de datos, una estructurada y otra no.

[1] https://en.wikipedia.org/wiki/Wikipedia:Modelling_Wikipedia%27s_growth#Old_exponential_model_for_article_count_of_Wikipedia

Los datos estructurados son convenientes para mostrar relaciones de clases, herencia o partes de un concepto. Los datos no estructurados en forma de textos descriptivos a menudo aportan mayor contextualización. Pasando por un filtrado exhaustivo de los conceptos y datos que mostrar, y usando técnica de Procesamiento del Lenguaje Natural y Clustering, se devolverá al usuario una versión lo más simplificada y relevante posible, que evitarán sentimientos de abrumación o confusión, sin perder el sentido principal del concepto o semilla. La estructura o paquete de datos resultante además será usada y probada en un sistema inteligente, demostrando así la fiabilidad del uso de las bases de conocimiento abiertas, y la necesidad y eficacia del filtrado de los datos para la eliminación del ruido.
Para realizar el sistema se ha llevado a cabo un estudio profundo sobre la Web Semántica y sus funcionalidades, sobre los Datos Enlazados y las Bases de Conocimiento, sobre los métodos de aprendizaje de las personas, sobre cómo se produce el desarrollo cognitivo y sobre la representación del conocimiento. Este estudio no es sólo necesario para el aprendizaje y desarrollo del trabajo, sino que permite valorar el grado de madurez y la utilidad de todas las herramientas y tecnologías escogidas para el proyecto.
Cabe recalcar que en esta memoria, cuando se haga referencia a 'aprendizaje', se estará hablando del aprendizaje en los seres humanos, y no al aprendizaje de la Inteligencia Artificial, conocido como 'Machine Learning'.

1.1. Objetivos e hipótesis

En este trabajo se va a crear un sistema mediador entre un agente humano o una aplicación inteligente y las bases de conocimiento de Wikidata y Wikipedia. El sistema de descubrimiento estará dirigido a la conceptualización y el aprendizaje en el enfoque al usuario. A su vez en el enfoque para la Inteligencia Artificial se centrará en la creación de un conjunto de datos fiable, filtrado y sin atributos irrelevantes, que aporte contextualización a un concepto. Para ello se presentará además una propuesta estratégica de etapas y sintonización de parámetros (confirmados posteriormente en el apartado de experimentos) que garantizan el balance entre la cantidad de datos y la eficacia del sistema, en términos de relevancia de los conceptos mostrados. Como resultado final no sólo se devolverá la representación y visualización de los conceptos y ramas relacionadas con el concepto-semilla del usuario, sino que también se proporcionará un dossier o portfolio de conceptos, donde tal representación será transformada en una estructura de datos capaz de ser exportada/importada en diferentes entornos de aprendizaje. Este sistema mediador se basa en varios supuestos:

- La visualización estructurada haciendo hincapié en las relaciones permite crear un conocimiento más profundo sobre un concepto dado. Además de que la navegación y la interacción de forma visual con datos favorece al descubrimiento y afianza el aprendizaje en el usuario.
- La recomendación de datos incrementales a partir de una semilla favorece y crea las bases para un sistema de autoría propio.

- El filtrado automático, así como la entrada manual del usuario son útiles para asentar las bases de un sistema de recomendación o clasificación inteligente.

A continuación se enumeran los objetivos de manera resumida:
- El estudio del desarrollo cognitivo en los seres humanos.
- El estudio y valoración de la Web Semántica y sus tecnologías (Datos Enlazados, RDF, SPARQL y Bases de Conocimiento con especial enfoque en Wikipedia y Wikidata) en aplicaciones dirigidas al aprendizaje en usuarios y en sistemas inteligentes como fuente fiable de datos.
- El estudio de las formas de representación del conocimiento, tanto visualmente como en términos de programación.
- El estudio y la integración de datos estructurados y no estructurados en un mismo sistema híbrido y polivalente.
- La definición y el desarrollo de las estrategias de aprendizaje en las personas mediante fases.
- La sintonización de parámetros que garanticen la relevancia de datos.
- La creación de un sistema fundamentado en Bases de Conocimiento y dirigido a la conceptualización y el aprendizaje, donde el usuario o agente inteligente tras introducir un concepto como semilla reciba un paquete de datos o dossier con un conjunto de conceptos relacionados a la semilla.

1.2. Alcance y aplicaciones

Llevar a cabo los objetivos de una manera eficaz se podrá traducir en una serie de casos de uso que van, desde las valoraciones y conclusiones de los estudios llevados a cabo, hasta las posibles aplicaciones del sistema desarrollado.

Tras los estudios y valoraciones sobre la Web Semántica se espera reafirmar su uso como fuente útil y fiable de datos. Se espera que este estudio apoye, afiance y generalice el uso de la Web Semántica. Integrando este conocimiento con los estudios sobre el desarrollo cognitivo, el desarrollo de las estrategias de aprendizaje de usuarios y la sintonización de parámetro, se crea un proceso o metodología a seguir que pudiese servir de referencia a trabajos futuros centrados en el aprendizaje.
Este trabajo facilita el proceso de consulta de conceptos de manera interactiva y poniendo al alcance del usuario la gran cantidad de datos 'escondidos' en la Web Semántica, por lo que resultará útil como herramienta dentro de un **Entorno Personal de Aprendizaje** (conjunto de elementos como recursos, actividades y fuentes de información utilizados para la gestión del aprendizaje personal. En inglés: Personal Learning Environment, PLE**)**. De la misma manera podrá ser usado en **sistemas de autorías propios**, donde el usuario pretenda crear contenido y necesite de una guía o de recomendaciones de conceptos que deben ser incluidos en su tema. La aplicación pues podría ser usada en los entornos virtuales de la UNED, tanto enfocado a alumnos como a profesores.
El paquete de datos o dossier resultante además resulta de utilidad para otros múltiples usos en **aplicaciones inteligentes donde la contextualización sea**

importante, como son los sistemas de recomendación en frío, descubrimiento de información en textos, inferencia o pertenencia a temáticas, clusterización… Esto quedará probado en la parte de experimentación del trabajo.

1.3. Estudio de la cuestión

Madurez de la tecnología

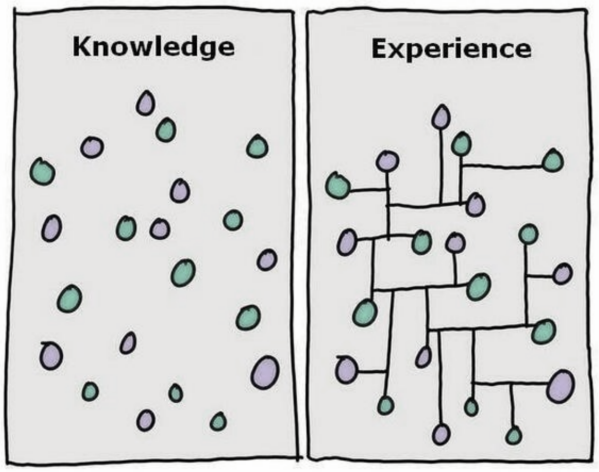
A fecha de Julio de 2018, las grandes bases de conocimiento abiertas y disponibles en la Web Semántica son suficientemente maduras y robustas como para haber parado de crecer añadiendo contenido nuevo, centrándose más en expandir y mejorar el ya existente. Se pone de ejemplo el caso de Wikipedia, el trabajo de referencia general y la base de conocimiento más grande y popular en Internet, con 5.685.718 artículos. La tasa de nuevos artículos dentro de la Wikipedia en inglés creció exponencialmente hasta alrededor de 2007, pero este ya no es el caso, la tasa está disminuyendo muy lentamente a un ratio de 60000 artículos menos cada año. Los dos modelos de crecimiento más probables para el futuro de Wikipedia son un modelo de función de Gompertz que predice que la creación de artículos se acercará asintóticamente a cero, y un modelo de Gompertz modificado que predice que el crecimiento continuará indefinidamente, pero a un ritmo significativamente menor que en los primeros días de Wikipedia[2]. Por otro lado, la cantidad total de texto en los artículos de Wikipedia se ha mantenido básicamente de forma lineal, y la tasa de crecimiento se mantuvo prácticamente sin cambios desde 2006. Esto implica que la contribución a Wikipedia no se desvanece con el tiempo, pero el trabajo está en expandir artículos existentes o incluso fusionar artículos que son similares en alcance en lugar de crear nuevos. Con tal madurez y robustez conseguida, se entiende que estas bases de conocimiento, aún siendo abiertas y de libre edición, son suficientemente fiables como para asentar las bases de sistemas inteligentes.

La importancia de crear una red de relaciones

El conocimiento, según Vergnaud (Vergnaud, 1990), está organizado en campos conceptuales cuyo dominio por parte del sujeto ocurre a lo largo de un extenso periodo de tiempo, a través de experiencia, madurez y aprendizaje. Vergnaud propuso la teoría de los campos conceptuales con la idea de que sirva de marco teórico en investigaciones relacionadas con actividades cognitivas, por lo que cobra importancia directa en el campo de la Inteligencia Artificial. Los conocimientos sólo adquieren generalidad si los elementos que los definen son aprehensibles por el sujeto, al margen de referencias a situaciones particulares. Esto implica que deben estar integrados en una red de conceptos.

[2] https://en.wikipedia.org/wiki/Wikipedia:Modelling_Wikipedia%27s_growth#Old_exponential_model_for_article_count_of_Wikipedia

La teoría de los campos conceptuales supone que **"el amago del desarrollo cognitivo es la conceptualización"**. Ella es la piedra angular de la cognición. Luego, se debe prestar toda la atención a los aspectos conceptuales de las relaciones y al análisis conceptual de las situaciones para las cuales los estudiantes desarrollan sus esquemas, en la escuela o fuera de ella (Moreira).
Se entiende entonces la importancia de la creación de una red de conceptos para el correcto aprendizaje y la creación de conocimiento. Basándonos en la **teoría de los campos conceptuales** se puede justificar un mayor estudio y esfuerzo en este tema.



Como ejemplo o caso de estudio citamos al profesor Beveridge. En "The Art of Scientific Investigation", el profesor W. Beveridge de la Universidad de Cambridge escribió que los científicos exitosos "han sido personas con amplios intereses", lo que ha llevado a su originalidad: "La originalidad a menudo consiste en vincular ideas cuya conexión no se sospechaba previamente.". También sugirió que los científicos deberían expandir su lectura fuera de su propio campo, para agregar conceptos nuevos a su conocimiento (y tener así más puntos cuando llegue el momento de conectarlos más adelante).

La visualización es un componente vital del proceso analítico y de aprendizaje, pero la información presentada sin contexto pierde parte de su idea. Los datos son más que un solo conjunto de números. Las ideas se pueden (y deben) tomar de varias partes de un conjunto o combinar diversas fuentes en una comprensión procesable. Sin embargo, sin la capacidad de ordenarlo rápida y lógicamente, transmitir la información es un proceso menos directo, por lo que la Inteligencia Artificial cobra importancia en este aspecto. Según Jorn Hees y su grupo de trabajo (Jorn Hees, 2010), simular asociaciones de conceptos humanas podría mejorar las capacidades de comprensión de texto de las máquinas. "Gracias a los Datos Enlazados de la Web Semántica, tenemos un conjunto de datos muy grande y prometedor para simular asociaciones humanas". Sin embargo, las asociaciones humanas tienen diferentes fortalezas, mientras que los Datos Enlazados trata a todas los tripletes por igual, lo que Jorn Hees resuelve mediante la asignación de ponderaciones. En este trabajo se decide solventar el problema de la relevancia mediante un filtrado exhaustivo.

Se deriva pues la importancia en la esquematización y en la correcta visualización de conceptos, que ayuden a los estudiantes a tener una vista amplia de conceptos y al agente inteligente a inferir concepto de manera 'humana'. Una vez queda clara la importancia, debemos pasar pues a la tarea de cómo representar dicha red de conceptos.

La importancia de la visualización

La representación del conocimiento y el razonamiento es un área de la Inteligencia Artificial cuyo objetivo fundamental es representar el conocimiento de una manera que facilite la inferencia (sacar conclusiones) a partir de dicho conocimiento.
Qué es la representación del conocimiento se entiende mejor en términos de cinco roles fundamentales que juega, cada uno crucial para la aplicación (R. Davis, 1993):

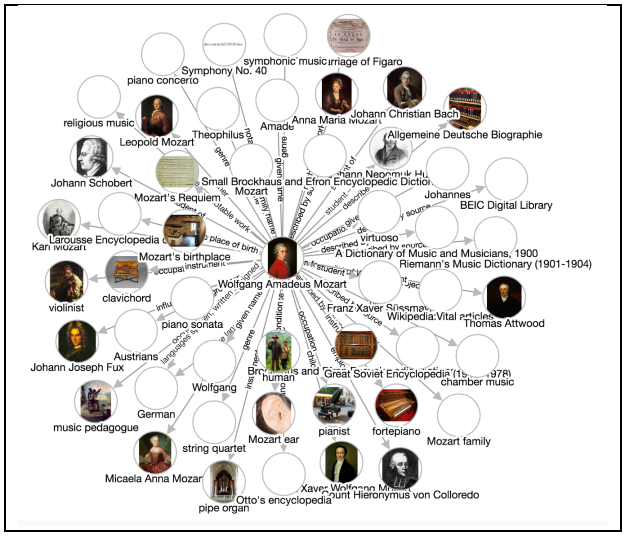
- Una representación del conocimiento es fundamentalmente un sucedáneo, un sustituto para el objeto en sí, usado para activar una entidad a efectos de determinar las consecuencias, pensando en lugar de actuando, o sea, razonando acerca del mundo en lugar de tomando acción en él.
- Es un grupo de compromisos ontológicos, una respuesta a la pregunta sobre los términos en que se debe pensar acerca del mundo.
- Es una teoría fragmentaria del razonamiento inteligente, expresado en términos de tres componentes: (i) El concepto fundamental de la representación del razonamiento inteligente; (ii) El conjunto de inferencias que la representación sanciona; y (iii) El conjunto de inferencias que recomienda.
- Es un medio para una computación pragmáticamente eficiente, el entorno computacional en que el pensamiento tiene lugar. Una contribución para esta eficiencia pragmática viene dada por la guía que una representación provee para organizar información, de modo que facilite hacer las inferencias recomendadas.
- Es un modo de expresión humana, un lenguaje en el que se dicen cosas sobre el mundo.

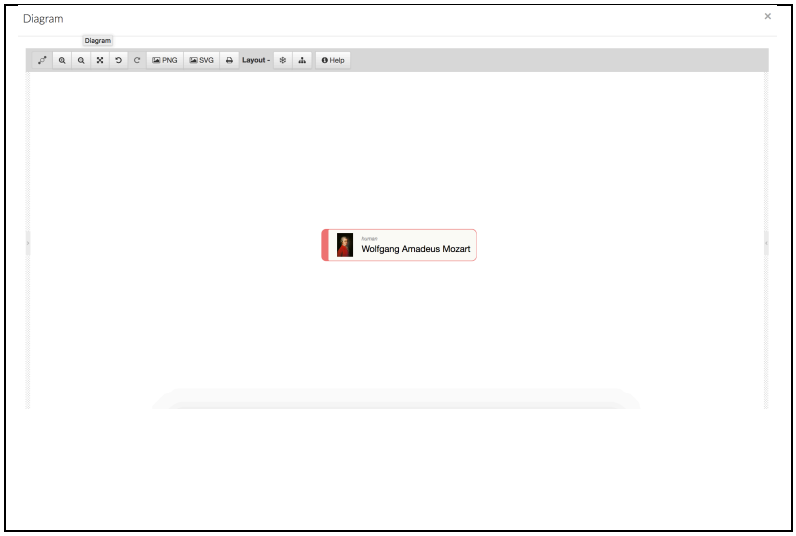
Todas las representaciones son aproximaciones imperfectas de la realidad, cada aproximación atiende algunas cosas e ignora otras. El compromiso que hacemos al seleccionar una u otra ontología puede producir una visión muy diferente de la tarea en cuestión, por lo que se dedicará especial atención a la elección y construcción de la ontología de este proyecto, con un apartado dedicado a ello.

Un desafío clave en el área de la Web Semántica es proporcionar un fácil acceso de los usuarios a la gran cantidad de datos ocultos en los repositorios de Datos Vinculados. El acceso directo a los datos requiere una comprensión de los lenguajes semánticos de consulta y los conjuntos de datos específicos. Una forma de abstraer a los usuarios de los modelos de datos son las interfaces de lenguaje natural para los Datos Vinculados, que traducen las consultas en lenguaje natural a SPARQL y ocultan la complejidad. En muchos casos, sin embargo, el acceso visual a los datos es más intuitivo. Citando a Pavlov (Pavlov, et al., 2016) en el mundo de la Web Semántica todavía existe una gran necesidad de herramientas prácticas que sean accesibles a los personas sin conocimiento especial requerido (Graves & Hendle, 2013), (Dudas, Zamazal, & Svate, 2014) y (Ramakrishnan & Vijayan). De hecho, esta falta de herramientas apropiadas se menciona con frecuencia en publicaciones y debates abiertos (Janev & Vranes, 2011).
En cuanto a la gestión, visualización y transformación de datos abiertos vinculados (LOD, del inglés Linked Open Data) la selección de herramientas es aún más escasa.

Prácticamente, todas las opciones disponibles poseen una o varias de estas características: voluminosas y complejas (difícil de implementar, aprender y mantener), orientadas al desarrollador (solo un desarrollador puede iniciarlas y usarlas), costosas y ni intuitivas ni visuales. Las características cognitivas y la interacción visual del usuario son casi desconocidas entre los usuarios de LOD y, en ese sentido, están muy desprovistos de softwares avanzados de interfaces de usuario (UI) en comparación con los usuarios de bases de datos tradicionales.

Sí que existen sin embargo trabajos ambiciosos que intenta abarcar el problema de la representación de grandes conjuntos de datos relacionados. En el trabajo llevado a cabo por Mouromtsev y su equipo (Dmitry Mouromtsev 1, Haase, Pavlov, Emelyanov, & Morozov, 2018) se intentó presentar una herramienta visual de preguntas y respuestas (QA) sobre el gráfico de conocimiento de Wikidata basado en la representación y el razonamiento con diagramas y grafos. La demostración se basaba en la plataforma metaphacts[3] con la librería Ontodia[4] incrustada. Dicha herramienta consta de dos partes, la primera es de visualización de los datos extraídos de Wikidata y la segunda es la posibilidad de crear tu propio grafo o diagrama. En su trabajo y tras los experimentos, demostraron que los usuarios podían inferir conclusiones y contestar a las preguntas que les hicieron los encuestadores. Sin embargo, tal y como se explica en la publicación, algunos se sintieron perdidos a la hora de buscar el punto de partida, o de crear conexiones entre dos nodos. Se han sacado dos conclusiones a partir de este trabajo, primero, no se recomienda abrumar al usuario con una ingente cantidad de datos, de tal modo que se sienta perdido y segundo, si se pretende que el usuario navegue y descubra conceptos, se recomendaría un grafo con algunas conexiones ya hechas.





Ontodia y Metaphacts: en la primera imagen se observa la parte de visualización de datos, en la segunda imagen, la herramienta para que el usuario construya su propio grafo o diagrama. Se

[3] http://www.metaphacts.com/
[4] http://www.ontodia.org/

sugiere la integración de ambas partes en una sola, de tal manera que el sistema elige los primeros nodos, sin resultar abrumador para el usuario, que puede editar después.

Numerosos son también los estudios y proyectos sobre la creación de modelos para visualización de Datos Enlazados en general, como puede verse en los trabajos de Auguste Atemezing (Ghislain Auguste Atemezing, 2014), Klímek (Jakub Klímek, 2014) o Brunetti (Josep Maria Brunetti, 2012). Estos trabajos presentan una guía y siguen un proceso similar a la hora de construir sus modelos de visualización, sin embargo carecen de ontología, resultan abrumadores, o se centran únicamente en la visualización, sin la posibilidad de exportar los datos que han pasado por un usuario para su uso en aplicaciones inteligentes después.

Lo que se intenta demostrar no es sólo la importancia de la visualización de datos para el aprendizaje en las personas, sino del parte fundamental que desempeña a la hora de conectar conceptos e ideas de diferentes fuentes, aportando contexto y sentido global, además de su importancia en el uso de aplicaciones inteligentes, como constructor de datasets de manera supervisada.

La importancia de la contextualización

Según la Real Academia de la Lengua Española, el contexto se define como: (i) Entorno lingüístico del que depende el sentido de una palabra, frase o fragmento determinados. (ii) Entorno físico o de situación, político, histórico, cultural o de cualquier otra índole, en el que se considera un hecho. Sin embargo, la definición toman distintos matices y cumple distintos objetivos dependiendo del campo de la Inteligencia Artificial en el que se aplique (Akman, 2002), ya sea en el Procesamiento del Lenguaje Natural (palabras homónimas, oraciones que usan conectores..), Categorización (aplicar las correctas reglas), Recuperación de Información (aumentando el rendimiento y la eficiencia de las consultas) o en la Representación del Conocimiento (determinando la granularidad y precisión del conocimiento). La herramienta de recopilación incremental, desde una semilla hasta una red de conceptos estructurada y validada por el usuario que se propone para construir y desarrollar en este trabajo representaría la red de conceptos que se mencionan en los apartados anteriores. Esta red de conceptos creada además aporta luz sobre los distintos contextos en los que tenga sentido la semilla.

Una posible aplicación de este trabajo bien podría ser la formación de las bases de un sistema de recomendación que permita a los usuarios navegar y descubrir contenidos. Los típicos sistemas de recomendación suelen estar basados en etiquetas o en datos de entrada, las cuales el usuario nuevo ha escogido al entrar e inscribirse en ese sistema, como parte del proceso de incorporación o tutorial, o introducido de forma activa mediante su participación en la herramienta (input activo, sea el usuario consciente o no de la recopilación de sus costumbres y usos). Se podrían recomendar contenidos no sólo basándose en el etiquetado literal o en el input activo del usuario,

pero también en su relación con otros conceptos y el contexto. Como bien explican Adomavicius y Tuzhilin (Gediminas Adomavicius, 2011), los sistemas de recomendación conscientes del contexto (CARS, de las siglas en inglés Context Aware Recommender Systems) generan recomendaciones más relevantes al adaptarlas a la situación contextual específica del usuario, además de necesitar menos datos de entrada por parte del usuario. Esta herramienta sería útil pues para un 'cold start' o 'arranque en frío'[5] de cualquier aplicación (cuando se tiene mucha información de los elementos a recomendar, pero el usuario es nuevo y no se tiene feedback ni ninguna interacción previa de él, o viceversa). Los elementos a recomendar contienen atributos (en nuestro caso provenientes de la base de conocimiento escogida), tales atributos son muy útiles y los métodos de minería de datos se pueden usar para extraer conocimiento en formas de reglas y patrones que posteriormente se utilizan para la recomendación (algoritmos de relevancia o similitud). Incluso la descripción de texto largo y plano puede ser procesada por herramientas avanzadas de PNL (Procesamiento del Lenguaje Natural).

En lo que se refiere a la tarea de Categorización, Rosch asegura que se trata de uno de los procesos mentales básicos en la cognición (Rosch, 1978), y por lo tanto digno de estudio y análisis en la Inteligencia Artificial. Barwise y Seligman (Barwise, 1992) usaron las llamadas regularidades naturales para estudiar el papel del contexto en la categorización. Estas regularidades son confiables pero falibles (las aves vuelan por lo general, pero no siempre, como por ejemplo la avestruz, o un mismo concepto puede hacer referencia a varias categorías), por lo que el contexto resulta de vital importancia.

Cabe remarcar que las aplicaciones de Inteligencia Artificial a menudo son criticadas por su fragilidad y lentitud. Roy Turner (Turner, 1999) sin embargo, certifica que ambos problemas pueden mejorarse si el programa de IA es sensible al contexto.

1.4. Estructura del trabajo

El trabajo va a empezar realizando un estudio de las Bases del Conocimiento disponibles y abiertas más importantes y relevantes, valorando sobre todo la cantidad de datos con las que cuentan, la precisión de estos datos y su veracidad, la ontología en la que se basan y sus estructura interna, así como la dificultad o facilidad de acceder a esos datos. Por ejemplo, si hay tecnologías desarrolladas en su entorno y en su sistema en las que se pueda confiar para hacer consultas en vivo sin tener que volcar o replicar el conocimiento en nuestro propio sistema. A continuación se va a analizar la ontología de la base de conocimiento escogida para adaptarla a las necesidades específicas de este trabajo. Los siguientes apartados se van a centrar en el análisis de los datos y sus relaciones, así como en su búsqueda y recopilación para volcarla al interfaz del usuario. En esta fase se tomarán importantes decisiones de diseño, que garantizaran el correcto balance entre la cantidad de datos a presentar y la capacidad de estos en aportar significado y contextualización al usuario. Con técnicas de Minería

[5] https://medium.com/recombee-blog/recommender-systems-explained-d98e8221f468

de Datos se intentarán integrar los datos estructurados y no estructurados, ofreciendo siempre aquellos que resulten más relevantes.

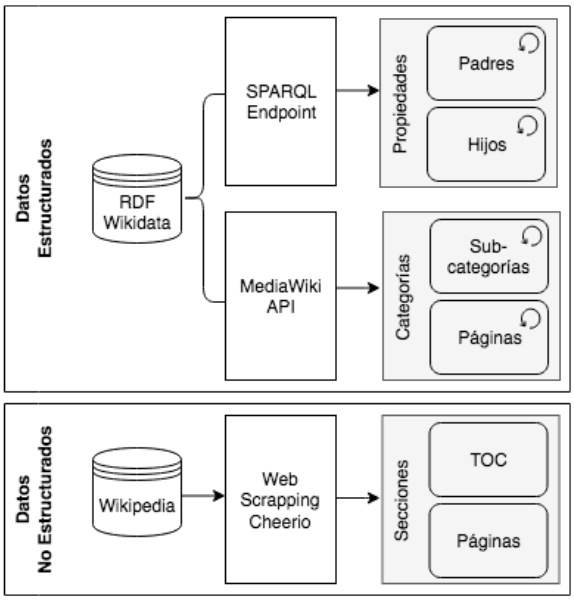


Ilustración 1: Esquema de la primera parte del trabajo en la que se recogen e integran datos desde las bases de conocimiento estructurada y no estructurada. Para una vista completa de la arquitectura, consultar el Anexo

La segunda parte del trabajo es aquella en la que el usuario va a tener acceso. Constará de la interfaz desde donde se introducirá la semilla o semillas de las que se quiere aprender. Ayudándose de tecnologías como D3.js, una librería para páginas basadas y dirigidas por datos. La experiencia de usuario se va a dividir en tres partes o fases. La primera fase será la de crecimiento con filtrado, en ella se le mostrarán al usuario una serie de nodos y conceptos enlazados iniciales a partir de la semilla introducida por él. La segunda fase es la de expansión, en ella se estudiarán métodos de navegación e interacción que faciliten el descubrimiento de conceptos y relaciones, se le permitirá al usuario interaccionar con los conceptos y la interfaz, añadiendo, quitando o editando la información inicial. Por último la tercera fase será la de cierre. Se usarán métodos de 'Clustering' para la integración final de los datos seleccionados por el usuario y se proporcionará la posibilidad de exportar la red semántica creada.

En la tercera y última parte del trabajo se definen un conjunto de experimentos y pruebas, que pretenden verificar la veracidad de las hipótesis descritas en el apartado anterior. Se mostrarán y discutirán los resultados obtenidos para llegar a las conclusiones finales. Por supuesto se adjuntarán ideas y propuestas de futuros trabajos que podrían surgir a partir de éste.

Capítulo 2:
Investigación y Diseño

El trabajo va a empezar realizando un estudio de las Bases del Conocimiento disponibles y abiertas más importantes y relevantes, valorando sobre todo la cantidad de datos con las que cuentan, la precisión de estos datos y su veracidad, la ontología en la que se basan y sus estructura interna, así como la dificultad o facilidad de acceder a esos datos. Por ejemplo, si hay tecnologías desarrolladas en su entorno y en su sistema en las que se pueda confiar para hacer consultas en vivo sin tener que volcar o replicar el conocimiento en nuestro propio sistema. A continuación se va a analizar la ontología de la base de conocimiento escogida para adaptarla a las necesidades específicas de este trabajo. Los siguientes apartados se van a centrar en el análisis de los datos y sus relaciones, así como en su búsqueda y recopilación para volcarla al interfaz del usuario. En esta fase se tomarán importantes decisiones de diseño, que garantizaran el correcto balance entre la cantidad de datos a presentar y la capacidad de estos en aportar significado y contextualización al usuario. Con técnicas de Minería de Datos se intentarán integrar los datos estructurados y no estructurados, ofreciendo siempre aquellos que resulten más relevantes.

2.1. Escoger Base de Conocimiento

Para representar el conocimiento primero debemos acceder a él, debemos buscar Bases de Conocimiento disponibles y abiertas. Una Base de Conocimiento, o en inglés, '*Knowledge Base*', es una base de datos diseñada para cumplir con los complejos requisitos de almacenamiento y recuperación de la gestión de conocimiento computarizada, especialmente contando con el apoyo de la Inteligencia Artificial o sistemas expertos. Una Base de Conocimiento en general consiste en una lista de entidades, información sobre cada una de esas entidades y datos sobre las relaciones entre las entidades. La base de conocimiento más conocida es probablemente Wikipedia (Wikipedia, n.d.). Una comparativa de otras bases de conocimiento estructurado populares se encuentra en el Anexo 'Bases de Conocimiento estructuradas'**.** Basándonos en ellas, en la cantidad de datos disponibles, abiertos y estructurados, la existencia de relaciones con datos no estructurados (permite el descubrimiento de datos nuevos mediante técnica de PLN) y la fiabilidad y veracidad de su contenido, se van a usar **Wikidata y Wikipedia** como Bases del conocimiento en este trabajo.

Conocimiento no estructurado, Wikipedia

Wikipedia es una enciclopedia multilingüe, basada en la web y gratuita que se basa en un modelo de contenido abiertamente editable. Es el trabajo de referencia general y la base de conocimiento más grande y popular en Internet. La Wikipedia, creada en

Enero del 20011, es propiedad y está respaldada por la Fundación Wikimedia. Contiene. En julio de 2018 Wikipedia cuenta con más de 48 millones de artículos/páginas (de los cuales 5,7 millones son en inglés y 1,4 millones en español). Las páginas en Wikipedia son principalmente de los siguientes tipos:

- Entidad: páginas sobre temas de la enciclopedia tradicional, personas, lugares, medios de comunicación, empresas, eventos y más. Son la base de conocimiento en sí, ya que tratan sobre entidades y conceptos.
- Desambiguación: se usa cuando el título de la página es ambiguo porque tiene múltiples significados. Contiene enlaces a diferentes títulos de página para cada significado.
- Redirigir: puntos de un título de página a otro y se usa cuando existe más de un título de página posible. También puede apuntar a una parte específica de una página.
- Administrativo: páginas de usuario, información interna de Wikipedia, plantillas.

Las páginas de Wikipedia se identifican de forma única por el título de su página. Una sola entidad puede tener múltiples títulos de página distintos a lo largo del tiempo y, por lo tanto, no tiene un identificador persistente. Los artículos de Wikipedia son documentos vivos, que cambian con el tiempo, es por eso que al trabajar con ella se deben hacer llamadas en tiempo real, y hacer uso de los identificadores y las relaciones que permiten la correcta navegación a la entidad deseada.
Por ejemplo, alguien añade la palabra 'lengua', refiriéndose al órgano muscular situado en la boca, se le asigna como identificador: https://es.wikipedia.org/wiki/Lengua. Más adelante se intenta añadir un nuevo artículo para lengua, está vez refiriéndose a 'lenguaje' o 'idioma', pero el identificador 'lengua' ya esta cogido. Se procede entonces a un trabajo de desambiguación, el anterior identificador https://es.wikipedia.org/wiki/Lengua llevará ahora a una página del tipo desambiguación de Wikipedia (donde se explican los diferentes significados de la palabra proporcionando los correspondientes enlaces), y se asignan nuevos identificadores para las dos entidades. 'Lengua' del lenguaje será https://es.wikipedia.org/wiki/Lenguaje y 'lengua' sobre el órgano muscular en la boca será https://es.wikipedia.org/wiki/Lengua_(anatomía). El identificador para la lengua de la boca ha sido modificado.

Conocimiento estructurado, Wikidata

Como se ha visto anteriormente, Wikipedia es la mayor base de conocimiento que existe, sin embargo todo ese conocimiento no está estructurado y los artículos e identificadores pueden cambiar constantemente. Es por eso que se creó Wikidata. Wikidata es una base de datos secundaria gratuita, colaborativa, multilingüe, que recopila datos **estructurados** para brindar soporte a Wikipedia, a Wikimedia Commons, a las otras wikis del movimiento Wikimedia y a cualquier persona en el mundo. La imposición de un alto grado de organización estructurada permite una fácil reutilización de datos por parte de los proyectos de Wikimedia y de terceros, y

permite que las computadoras los procesen y "entiendan". Wikidata ayuda también a Wikipedia con cajas de información más fáciles de mantener y enlaces a otros idiomas, lo que reduce la carga de trabajo de edición y mejora la calidad.

El repositorio Wikidata consta principalmente de *items*, cada uno con una etiqueta (*label*), una descripción y cualquier cantidad de alias. Los elementos se identifican de forma única con una Q seguida de un número, como por ejemplo Douglas Adams (Q42). Las declaraciones describen características detalladas de un artículo y consisten en una propiedad y un valor. Las propiedades en Wikidata tienen una P seguida de un número, como con el término 'educado en' (P69).

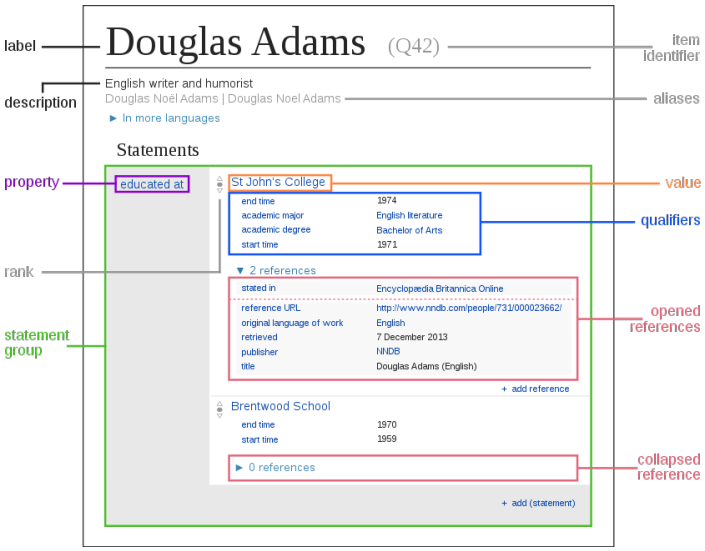


Ilustración 2: Estructura de un 'item' en Wikidata

La mejor forma de hacer uso de datos de Wikidata es mediante SPARQL. SPARQL (acrónimo de Simple Protocol And RDF Query Language) es un lenguaje de consulta RDF, es decir, un lenguaje de consulta semántica para bases de datos. Con SPARQL se puede extraer cualquier clase de datos, con una consulta compuesta de combinaciones lógicas de tripletes (*triplets*), elemento – propiedad – valor, (*item – property - value*).

2.2. Definir Representación del Conocimiento

Una vez que se tiene el conocimiento, éste se debe de representar. En la problemática de la representación del conocimiento debemos separar dos partes. La primera, la representación al nivel del programador, es decir, cómo se va a representar el conocimiento para que se pueda manipular, programar y hacer uso de él. La segunda, la visualización de dichos datos de manera gráfica y representativa, que ayude al usuario a inferir conocimiento a partir de éste.

A nivel de programación

En IA debes ser capaz de representar de manera precisa el conocimiento para poder explorar ese conocimiento para resolver problemas. Esto se debe a que, en contraste con los humanos, las computadoras solo pueden manipular el conocimiento que tiene una sintaxis y semántica precisas, e incluso si los requisitos se relajan un poco, aún debe haber algún lenguaje subyacente que sea preciso. Aunque la representación del conocimiento puede hacerse usando diferentes lenguajes formales, los dos lenguajes que han llegado a ser dominantes en la IA en el transcurso del tiempo son la teoría lógica y de probabilidad.

En el anexo 'Representación del conocimiento a nivel de programación en la Inteligencia Artificial.' se van a describir a modo de sumario todas las distintas formas de representación de conocimiento.

En este trabajo se ha optado por la representación en forma de **Redes Semánticas o grafos con nodos y enlaces** almacenados como proposiciones. Ejemplos de los primeros grafos y redes semánticas en entornos de programación se pueden encontrar en el trabajo llevado a cabo por Stillings (Stillings, 1987). Las Redes Semánticas son una estructura gráfica que representan el conocimiento en forma de patrones de nodos y arcos interconectados, estos arcos representan relaciones semánticas entre conceptos en dicha red o estructura gráfica. Las implementaciones informáticas de redes semánticas se desarrollaron primero para la Inteligencia Artificial y la traducción automática, pero las versiones anteriores se han utilizado durante mucho tiempo en filosofía, psicología y lingüística (Saphiro, 1992). Se ha optado por esta estructura de datos porque es la que mejor representa relaciones semánticas entre conceptos, y por tanto la que mejor encaja en el objetivo de este proyecto. Se debe tener en cuenta que se está intentando lograr la rápida contextualización y conceptualización de un elemento semilla introducido por el usuario. Se está tratando con problemas de lenguaje natural y los grafos son la mejor forma de representarlos, ya que las conexiones entre conceptos pueden ser ilimitadas y en todas direcciones (descartando así la estructura en forma de árbol o lista por ejemplo). Por supuesto las normas y reglas subyacentes al grafo se dictarán en la definición de la ontología.

A nivel de representación:

La justificación para la representación del conocimiento es que el código de procedimientos convencional no es el mejor formalismo para resolver problemas complejos. La representación del conocimiento hace que el software complejo sea más fácil de definir y mantener que la programación por procedimientos y se puede usar en sistemas expertos. Por ejemplo, hablar con expertos en términos de reglas comerciales en lugar de código disminuye la brecha semántica entre usuarios y desarrolladores y hace que el desarrollo de sistemas complejos sea más práctico. La representación del conocimiento va de la mano con el razonamiento automático porque uno de los propósitos principales de representar explícitamente el

conocimiento es poder razonar sobre éste, hacer inferencias, afirmar nuevos conocimientos, etc.
Existen diferentes maneras de representar y visualizar datos que van a depender del concepto que se intenta mostrar con dicho gráfico, ya sea una comparación, una relación, una composición o una distribución:

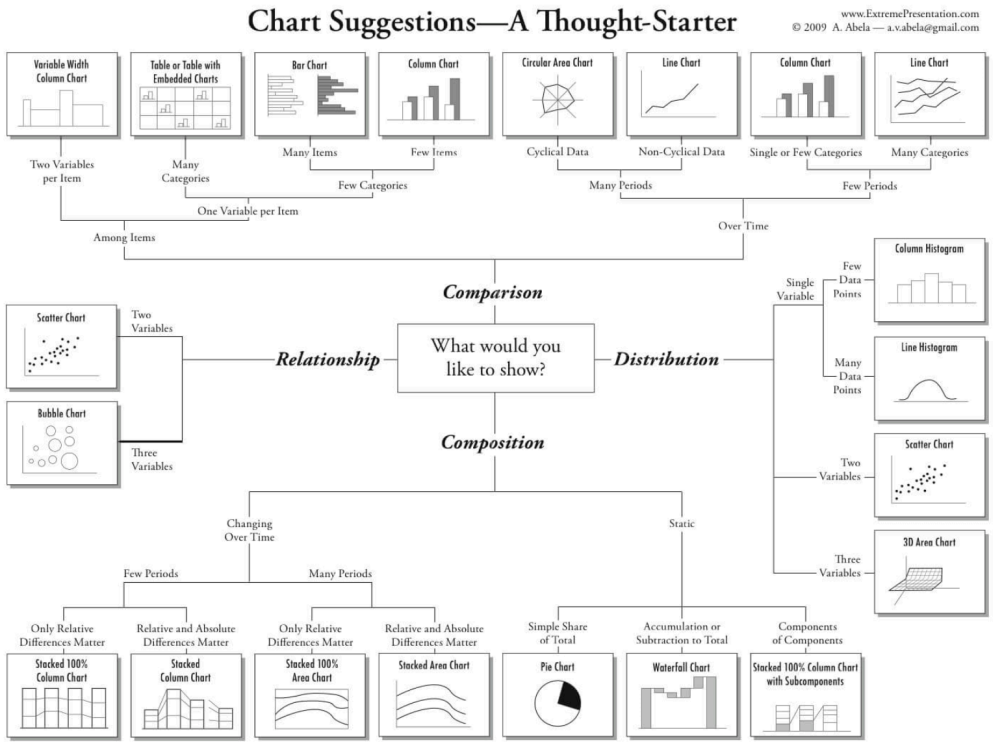


Ilustración 3: Alternativas de visualización dependiendo de la intención que se quiere mostrar de los datos. Diagrama no original, de fuente desconocida.

La visualización de los datos ha resultado ser fundamental para el desarrollo de la IA, ya que puede ayudar a los desarrolladores de IA y a las personas preocupadas por la adopción de los sistemas de IA a explicar y comprender estos sistemas. Los respetados investigadores de Google Fernanda Viégas y Martin Wattenberg llegaron incluso a llamar a su discurso principal de EuroVis 2017 Visualización: el arma secreta del aprendizaje automático. Uno de los mayores beneficios de los algoritmos de aprendizaje automático es acelerar el proceso de descubrimiento de datos y el proceso de encontrar conexiones e ideas más profundas en grandes conjuntos de datos. Otro ejemplo práctico del poder de la visualización en la Inteligencia Artificial es Watson Analytics: Un servicio de visualización y análisis predictivo de datos inteligente que se utiliza para descubrir patrones y significado en datos. Con el descubrimiento de datos guiado, la analítica predictiva automatizada y las funcionalidades cognitivas como el diálogo en lenguaje natural, se puede 'conversar' con los datos y obtener respuestas comprensibles.

En este trabajo se va a construir una interfaz con barra de búsqueda, grafo de relaciones entre conceptos, y listas de conceptos o enlaces importantes que se podrán

añadir con un solo click. Se mostrará la máxima cantidad de información posible, de forma ordenada y sin resultar demasiado abrumador para el usuario. Pasar el ratón o cursor por encima de los conceptos desplegará una alerta con más información sobre este, con descripciones a modo de extracto sacados directamente de la API de MediaWiki. Cuando sea posible se mostrarán también fotografías. El grafo principal que se podrá exportar estará impulsado por la librería D3.js que permite un sencillo uso y customizado en los elementos <svg> de HTML y que hemos explicado anteriormente en la parte de justificación.

2.3. Definición y desarrollo de la Ontología

Una ontología en ciencias de la comunicación y en Inteligencia Artificial, y según la Real Academia Española, es una red o sistema de datos que define las relaciones existentes entre los conceptos de un dominio o área del conocimiento. Para apoyar el intercambio y la reutilización del conocimiento formalmente representado entre los sistemas de Inteligencia Artificial, es útil definir el vocabulario común en el que se representa el conocimiento compartido (Gruber T. R., 1993). Una ontología es pues un gráfico (estructura de datos). Cada nodo de este gráfico representa un "concepto" (Gruber T. , n.d.). Un concepto es una unidad en la que uno puede pensar y que en nuestro caso serán entidades de Wikidata o páginas de Wikipedia que servirán de semilla para el posterior desarrollo. Los conceptos serán pues palabras o frases cortas que típicamente corresponderán a sustantivos o frases nominales, pero no tienen por qué hacerlo (Gruber T. R., 1993).

Los componentes comunes de las ontologías incluyen:

Individuos: instancias u objetos (los objetos básicos o de "nivel del suelo"). El equivalente en Wikidata será el Ítem, que se refiere a un objeto, concepto o evento del mundo real al que se le da un identificador (un equivalente de un nombre) en Wikidata junto con información al respecto. Cada Ítem tiene una Wikipage correspondiente en el espacio de nombres principal de Wikidata. Los elementos se identifican mediante una identificación prefijada (como Q5), o mediante un enlace a una página externa, o mediante una combinación única de etiqueta y descripción multilingüe. Los elementos también pueden tener alias para facilitar la búsqueda. La parte principal de datos de un Ítem es la lista de afirmaciones ('Statements') sobre el elemento. Un elemento se puede ver como el parte del **sujeto de un triplete** en datos vinculados. En el caso de Wikipedia los Individuos serán los títulos de las páginas, ya que estos además funcionan como identificadores y son únicos para cada concepto.

Clases: conjuntos, colecciones, conceptos, tipos de objetos o tipos de cosas. Existen incontable número de clases en Wikipedia y Wikidata, lugar, persona, disciplina académica… En nuestro caso se usarán las categorías del atributo 'categoría en Commons' (P373) para clasificar, además de la propiedad 'subclase de' (P279) que indica una jerarquía de clases directamente.

Atributos: aspectos, propiedades, características o parámetros que los objetos (y las clases) pueden tener. El equivalente en Wikidata es Propiedad (también atributo) o 'Property' en inglés, que es el descriptor para un valor de datos, o alguna otra relación o compuesto, pero no el valor o los valores de los datos en sí. Cada instrucción en una página de Ítems se vincula a una Propiedad y le asigna a dicha Propiedad uno o varios valores. En comparación con los datos vinculados, la Propiedad representa el **predicado de un triplete.** Con respecto a Wikipedia es más complicado, Wikipedia no está estructurada, la información se despliega en forma de texto, oraciones y descripciones. Se ha decidido que la forma de buscar atributos de una manera segura será analizando el TOC, Table Of Contents o Tabla de Contenidos. El TOC divide la página en secciones por temáticas, por ejemplo, la página de Paris tiene secciones distintas: Historia, Economía, Demografía, Cultura… Se puede entender cada una de estas secciones del TOC como un atributo, y el contenido de estas secciones será el valor de dicho atributo.

Relaciones: formas en que las clases y los individuos se pueden relacionarse entre sí. En Wikidata se llaman '*Claims*', Un Claim consiste en una propiedad (como ubicación (P276)) y un valor (por ejemplo, Alemania (Q183)) asociadas a un sujeto (Muro de Berlín (Q5086)). Un reclamo, Claim, puede tener calificadores, como por ejemplos calificadores temporales que dicen que el reclamo es válido dentro de un marco de tiempo específico. Los Claims son los tripletes de Wikidata cuyo formato, sujeto-propiedad-valor, equivale al triplete del lenguaje RDF formal sujeto-predicado-objeto. En Wikipedia se formarán las relaciones mediante el nombre de la página, sujeto, el nombre de la sección en el TOC, propiedad, y el contenido en dicha sección, valor.

Restricciones: descripciones formalmente establecidas de lo que debe ser cierto para que se acepte alguna afirmación como entrada, en Wikidata llamados 'Constraints'.

Reglas: declaraciones en forma de una oración if-then (antecedente-consecuente) que describe las inferencias lógicas que pueden extraerse de una afirmación en una forma particular. Las propiedades básicas de membresía por ejemplo son transitivas. Por ejemplo: Si A es subclase de B y B es subclase de C, entonces A es subclase de C.

Axiomas: aserciones (incluidas las reglas) en una forma lógica que juntas comprenden la teoría general que la ontología describe en su dominio de aplicación. Esta definición difiere de la de los "axiomas" en la gramática generativa y la lógica formal. En estas disciplinas, los axiomas incluyen solo afirmaciones afirmadas como conocimiento a priori. Como se usa aquí, los "axiomas" también incluyen la teoría derivada de enunciados axiomáticos.

Eventos: el cambio de atributos o relaciones. En Wikidata y Wikipedia la edición es libre y está abierta para cualquier usuario. Ambas se basan en el paquete wiki de MediaWiki, lo que significa que el contenido de las páginas se puede agregar, modificar o eliminar en colaboración con otros. A diferencia de Wikipedia, Wikidata

también usa el software Wikibase que permite la edición colaborativa de datos estructurados.

Para acotar el proyecto se va a enfocar el estudio y trabajo en las ontologías relacionales y las reglas de jerarquía, enfocándonos en las propiedades que permitan tanto ramificaciones tipo 'padre', generalizando, como ramificaciones tipo 'hijo', especificando. Esto permitirá navegar al usuario de una manera intuitiva entre conceptos, yendo tan profundo y detallado como desee.
Las propiedades/atributos escogidas se van a detallar en la siguiente tabla:

Nombre 'Padres'	Identificador	Descripción	Alias
instancia de / instance of	(P31)	Este elemento es un ejemplar del otro elemento. Propiedad más popular de Wikidata apareciendo en 45 millones de items	es un es una instancia de es miembro de es una ocurrencia de es un ejemplo de ∈ rdf:type
subclase de / subclass of	(P279)	Todas las instancias de estos elementos son instancias del otro elemento. Este elemento es una subclase del otro. No debe confundirse con P31 (instancia de). Usado 1,4 millones de veces	subtipo de ⊆ ⊂ rdfs:subClassOf
forma parte de / part of	(P361)	Es parte del elemento. Propiedad inversa de "compuesto de" (P527).	parte de es parte de incluso en contenido en sección de componente de sistema de
tiene partes de la clase / has parts of the class	(P2670)	Este elemento está compuesto de elementos de otra clase	tiene parte del tipo comprende los elementos del tipo
precedido por / follows	(P155)	Inmediato predecesor en alguna serie de la que el elemento forma parte	antecedida por antecedido por precuela sucede a sigue a

			precedida por
utilizado por / used by	(P1535)	Elemento, concepto, persona u organización que utiliza el objeto	usado por
basado en / based on	(P144)	El trabajo utilizado como base para el elemento	basada en derivado de adaptado de es adaptación de
categoría en Commons / Commons category	(P373)	Nombre de la categoría de Wikimedia Commons con archivos relacionados con este elemento	categoría Commons Commons categoría

Nombre 'Hijos'	Identificador	Descripción	Alias
compuesto de / has part	(P527)	Entidad que forma parte de este elemento. Inverso de " forma parte de " P361	compuesto por conformado por consta de consiste en formado por
sucedido por / followed by	(P156)	Inmediato sucesor en alguna serie de la cual forma parte el elemento	seguido por secuela precede a antecede a
usa / uses	(P2283)	Objeto usado por el sujeto	utiliza emplea elemento usado hace uso de
obra derivada / derivative work	(P4969)	Elemento creado a partir de la mayor parte de este trabajo	

Las propiedades citadas en la tabla han sido cuidadosamente escogidas por aportar y reflejar mayor sentido contextual. Por ejemplo, se han evitado las propiedades cuyo valor suele ser cuantitativo (date of birth (P569), educated at (P69), population (P1082)…) y que no aportan en la creación de una red de conceptos. Se ha añadido la propiedad categoría en Commons (P373) que resulta muy interesante para esquematizar, y que revela datos importantes que de alguna manera se podrían haber perdido, por ejemplo, se ha omitido la propiedad capital (P36), porque existe la categoría Capitales, que aparecerá si ese elemento es una capital importante. Se han incluido también las propiedades básicas de membresía (Basic membership properties, 2018), instancia de (P31), subclase de (P279) y forma parte de (P361), que

además de ser de las propiedades más populares (Wikidata:Database reports/List of properties/Top100, 2018), son las más indicadas para mostrar conceptos de pertenencia y jerarquía, lo que ayuda a una visualización contextual. Como proyecto futuro, esta elección de propiedades se podría llevar a cabo metódica y automáticamente, mediante algoritmos de Inteligencia Artificial, en este trabajo sin embargo, se considera que los objetivos son específicos al área de conceptualización y contextualización, por lo que merece la pena el esfuerzo de búsqueda y elección manual de estas propiedades. Para una explicación más detallada sobre las propiedades de Wikidata y los factores que han influido en la elección de éstas para el filtrado se recomienda revisar el Anexo: Propiedades de Wikidata.

2.4. Análisis y extracción de datos estructurados

Como se ha mencionado anteriormente, Wikidata esta profundamente ligado a SPARQL mediante WDQS (Wikidata Query Service), el servicio de consultas Wikidata. SPARQL es una de las tecnologías aprobadas y estandarizadas por el consorcio W3C y forma parte de la pila arquitectónica de la Web semántica. La Web Semántica a su vez se considera como un integrador fundamental de diferentes contenidos mediante datos enlazados, basados y potenciados además por otras tecnologías como RDF, OWL y SKOS. En el ANEXO sobre el World Wide Web Consortium (W3C) y la Web Semántica se explica cómo se interrelacionan estos conceptos y tecnologías de manera integrada para poder acceder a datos estructurados de manera estandarizada.

Servicio de consultas de Wikidata

El servicio de consulta de Wikidata (WDQS) es un paquete de software y servicio público diseñado para proporcionar un punto de acceso final a SPARQL permitiendo realizar consultas en el conjunto de datos de Wikidata. WDQS entiende muchas abreviaciones de acceso directo, conocidas como prefijos. Algunos son internos a Wikidata wd, wdt, p, ps, bd, etc. y muchos otros son prefijos externos comúnmente utilizados, como rdf, skos, owl, schema, etc.
Para triples WDQS simples, los elementos deben estar prefijados con wd:, y las propiedades con wdt:. Esto solo se aplica a valores fijos ya que las variables no tienen prefijo en las consultas.

Búsqueda y recopilación de nodos padre e hijos

Una vez entendidos los conceptos y las tecnologías con las que se van a trabajar, se procede al desarrollo del servidor que se encargará de hacer las consultas pertinentes y devolver los datos procesados a la interfaz del usuario en la página web.

El primer paso es crear un método que nos permita hacer consultas sobre un nodo y sobre un tipo de propiedades determinadas. Las propiedades se han guardado con su prefijo inicial wd:

```
const sparqlParent = {
    instanceOf: 'wdt:P31',
    partOf: 'wdt:P361',
    hasPartsOf: 'wdt:P2670',
    subclassOf: 'wdt:P279',
    follows: 'wdt:P155',
    usedBy: 'wdt:P1535',
    basedOn: 'wdt:P144',
    category: 'wdt:P373'
}
```

```
const sparqlChild = {
    hasPart: 'wdt:P527',
    followedBy: 'wdt:P156',
    uses: 'wdt:P2283',
    derivatesFrom: 'wdt:P4969'
}
```

```
var generateQuery = (node, propertiesArray) => {
    const nodeId = node.id.split("/").pop()
    var selectQuery = ""
    var whereQuery = []
    propertiesArray.forEach(function(property) {
        selectQuery += "?" + property.title + " ?" + property.title + "Label " + " ?" + property.title + "Description "
        whereQuery.push(" { wd:" + nodeId + " " + property.value + " ?" + property.title + " .} ")
    })
    return `#` + node.label + `
    SELECT ` + selectQuery + ` WHERE {
      ` + whereQuery.join("UNION") + `
      SERVICE wikibase:label { bd:serviceParam wikibase:language "[AUTO_LANGUAGE],en". }
    }`
}
```

La consulta SPARQL generada nos sirve de plantilla tanto en la búsqueda de conceptos generalizadores, como en la de hijos y ramificaciones con la intención de especificar, a partir de un nodo central, dependiendo de la lista de propiedades introducida como parámetro.

El siguiente paso consiste en realizar sucesivas llamadas a esta plantilla, de manera recursiva, a partir del nodo semilla introducido por el usuario. Surge la problemática de acotar el problema, ya que se necesita mostrar al usuario datos relevantes al nodo, ¿a partir de qué ramificación los nodos empiezan a perder relevancia y sentido respecto al nodo semilla? ¿debemos buscar hijos y padres de cada nodo nuevo que surja?. El crecimiento del grafo puede crecer de manera exponencial, por lo que rápidamente se convierte en un caos de información provocando la reacción opuesta de la que se busca del usuario. Se necesita llegar a un equilibrio, mostrando la suficiente información sin resultar abrumadora. Se debe dejar al usuario navegar y descubrir el contenido por sí mismo y si así lo desea, a partir de un grafo de conceptos inicial que se le presente. El esfuerzo entonces debe centrarse en construir ese grafo base desde el cuál el usuario podrá interactuar.
Se decide por defecto analizar recursivamente 2 niveles o ramificaciones de relaciones entre elementos nodos. Es decir, se subirá 2 niveles en la búsqueda de padres, y se bajará 2 niveles en la búsqueda de hijos. Ésta variable se ha convertido además en un parámetro que se puede pasar al hacer consultas con el servidor, esto significa que a la hora de hacer experimentos podemos alterar la profundidad de las ramificaciones según se desee. Si el parámetro se obvia, entonces se ejecutará la búsqueda con el valor por defecto, en este caso 2.

Otra particularidad y decisión del diseño que se debe tomar en cuenta es que sólo el nodo semilla cuenta con los dos tipos de ramificación, del tipo padre y del tipo hijo. El resto sólo cuenta con su tipo correspondiente, es decir, si un nuevo nodo ha surgido al realizar una consulta del tipo padre, entonces las nuevas consultas sobre el nodo nuevo serán también y únicamente del tipo padre. Esto se ha hecho siguiendo el principio de claridad y concisión que llevamos siguiendo durante todo el trabajo (no abrumar al usuario con demasiada información). Si el usuario desease conocer la información completa sobre uno de los nodos creados, entonces una nueva consulta sobre dicho nodo debe realizarse.

La última parte es referente a las categorías, que son accedidas realmente por medio de la API de Wikipedia, en vez de por consultas SPARQL.

Búsqueda y recopilación de Categorías

La API de acción de MediaWiki, (MediaWiki API), es un servicio web que proporciona un acceso a funciones wiki, datos y metadatos a través de llamadas de HTTP y una URL generalmente en api.php. Los clientes solicitan "acciones" particulares al especificar un parámetro de acción, principalmente ´acción = query´ para obtener información. Se lo conocía como la API de Wikipedia, pero ahora hay otras API web disponibles, como la REST API, (Wikimedia REST API, s.f.), y el servicio de consulta Wikidata que se ha usado anteriormente. En este trabajo se va a usar la API cada vez que uno de los nodos encontrados en los pasos anteriores sea del tipo categoría Commons, más concretamente la API:Categorymembers (API:Categorymembers).

```
const url =
"https://en.wikipedia.org/w/api.php?action=query&list=categorymembers&cmlimit=40&format=jso
n&cmtype=subcat|page&cmtitle=Category:" + category
```

Se recopilan tanto las subcategorías como los links a páginas, ya que éstas también pueden resultar interesantes al usuario, y pueden enlazarse luego de una manera similar a la que analizaremos las páginas de Wikipedia más adelante. Cabe notar que se ha reducido el número de posibles enlaces a 40, esto quiere decir que no se quiere recibir más de 40 links, sumando las subcategorías y páginas juntas, por los mismos motivos que se han ido explicando de no abrumar al usuario. Cabe la posibilidad de ampliar y continuar la consulta por donde se dejó, mediante 'cmcontinue' que proporciona un identificador para continuar una consulta anterior sin repetir los links ya proporcionados.

2.5. Análisis y extracción de datos no estructurados

En este apartado se realiza la exportación de datos y conceptos relevantes a la semilla a partir de una fuente de datos no estructurada como es Wikipedia. Prácticamente todos los conceptos de Wikipedia tienen su concepto equivalente en Wikidata, y

viceversa. Accediendo a los datos no estructurados se pueden encontrar y relacionar conceptos nuevos que de otra manera no aparecerían o se perderían, por ejemplos datos referentes a la historia o procedencia y orígenes, a discusiones, argumentos, opiniones, puntos de vista, referencias, datos económicos o geográficos no cuantificables…

Medida de relevancia

Para extraer estos conceptos se debe especificar primero qué datos se entenderán como relevantes y cuáles se omitirán por irrelevantes. Aunque existen muchas técnicas de extracción y análisis de información en textos, se ha intentado mantener un método sencillo a la par que eficaz. Se van a escoger aquellos elementos o conceptos que tengan sus propias páginas de Wikipedia, a menudo representados en la propia página como links. Esto tiene varias ventajas: primero, si un concepto tiene su propia página de Wikipedia, éste concepto debe ser relevante y digno de descripción, la problemática de la relevancia viene resuelta sin tener que recurrir a modelos o funciones de pesado; segundo, se eliminan las barreras lingüísticas al hacer el análisis. El formato de link se mantiene siempre igual y es descubrible a lo largo de todos los contenidos de la Web Semántica, sin embargo si fuésemos a hacer un estudio típico de descubrimiento de información de textos, deberíamos usar por ejemplo para cada lengua su conjunto determinado de stop-words (palabras que carecen de significado en un vocabulario dado), o sin tener que recurrir a corpus especializados para realizar los etiquetados léxicos o sintácticos.

Contents [hide]

Ilustración 4 Ejemplo de tabla de contenidos para la página de Blockchain. Si un concepto, por ejemplo Proof of Work, se encuentra bajo la sección 4.1, la uniremos a la semilla mediante dos relaciones: Blockchain (la semilla) → Tipos → Públicas → PoW

Inferencia de relaciones

Una vez extraídos los nuevos datos, se debe inferir qué tipo de relación los une con la semilla. Este apartado puede llegar a ser muy subjetivo, ya que un mismo concepto puede estar relacionado de múltiples maneras, además de que no existe una lista específica y limitada de relaciones, la propia Wikidata está en constante actualización. Para poder inferir el tipo de relación se puede realizar técnicas de procesamiento de lenguaje natural, entendiendo el significado que une ambos conceptos, mayormente buscando los verbos de las oraciones después de la correspondiente lematización del texto. Para agilizar el proceso, y de manera específica y única al análisis de textos de la Wikipedia se pueden clasificar los conceptos dependiendo de las secciones a las que pertenezcan. Mirando las tablas de contenidos se puede inducir la pertenencia y relación de cada concepto.

Las tablas de contenido (TOC) contienen enlaces a secciones dentro de un artículo, lo que proporciona una navegación en

la página fácil y automática. Los artículos o páginas de Wikipedia incluirán una tabla de contenidos predeterminada cuando aparezcan más de tres encabezados de sección. La tabla de contenido predeterminada incluirá todos los encabezados de sección y aparecerá justo antes del encabezado de la primera sección. La TOC siempre vendrá identificada con el id "toc" y estarán organizadas en niveles, secciones y subsecciones. Una vez que se obtengan los nombres e identificadores de todas las secciones, es posible buscar en el texto de la Wikipedia dichas secciones, entendiendo y buscando los links y enlaces que pertenezcan a cada una de ellas. Cada sección contendrá uno o varios elementos <p> envolviendo todo el contenido de la sección. Dentro de tales elementos es donde se centrará la búsqueda de links y conceptos.
Para unir cada concepto con la semilla se usará pues el nombre de la sección a la que pertenezca, teniendo en cuenta todos los niveles de subsecciones que se necesiten. Es decir, se podrán añadir tantos nodos y enlaces como sean necesarios hasta llegar al concepto encontrado.

Estructura del código fuente de una página típica de Wikipedia:

```
<div id="toc" class="toc">
<ul>
<li class="toclevel-1 tocsection-1"><a href="#Etymology"><span class="tocnumber">1</span>
<span class="toctext">Etymology</span></a></li>
<li class="toclevel-1 tocsection-2"><a href="#History"><span class="tocnumber">2</span> <span
class="toctext">History</span></a>
          <ul>
          <li class="toclevel-2 tocsection-3"><a href="#Ancient_Greeks"><span
          class="tocnumber">2.1</span> <span class="toctext">Ancient Greeks</span></a>
          </li>
          <li class="toclevel-2 tocsection-4"><a href="#Ancient_Romans"><span
          class="tocnumber">2.2</span> <span class="toctext">Ancient Romans</span></a>
          </li>
          </ul>
</li>
</ul>
</div>
<h2><span class="mw-headline" id="Etymology">Etymology</span></h2>
<p> Text with links that we'll use as concepts <a>Concept</a></p>
<h2><span class="mw-headline" id="History">History</span></h2>
<p> More text <a>This is a concept</a></p>
<h3><span class="mw-headline" id="Ancient_Greeks">Ancient Greeks</span></h3>
<p>More text </p>
<h3><span class="mw-headline" id="Ancient_Romans">Ancient Romans</span></h3>
<p> Last text example</p>
```

2.6. Entorno del servidor

Toda la parte del análisis y búsqueda de datos descrita en el apartado anterior se lleva a cabo en la parte del servidor. Se ha configurado el entorno en Heroku. Heroku es una plataforma en la nube que ofrece servicios bajo demanda (PaaS, del inglés Platform as a Service) y que es compatible con varios lenguajes de programación como son Java, Node.js, Scala, Clojure, Python, PHP, Ruby Go. Por esta razón, se dice que Heroku es una plataforma políglota, ya que permite al desarrollador crear, ejecutar y escalar aplicaciones de manera similar en todos los idiomas. En el caso de este trabajo

se ha usado el lenguaje Node.js, por su similitud con Javascript (usado en la parte de la interfaz) y por su sencillez a la hora de añadir librería o dependencias.

https://tfm-patriciamayotejedor.herokuapp.com/

La ventaja que tiene Heroku es que te permite alojar tus propias aplicaciones de manera gratuita en la versión 'light'. La versión light es más que suficiente para este proyecto, ya que simplemente queremos probar la funcionalidad y no está pensado como producto final. Debido a que el trabajo está en la versión la aplicación se pone en modo 'dormido' cuando lleva mucho tiempo sin usarse, esto puede hacer parecer al usuario que al principio no está funcionando, pero se recomienda paciencia y lanzar la primera consulta varias veces para 'despertar' al servidor.

Ajustes de seguridad

Se han añadido además sistemas de seguridad, garantizando el acceso por medio HTTPS en vez de HTTP, y configurando el uso compartido de recursos de origen cruzado (CORS). CORS es un mecanismo que permite que se soliciten recursos restringidos en una página web desde otro dominio fuera del dominio desde el que se sirvió el primer recurso. Ciertas solicitudes de "dominio cruzado", en particular las solicitudes de Ajax, están prohibidas por defecto por la política de seguridad del mismo origen y como desde nuestra interfaz las consultas se hacen usando AJAX se ha hecho necesario configurar este ajuste. CORS define una forma en que un navegador y un servidor pueden interactuar para determinar si es seguro o no permitir la solicitud de origen cruzado. La especificación para CORS se publicó originalmente como una Recomendación W3C (la institución en la que nos hemos estado basando prácticamente todo el trabajo) pero ese documento es obsoleto ya que la especificación actual mantenida activamente que define CORS es el estándar de vida de búsqueda. Se ha importado la librería 'cors'.

Rutas abiertas

Se han abierto tres rutas o APIs de consultas al servidor. La primera ruta es para acceder a la parte de datos estructurados mediante SPARQL y Wikidata. Es accesible mediante la extensión: **'/wikiquery?id=<WikidataID>'**
También es posible especificar el nivel de profundidad en la búsqueda de nodos padres e hijos modificando la extensión:
'/wikiquery?id=<WikidataID>&deep=<nivel de profundidad>'
Éste nivel de profundidad tiene como valor predeterminado '2', ya que se ha considerado que a partir de más niveles la relación con el concepto semilla empieza a perder relevancia, además de corroborarse posteriormente en los test con los usuarios. La razón por la que se permite cambiarlo es sobre todo como testeo, además de simplemente dar al usuario mayor control sobre las consultas. Pudiese resultar más útil si el objetivo de la consulta no es para mostrar los datos en la interfaz como metodología de aprendizaje, sino que el motivo reside en usarlo como dataset para

una aplicación de Inteligencia Artificial. Éstas aplicaciones pueden soportar gran cantidad de datos y entonces sí que podría resultar relevante.

Los resultados que se devuelven son los siguientes, en formato JSON: Un conjunto de nodos 'graphNodes', que se identifican por la propiedad 'id', y que constan de 'label' y 'description' para que también puedan identificarse por el usuario (inferir el sentido semántico del nodo). También esta compuesto por la propiedad 'group', que indica mediante índices la propiedad que se ha aplicado para que surja dicho nodo. Por último la propiedad 'position' indica la posición o el nivel de representación del nodo. El nodo semilla tendrá la posición 0, la primera línea de padres tendrá la -1, el segundo nivel de profundidad de padres tendrá el -2, la primera línea de hijos tendrá la 1, el segundo nivel de profundidad de hijos tendrá el 2… y así sucesivamente. La justificación en las reglas que se han seguido para esto es sencilla si uno se imagina una regla o metro puesto de manera horizontal (… -2, -1, 0, 1, 2 …), y tal como se espera, este valor tendrá implicaciones directas en la representación posterior en la interfaz.


```
{
  "graphNodes": [
    {
      "id": "http://www.wikidata.org/entity/Q18545",
      "label": "ball",
      "description": "round object",
      "group": 0,
      "position": 0
    }],
  "graphLinks": [
    {
      "source": "http://www.wikidata.org/entity/Q768186",
      "target": "http://www.wikidata.org/entity/Q18545",
      "type": "subclassOf"
    }],
  "graphCategories": [
    {
      "id": "Balls",
      "subcategories": [
        {
          "id": 996038,
          "title": "Category:Ball games"
        }],
      "pages": [
        {
          "id": 3928,
          "title": "Ball"
        }]
     }]
}
```


La segunda ruta es para la búsqueda y análisis de páginas web de Wikipedia. **'/page?url=<Wikipedia URL>'**

Se ha importado la librería 'cheerio' para las técnicas de raspado web y el análisis de textos. Con Cheerio[6] se analiza el marcado propio de HTML en páginas web y proporciona una API para atravesar / manipular la estructura de datos resultante. Es importante destacar que no interpreta el resultado como lo hace un navegador web. Específicamente, no produce una representación visual, aplica CSS, carga recursos externos o ejecuta JavaScript, simplemente facilita la navegación y la interacción entre etiquetas.
Los datos devueltos seguirán el siguiente formato: el identificador del elemento junto a su tabla de contenidos, TOC, correspondiente. La TOC estará formada por un conjunto de secciones, que a su vez contienen los datos relevantes para identificar la sección y la lista de enlaces que se encuentran en ella.

```
{
  "wikidataID": "Q18545",
  "toc": [
    {
      "section": {
        "id": "#",
        "title": "Extract",
        "number": 0
      },
      "content": [
        {
          "link": "https://en.wikipedia.org/wiki/Sphere",
          "title": "Sphere"
        }]
    }
}
```

Por último, la tercera ruta, que tiene un sentido mayormente auxiliar.
'/entity? =<Wikipedia URL>'
Con esta ruta se pretende convertir cualquier enlace a Wikipedia en un nodo posible de usar en el contexto de este proyecto, siguiendo el formato de la librería D3.js. Esto es útil y necesario en la interacción del usuario con la plataforma, ya que se permite añadir nodos de manera individual al grafo o resultado final, sin tener que por ello cargar todo el proceso cognitivo. En el siguiente apartado se explicarán las formas en las que el usuario puede hacer tales acciones, interaccionando con la página.

[6] https://cheerio.js.org/

Capítulo 3:
Experiencia de usuario

La segunda parte del trabajo es aquella en la que el usuario va a tener acceso. Constará de la interfaz desde donde se introducirá la semilla o semillas de las que se quiere aprender. Ayudándose de tecnologías como D3.js, una librería para páginas basadas y dirigidas por datos. La experiencia de usuario se va a dividir en tres partes o fases. La primera fase será la de crecimiento con filtrado, en ella se le mostrarán al usuario una serie de nodos y conceptos enlazados iniciales a partir de la semilla introducida por él. La segunda fase es la de expansión, en ella se estudiarán métodos de navegación e interacción que faciliten el descubrimiento de conceptos y relaciones, se le permitirá al usuario interaccionar con los conceptos y la interfaz, añadiendo, quitando o editando la información inicial. Por último la tercera fase será la de cierre. Se usarán métodos de 'Clustering' para la integración final de los datos seleccionados por el usuario y se proporcionará la posibilidad de exportar la red semántica creada.

La interfaz se encuentra ejecutándose en Netlify. Netlify es una plataforma que permite realizar depliegues continuos y de manera gratuita de sitios web estáticos.

https://friendly-banach-15e173.netlify.com/

3.1. Interfaz e interacción con el usuario

Esta parte del trabajo constará de la interfaz desde donde se introducirá la semilla o semillas de las que el usuario quiere aprender ayudándose de tecnologías como D3.js, una librería para páginas basadas y dirigidas por datos. La experiencia de usuario se va a dividir en tres partes o fases. La primera fase será la de crecimiento con filtrado, en ella se le mostrarán al usuario una serie de nodos y conceptos enlazados iniciales a partir de la semilla introducida por él. La segunda fase es la de expansión, en ella se estudiarán métodos de navegación e interacción que faciliten el descubrimiento de conceptos y relaciones, se le permitirá al usuario interaccionar con los conceptos y la interfaz, añadiendo, quitando o editando la información inicial. Por último la tercera fase será la de cierre. Se usarán métodos de 'Clustering' para la integración final de los datos seleccionados por el usuario y se proporcionará la posibilidad de exportar la red semántica creada.

Fase 1: Crecimiento con filtrado

El usuario quiere aprender sobre un concepto, desea entender los distintos contextos en los que se puede aplicar y resulta relevante, desea investigar sus usos o sus partes, necesita de un esquema o guía para escribir en sistema de autoría propio, busca construir un dataset o un conjunto de conceptos para un sistema inteligente… El

usuario conoce de nuestro trabajo y abre la aplicación. La página aparece vacía ya que lo primero que debe hacer es introducir en la barra de búsqueda el concepto del que desea conocer más. A éste concepto se le denomina semilla, ya que a raíz de él van a surgir muchos más conceptos en formas de ramas relacionadas semánticamente.

La barra de búsqueda tiene una función integrada de autocompletado. Esto es muy importante ya que las bases de conocimiento que se están usando en este trabajo son muy específicas, (Wikipedia y Wikidata). El autocompletado mostrará una lista de sugerencias filtradas por el texto introducido por el usuario. Esta lista estará formada de conceptos que tengan su propia página en Wikidata, y estarán unidos a su propio identificador único (los identificadores en Wikidata siempre comienzan con 'Q' seguido de números enteros). Por ejemplo si el usuario escribe ball le aparecerá una lista de posibilidades: round object, dance party, in mathematics - space inside a sphere, family name… Ya que no es posible entender los distintos significados basándose únicamente en el identificador, las propiedades de 'label' y 'description' se vuelven indispensables. Se permiten también otras formas de búsqueda, por ejemplo si el usuario conoce de antemano el identificador de la semilla entonces puede ponerlo directamente. Por último se ha integrado la posibilidad de buscar directamente mediante un enlace a una página de Wikipedia.
https://www.wikidata.org/w/api.php?origin=*&action=wbsearchentities&format=json&language=en&type=item&continue=0&search=ball

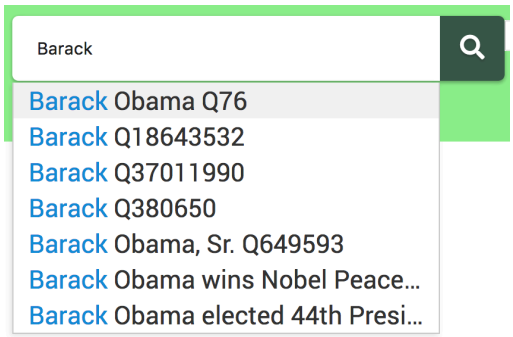


Ilustración 5 Función de autocompletado con sugerencias

Una vez la semilla haya sido elegida y seleccionada por el usuario, todo el mecanismo interno se pone en marcha. Desde la interfaz se hacen dos llamadas de manera simultánea al servidor. La primera tiene como objetivo la creación del grafo inicial a partir de Wikidata, además de la búsqueda de categorías con páginas relacionadas y la segunda tiene como motivo el análisis de los datos no estructurados de la página de Wikipedia correspondiente a la semilla.

https://tfm-patriciamayotejedor.herokuapp.com/wikiquery?id=Q18545
https://tfm-patriciamayotejedor.herokuapp.com/page?url=https://en.wikipedia.org/wiki/Ball

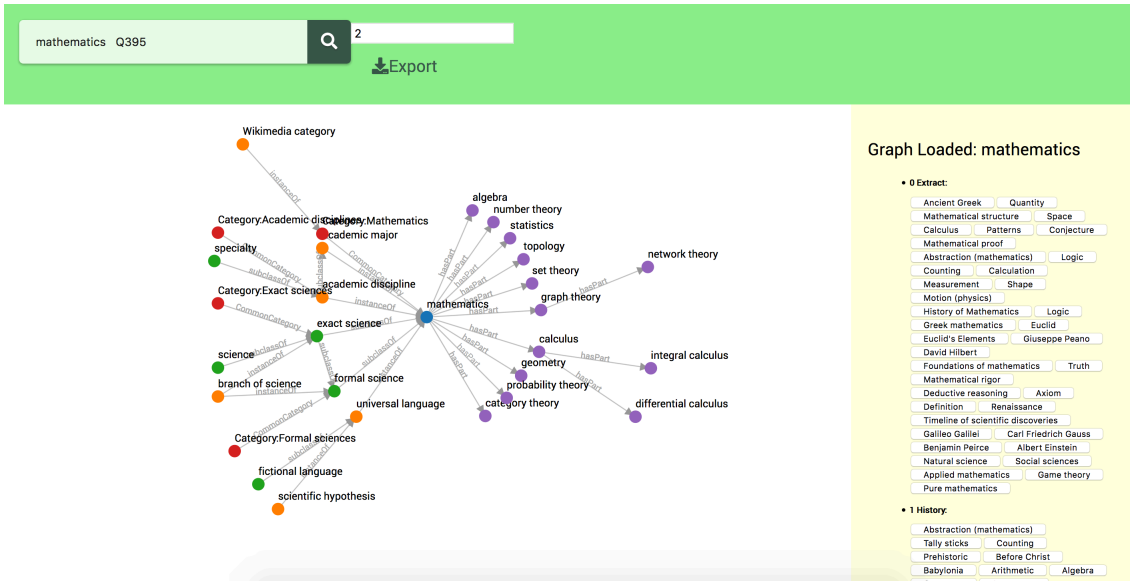


Ilustración 6 Vista general de la aplicación, con el grafo construido y los conceptos del TOC visibles

Estos resultados se muestran por pantalla de manera clasificada y ordenada. La parte de Wikidata se muestra en forma de grafo dirigido, posicionando en el centro el nodo semilla. A la izquierda aparecerán los nodos padres, ordenados por niveles. A la derecha aparecerán los nodos hijos, también ordenados por niveles. También se ha de destacar la inclusión de nodos 'Categoría', que son tratados como padres, pero que necesitan de especial atención ya que contienen mucha información extra. Todos los datos referentes a la navegación por categorías se muestran en tablas justo debajo del grafo principal. La parte de Wikipedia aparecerá en la parte derecha, clasificada y ordenada por secciones extraídas directamente desde la Tabla de Contenidos.

Fase 2: Exploración y Expansión

El usuario ha recibido por pantalla mucha información tras introducir la semilla. Ya empieza a estructurar los conceptos y a entender como se relacionan entre ellos, sin embargo quiere interactuar y experimentar, quiere descubrir más, y ¿por qué no? quiere corregir el modelo resultante. Prácticamente todos los conceptos que se muestran son accionables.

En la parte de Wikipedia, pasar el ratón por encima de los conceptos mostrará una pequeña alerta, con una imagen si procede, y un pequeño extracto de información sobre el concepto. Este modo de interacción con el usuario es muy útil y rápido, de primeras no se muestran los detalles específicos a cada concepto, ya que esto podría resultar abrumador. Sin embargo simplemente pasando el ratón por encima desencadena la alerta, que se elimina tan pronto como el cursor se mueve hacia otro lado. Se permite así mostrar información sólo cuando sea útil y relevante, sin necesidad de actualizar toda la página para entender un nuevo concepto. Si se desea conocer incluso más detalles, se puede ver el enlace, lo que producirá, esta vez sí, una actualización de toda la página, tomando como referencia y semilla el concepto seleccionado guardando la referencia a la semilla previa, mediante una relación inferida directamente desde la sección en la que se encuentra dentro de la Tabla de Contenidos. La semilla previa aparecerá como padre del nuevo concepto que se va a visualizar.

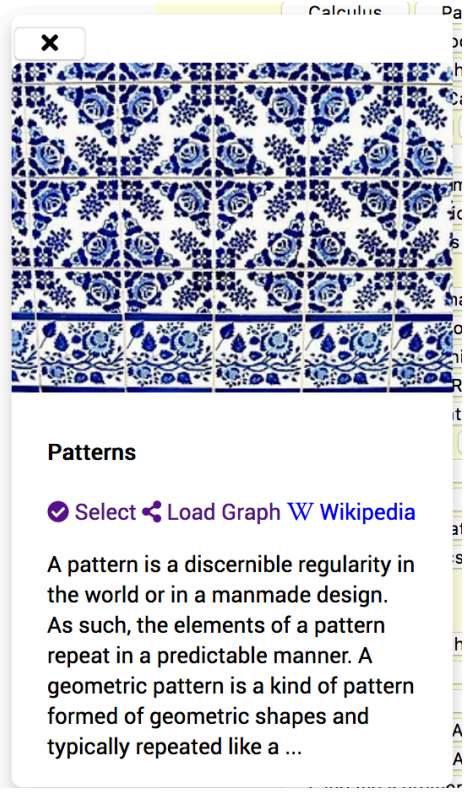


Ilustración 7 Ejemplo de Alerta. Contiene la imagen y el extracto del concepto, así como las acciones disponibles

En la parte de Wikidata encontramos un comportamiento similar. Pasar por encima de los nodos nos mostrará el mismo tipo de alerta con una imagen y una pequeña descripción del nodo a modo de extracto. Cada nodo a su vez es seleccionable o no para el resultado final. Cada nodo además puede llevar a una actualización de la página convirtiéndose en semilla. Las categorías y páginas sin embargo se comportan de un modo diferente, están ahí para proporcionar un sentido globalizador a la semilla, para aportar contexto y estructura. Se pueden seleccionar y navegar entre ellas, pero éstas no refrescarán la página ni el grafo original, ya que no se tratan de conceptos o entidades en sí mismos y sólo cobran sentido cuando se relacionan con un nodo en concreto.

Como resumen, hay tres tipos de interacciones:
Vista Previa: a modo de alerta, contiene una imagen si procede y un pequeño extracto en relación al concepto o nodo al que hacen referencia. Los datos se sacan directamente de la API de Wikipedia
https://en.wikipedia.org/api/rest_v1/page/summary/Ball

Explorar: se recarga la página con el nodo seleccionado como nueva semilla. Se guardará siempre la relación con la semilla previa de la que proviene. Se crea un nuevo grafo y se hace un nuevo análisis en Wikipedia procesando la nueva página seleccionada. Esta acción funciona de manera distinta en la sección de categorías y páginas, ya que no recargan la página entera, sólo la tabla en la que aparecen (no se buscan grafos ni páginas de Wikipedia con las categorías)

Seleccionar: Todos los conceptos de la página se pueden seleccionar y añadir al grafo final. También se puede deseleccionar. En el modo automático, todos los conceptos, tanto de Wikidata como de Wikipedia se interpretarán como seleccionados por defecto. En el modo de interacción con el usuario, sólo aquellos que aparecen inicialmente en el grafo dirigido.

Fase 3: Cierre

La última fase es la fase del cierre, el usuario ha experimentado, ha jugado con la aplicación, ha explorado lo suficiente y siente que ya ha conseguido recolectar todos los datos necesarios y relevantes que buscaba. Seleccionando el botón de exportar el usuario podrá descargar un resumen de todos los nodos y datos seleccionados, en formato JSON, de manera muy similar a la forma en la que el servidor realiza la primera búsqueda. El documento incluirá un conjunto de nodos, otro de links y otro de categorías y páginas. Si se tratase de simplemente una interacción automática para dar soporte a una aplicación de IA entonces el documento incluirá los datos procedentes de la primera semilla, si por el contrario, ha habido interacción con el usuario, entonces se incluirán los datos customizado y personalizados por éste.

3.2. Estadísticas y análisis

Tal y como se ha mencionado anteriormente, durante la definición de la ontología y durante el apartado de análisis de los datos estructurados, se ha realizado un filtrado a la base de conocimiento de Wikidata, antes de mostrarla y enviarla al usuario. Las propiedades o tipo de relaciones filtradas han sido escogidas cuidadosamente por aportar y reflejar mayor sentido contextual. Este filtrado no solo tiene como objetivo resultar más esclarecedor para el usuario, sino que resulta totalmente necesario si se quieren reflejar relaciones con la semilla más haya de las estrictamente directas a un solo nivel. El número de relaciones, propiedades y valores crece exponencialmente según se avanza en padres/hijos. Cabe por último añadir una última razón y justificación al filtrado de propiedades; Wikidata también se usa como base de datos de identificadores de una misma entidad entre múltiples estándares. Esto quiere decir que la mayor parte de las propiedades de la 'semilla' simplemente serán identificadores a bases de conocimiento ajenas a Wikidata, que contienen una entidad con el mismo significado.

Aunque el filtrado se lleve a cabo con plena justificación se ha decidido incluir en los datos (junto a los nodos, enlaces y categorías) un apartado dedicado expresamente a mostrar la cantidad de nodos y tipos de enlace que se habría mostrado en caso de no usar el filtrado. Esto puede ser útil para dos casos desde el punto de vista del usuario: i) el usuario valora más la aplicación pudiendo ver el esfuerzo que hay detrás de ella, ii) en caso de que no se muestre por pantalla una propiedad concreta buscada, se entenderá que es porque ha sido filtrada y no por fallo de la aplicación misma. Desde

el punto de vista científico, los datos mostrados justifican y muestran el comportamiento intrínseco de las bases de datos enlazados, donde las relaciones con otras entidades crece de manera exponencial.
Se han usado llamadas a la API de Wikidata.

```
const url =
"https://www.wikidata.org/w/api.php?action=wbgetentities&props=claims&languages=en&format=j
son&ids=
+ id1|id2|id3…
```

Los tipos de enlaces referentes una entidad se llaman 'claims' en la API de Wikidata. Dentro de cada 'claim' puede haber referencia a otra entidad o un conjunto de ellas. Se analizarán entonces tanto el conjunto de 'claims' (tipos de enlace), como el conjunto de entidades integradas en cada 'claim' (número de nodos que aparecerían en el grafo final si no fuese por el filtrado). Se lleva a cabo también un trabajo de eliminación de duplicados, ya que se intentan mostrar los datos tal y como aparecerían en la aplicación. En el ejemplo de abajo se muestran los resultados tras hacer la llamada al servidor de 'Dog Q144' con un nivel Deep 2.


```
{"analytics": {
    "filteredNodes": 15,
    "unfilteredNodes": 83,
    "filteredLinkTypes": 4,
    "unfilteredLinkTypes": 97
  }}
```

Capítulo 4:
Experimentación y testeo

En este apartado se van a realizar un conjunto de experimentos con tres propósitos principales y que a su vez están directamente relacionados con la estrategia de interacción seguida basada en fases: (i) corroborar la sintonización correcta de parámetros en la primera fase de filtrado, (ii) corroborar la utilidad de la plataforma creada en contextos de aprendizaje en la segunda fase de interacción y (iii) corroborar la utilidad de la plataforma en aplicaciones inteligentes en la última fase de cierre.

En la primera parte de preparación se van a definir y asentar los datos usados, a continuación se van a describir las características del conjunto de usuarios con los que se van a llevar a cabo los test y por último se describen los experimentos, con una breve descripción de la metodología a seguir y los pasos de preparación necesarios.
En la segunda parte de este capítulo se van a analizar y discutir los resultados de los experimentos, evaluando así si se han alcanzado o no los propósitos buscados.

4.1. Datos

Wikidata se está convirtiendo en una base de conocimiento cada vez más importante cuyo uso se está extendiendo en la comunidad de investigación. Sin embargo, la mayoría de los conjuntos de datos de evaluación de sistemas de preguntas y respuestas se basan en Freebase o DBpedia. En el trabajo llevado a cabo por Dennis Diefenbach y su equipo (Diefenbach, Tanon, Singh, & Maret, 2017) se presentan dos nuevos conjuntos de datos con el fin de capacitar y comparar sistemas de Preguntas y Respuestas sobre Wikidata. El primer conjunto de Diefenbach es una traducción del popular conjunto de datos SimpleQuestions[7] a Wikidata, el segundo es un conjunto de datos creado mediante la recopilación de comentarios de los usuarios. Se puede usar el mismo sistema y conjunto de datos en nuestro trabajo. De nuevo el problema con este tipo de conjunto de datos es que en general se experimenta con preguntas cuantitativas y medibles, del tipo, ¿quién es el alcalde de Madrid? ¿Cuánta gente vive en Polonia? ¿Cuál es la capital de Francia? … En este proyecto el enfoque reside en proporcionar contexto en torno a un ejemplo, por lo que se han omitido todo tipo de propiedades que proporcionasen datos específicos como éstos. Se escogerán pues de este conjunto de datos aquellos que concuerden con los objetivos de este trabajo. Esta elección tendrá que ser manual pero también hay posibilidad de automatización, ya que se pueden escoger sólo aquellas preguntas del dataset que contengan una sola palabra (al no ser preguntas, se entiende que el usuario busca conocimiento general y contextualización sobre esa palabra-concepto). El dataset se encuentra en github[8] con

[7] https://research.fb.com/downloads/babi/
[8] https://github.com/WDAqua/WDAquaCore0Questions

acceso abierto y libre. Una demo en vivo también se puede encontrar en la web www.wdaqua.eu/qa.
Para el último experimento sobre el uso de la plataforma en una aplicación inteligente se han recopilado textos de manera aleatoria por internet. Éstos textos se encuentran en su totalidad en el Anexo y cumplen ciertas características que se describen con mayor detalle en el apartado de preparación del experimento.

4.2. Usuarios

Se han realizado los tests con 24 usuarios. EL rango de edad variaba entre los 20 y los 36 años. En total han participado 15 mujeres y 9 hombres. Las ocupaciones de estos usuarios varían y son diversificadas, desde Humanidades y Ciencias Sociales, Marketing, Abogacía, Industria del Juego y Apuestas, Economía, Finanzas, Ingeniería Industrial, Ingeniería Informática, Desarrollador de aplicaciones móviles o Inteligencia Artificial. Las entrevistas se han llevado a cabo en persona, uno a uno. En el test sobre la profundidad, no se les ha explicado con detalle la plataforma ni los propósitos del trabajo, ya que simplemente se evaluaba la relevancia de los datos. En las preguntas sobre conocimiento se les ha explicado brevemente el propósito y las intenciones, de tal manera que interactúen con la plataforma con el propósito de aprender o con vista a la consecución de un grafo útil personalizado. En cada pregunta se les ha evaluado el conocimiento previo sobre la semilla o concepto a consultar. De esta manera se evalúa lo aprendido nuevo, y no lo que ya se sabía.

4.3. Experimentos

Test 1: Fase 1 de visualización con filtrado

Descripción: El filtrado inicial de los nodos se hace absolutamente necesario y se convierte en requisito cuando se usan bases de conocimiento tan grandes y robustas. La profundidad de datos a partir de 3 niveles pierde el sentido y se convierte en irrelevante.
Metodología: En este testeo se van a cuantificar primero el número de links, conceptos y categorías que aparecen con el cuidadoso filtrado estudiado en este proyecto. Estos datos se van a comparar a continuación con una búsqueda limpia de la semilla del usuario, es decir, se van a recoger todos y cada uno de las propiedades y links asociadas a ella. Se comprobará entonces la necesidad del filtrado inicial, ya que la información crece exponencialmente si no es tratada (además de incluir datos no relevantes). Para medir estos datos se ha usado el apartado de 'analytics' del servidor. Para medir el nivel de profundidad se harán 3 intentos con cada usuario, incrementando la profundidad de las relaciones en cada intento, 1, 2 y 3. Se comprobará así hasta que punto deja de tener sentido mostrar más información preguntando con qué nivel se sienten más cómodos.
Preparación: Se procede a la elección de preguntas sacadas directamente de la base de preguntas desarrollada por Dennis Diefenbach y su equipo. Se ha creado la siguiente

plantilla a rellenar por los usuarios. No se les va a dejar a los usuarios ningún tipo de interacción con la herramienta para esta primera etapa.

	Nodos/Tipos de Enlace Con Filtro			Nodos/Tipos de Enlace Sin Filtro		
Pregunta	Deep = 1	Deep = 2	Deep = 3	Deep = 1	Deep = 2	Deep = 3
Barack Obama Q76						
Paris Q90						
Ottawa Q1930						
UN Q1065						
Apple (empresa) Q312						
Apple (fruta) Q89						
Moon Q405						
Car Q1144312						
J. K. Rowling Q34660						
Aluminium Q663						
Eiffel Tower Q243						
Mathematics Q395						

Tabla 1- Plantilla de análisis de la cantidad de nodos, con filtro y sin filtro, a distintos niveles de profundidad

Test 2: Fase 2 de interacción

Descripción: La visualización estructurada haciendo hincapié en las relaciones permite crear un conocimiento más profundo sobre un concepto dado. Por otra parte, la navegación y la interacción de forma visual con datos favorece al descubrimiento y afianza el aprendizaje en las personas.
Metodología: Usando la plantilla de preguntas del apartado anterior comprobaremos si el usuario siente que conoce más un concepto comparándolo con la demo en la web del dataset de Dennis Diefenbach y con nuestro sistema, filtrado, basado y enfocado en las relaciones. Se comprobará por medio de preguntas si el usuario ha conseguido arraigar lo aprendido y si ha conseguido conceptualizar y contextualizar la ´semilla´.
Preparación: Igual que en el test anterior, con la misma tabla de preguntas. Se intenta cuantificar el conocimiento previo, el posterior tras la interacción con la herramienta (3 minutos), el sentimiento de abrumación (0 es muy abrumado, 10 es muy sencillo) y la puntuación general a la plataforma (valoración subjetiva de la utilidad, la facilidad de uso…)

Pregunta	Conocimiento Previo	Conocimiento Posterior	Sentimiento: Abrumado - Sencillo	Puntuación 0 - 10
Barack Obama				
Paris				
Ottawa				
UN				
Apple (empresa)				

Apple (fruta)				
Moon				
Bike				
J. K. Rowling				
Aluminium				
Eiffel Tower				
Mathematics				

Test 3: Fase 3 de cierre con aplicación práctica

Descripción: El filtrado automático, así como la entrada manual del usuario son útiles para asentar las bases de un sistema de recomendación o clasificación inteligente.
Metodología: Se va a testear y usar el grafo resultante de la aplicación en un ejemplo real, un artículo o pieza de texto donde el tema principal sea el concepto o semilla estudiado y se mencione a ésta de manera literal. También se testeará con una pieza de texto que resulte totalmente irrelevante a la semilla. Por último se comparará también con un texto que brevemente mencione la semilla o concepto, pero cuyo tema principal sea algo diferente y viceversa, un texto que no contenga la semilla de forma literal escrita, pero que se entienda que ésta sea el tema principal por medio del contexto. Se medirá la correlación entre el concepto y los textos, intentando averiguar si el tema principal del texto es o no la semilla o concepto estudiado.

```
is this text related to Apple?:    false


 ==========================


 Conquer The Big Apple: 30 Things To Do In New York City. Home to over 8 million people,
New York is the most populous city in the United States. Having been depicted in numerous
films like Breakfast and Tiffany's and Goodfellas, New York is now often associated with
Wall Street's soaring skyscrapers and monuments, the neon signs of Times Square and the
greenery of Central Park, all contributing to the unfading energy of the city. From
Broadway shows, to world-class museums like the Museum of Modern Art, to the iconic Statue
of Liberty, this exciting city is just brimming with activities to entertain every
traveler. It's hard to cover all that New York has to offer with so many attracting
options to choose from, and it can make planning your trip a bit baffling. That's why this
list will come in handy. Here are the 30 must-dos, ranging from the iconic landmarks to
local favorites


 ==========================
```

Ilustración 8 Ejemplo del test 3, donde se refleja que el texto no habla sobre la empresa Apple, aunque se haga mención a 'The Big Apple', refiriendose a la ciudad de Nueva York

Preparación: Para llevar a cabo esta prueba se ha usado la librería 'lda'[9] que lleva a cabo el modelo probabilístico generativo de Latent Dirichlet allocation (LDA) que se realiza sobre una colección de compuestos formados por partes. En el procesamiento del lenguaje natural, el LDA y más concretamente en términos de modelado de temas, los compuestos son documentos y las partes son palabras y / o frases (n-grams). LDA fue presentado por primera vez como un modelo gráfico para el descubrimiento de temas por David Blei, Andrew Ng y Michael I. Jordan en 2003.

[9] https://github.com/primaryobjects/lda

Si se ve el número de temas como el número de clústeres posibles en un texto y las probabilidades como la proporción de pertenencia al clúster, entonces el uso de LDA es una forma de suave de agrupación de sus componentes y partes. Contraste esto con otros métodos como k-means donde cada entidad solo puede pertenecer a un clúster. Estas membresías difusas proporcionan una forma más matizada de recomendar artículos similares o encontrando duplicados.

El LDA se configurará de la siguiente manera. Se centrará en la búsqueda de un solo tema principal, y de 5 parámetros o conceptos importantes y relevantes que reflejen dicha tenacidad. Cada parámetro vendrá acompañado de la probabilidad de que el texto analizado hable sobre él. Estos parámetros también vendrán ordenados de mayor a menor, siendo entonces el tema o concepto más probable el primer parámetro de la lista.

Este modelo se ha combinado con el modelo de bolsa de palabras en una representación simplificadora utilizada en el procesamiento del lenguaje natural y la recuperación de información (IR). En este modelo, un texto (como una oración o un documento) se representa como la bolsa (multiset) de sus palabras, sin tener en cuenta la gramática e incluso el orden de las palabras, pero manteniendo la multiplicidad. El modelo de bolsa de palabras se usa comúnmente en métodos de clasificación de documentos donde la (frecuencia de) ocurrencia de cada palabra se usa como una característica para entrenar un clasificador. En el caso de este proyecto la bolsa de palabras es la lista de conceptos sacadas de los nodos, enlaces Wikipedia, categorías y páginas. Los textos usados se han añadido en el Anexo.

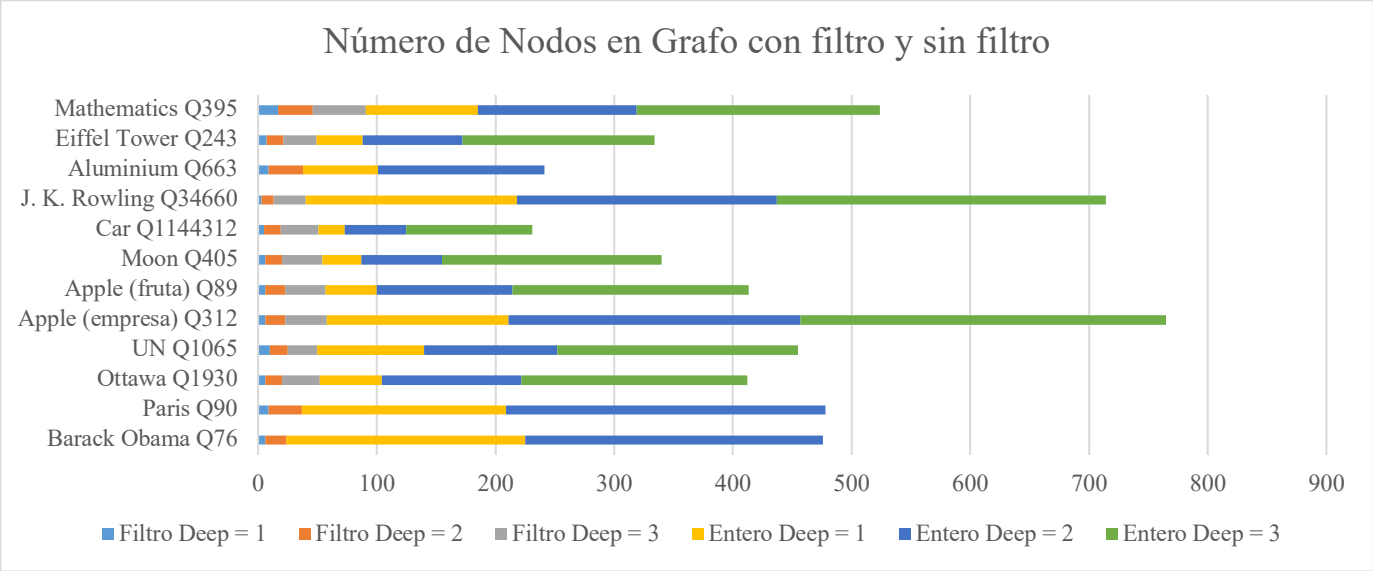
4.4. Discusión y resultados

Resultados Test 1

La primera parte a evaluar es la de la efectividad del filtrado. Se empieza entonces demostrando la necesidad imperativa de hacer un filtrado y escoger los tipos de enlaces a la hora de conceptualizar una semilla. Se ha llamado al servidor para las doce preguntas diseñadas anteriormente en la plantilla. Además se evalúa cada semilla en tres niveles de profundidad diferentes, demostrando así el crecimiento exponencial típico de bases de conocimiento relacionales.

	Nodos Con Filtro			Nodos Sin Filtro		
Pregunta	Deep = 1	Deep = 2	Deep = 3	Deep = 1	Deep = 2	Deep = 3
Barack Obama Q76	6	18	-	201	251	-
Paris Q90	9	28	-	172	269	-
Ottawa Q1930	6	14	32	52	118	190
UN Q1065	10	15	25	90	112	203

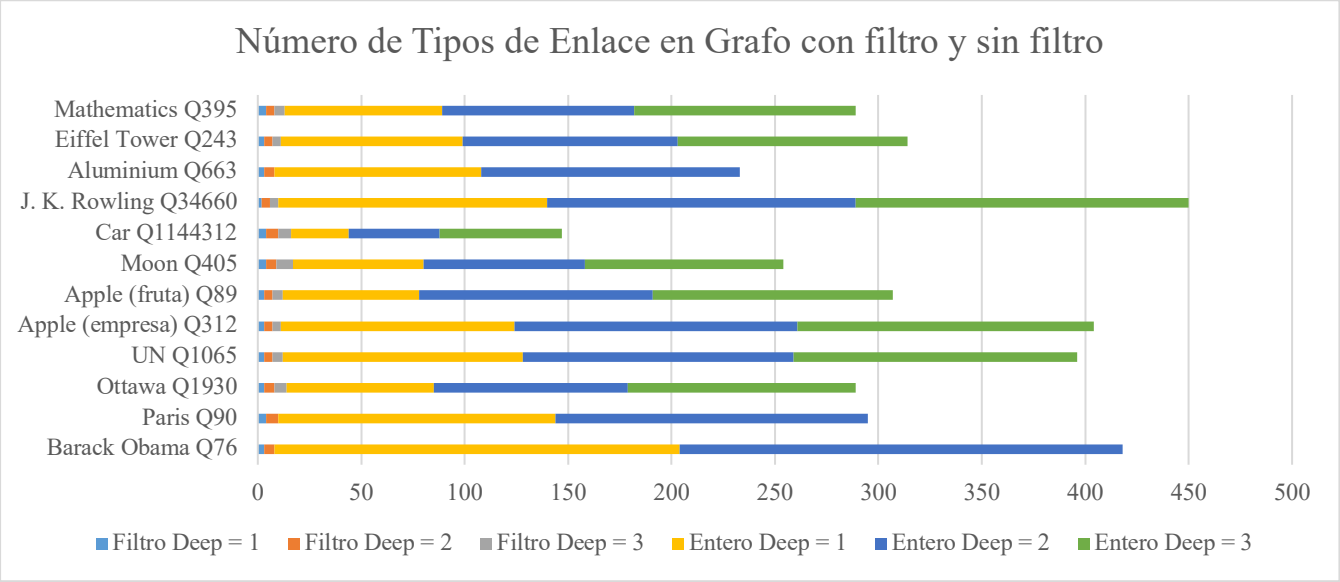
Apple (empresa) Q312	6	17	35	153	246	308
Apple (fruta) Q89	6	17	34	43	114	199
Moon Q405	6	14	34	33	68	185
Car Q1144312	5	14	32	22	52	106
J. K. Rowling Q34660	3	10	27	178	219	277
Aluminium Q663	9	29	-	63	140	-
Eiffel Tower Q243	7	14	28	39	84	162
Mathematics Q395	17	29	45	94	134	205



Si se hace un enfoque en los datos referentes a los nodos de las tablas de arriba se ve primero que aumentar **los niveles de profundidad multiplica por un factor aproximado de 2 el número de nodos encontrados**. Ocurre de manera similar tanto en el caso de los nodos con filtro, como en el de los nodos encontrados mediante la búsqueda exhaustiva de Wikidata. Sin embargo, la cantidad de nodos se ha reducido considerablemente con el método propuesto en este trabajo. Se nota que el número de nodos es razonablemente más pequeño, simplificado y manejable cuando se han usado los filtros. Se entiende que un grafo representado de manera visual con más de 200 nodos pierde relevancia y sentido para el usuario, además de resultar abrumador e innecesario.

	Tipos de Enlace Con Filtro			Tipo de Enlace Sin Filtro		
Pregunta	Deep = 1	Deep = 2	Deep = 3	Deep = 1	Deep = 2	Deep = 3
Barack Obama Q76	3	5	-	196	214	-
Paris Q90	4	6	-	134	151	-
Ottawa Q1930	3	5	6	71	94	110
UN Q1065	3	4	5	116	131	137
Apple (empresa) Q312	3	4	4	113	137	143
Apple (fruta) Q89	3	4	5	66	113	116

Moon Q405	4	5	8	63	78	96
Car Q1144312	4	6	6	28	44	59
J. K. Rowling Q34660	2	4	4	130	149	161
Aluminium Q663	3	5	-	100	125	-
Eiffel Tower Q243	3	4	4	88	104	111
Mathematics Q395	4	4	5	76	93	107



El filtro esta basado en la reducción de los tipos de enlace disponibles y en efecto, se puede mostrar la efectividad del filtrado con los datos de la tabla de arriba, donde el máximo número ha sido sólo el uso de 8 tipos, con un nivel 3 de profundidad en la pregunta sobre la luna, 'Moon Q405'. De manera distinta que con el número de nodos, los tipos de enlace son limitados, y pocas veces se modifican en Wikidata, por lo que avanzar en los niveles de profundidad no descubre muchos nuevos tipos, usando prácticamente los mismos una y otra vez. De manera cuantitativa, **la cantidad de tipos de enlace crece simplemente con un factor aproximado de 1,2**.

Resultados Test 2

En este test se va a evaluar las experiencias y preferencias del usuario. Primero se evalúa el conocimiento ganado y la sensación de utilidad de la plataforma. Las preguntas se han realizado con un nivel de Deep=2.

Pregunta	Conocimiento Previo: 0 - 10	Conocimiento Posterior: 0 - 10	Sentimiento: Abrumado - Sencillo	Puntuación: 0 - 10
Barack Obama Q76	3	5	8	8
Paris Q90	4	6	8	8
Ottawa Q1930	2	7	8	8
UN Q1065	3	6	8	8

Apple (empresa) Q312	4	5	4	8
Apple (fruta) Q89	3	4	4	8
Moon Q405	3	5	8	8
Car Q1144312	6	7	8	8
J. K. Rowling Q34660	2	4	8	8
Aluminium Q663	0	3	7	8
Eiffel Tower Q243	5	6	8	8
Mathematics Q395	5	5	4	8
Media	**3,33333333**	**5,25**	**6,91666667**	**8**

Se puede comprobar que el conocimiento sobre un determinado concepto después de haber usado la herramienta ha incrementado en casi 2 puntos sobre 10. Los usuarios han aprendido datos nuevos, descubierto las distintas relaciones y contextos en los que se puede desarrollar el concepto. Al enseñar a los usuarios las páginas de referencia de donde se han extraído los datos, además de la plataforma de Metaphacts[10] (usada como referencia a mejorar), casi todos han estado de acuerdo que la herramienta creada en este trabajo ha simplificado de manera eficaz el proceso de aprendizaje del usuario.

A continuación se evalúa la preferencia del usuario al nivel de profundidad en la búsqueda de conceptos. Se presenta la misma pregunta con los tres niveles distintos.

Pregunta	Preferencia en la profundidad
Barack Obama Q76	2
Paris Q90	1
Ottawa Q1930	2
UN Q1065	2
Apple (empresa) Q312	2
Apple (fruta) Q89	1
Moon Q405	2
Car Q1144312	2
J. K. Rowling Q34660	3
Aluminium Q663	2
Eiffel Tower Q243	2
Mathematics Q395	1
Media	**1,83**

[10] https://wikidata.metaphacts.com/resource/app:Start

La media es un 1,83, por lo que se concluye que **la herramienta presenta mayor utilidad cuando se utilizan dos enlaces de profundidad**. Es interesante comprobar que existen preguntas, como la manzana o las matemáticas, que simplemente requieren un nivel. Se entiende que los conceptos más allá de las relaciones directas simplemente no aportan nueva información o quizás resulte redundante.

Resultados Test 3

Con este último test se intenta probar y demostrar la aplicación del proyecto usándose directamente con las técnicas de IA descrita en el apartado de preparación. Se ha preguntado al agente inteligente, ¿el texto <texto del tipo I, II, III o IV> trata sobre el concepto <elemento wikidata>?. Los textos usados están recogidos en el Anexo. Se han buscado de manera aleatoria por internet y siempre corresponden a uno de los 4 tipos de textos.

Pregunta	I. Texto sobre el concepto	II. Texto que no es sobre el concepto	III. Texto sobre el concepto sin incluir el término	IV. Texto que no es sobre el concepto pero incluye el término
Barack Obama Q76	Incorrecto	Correcto	Correcto	Correcto
Paris Q90	Correcto	Correcto	Correcto	Correcto
Apple (empresa) Q312	Correcto	Correcto	Incorrecto	Correcto
Aluminium Q663	Correcto	Correcto	Correcto	Correcto

Por brevedad, y por lo tedioso que resulta la búsqueda de textos de manera manual que se ajusten a los tipos de textos descritos, se ha optado por hacer los test en tan solo cuatro de los conceptos usados en el resto de plantillas. Tal y como se observa en la tabla superior, el algoritmo LDA en combinación con los nodos escogidos en el proyecto, sirve para identificar la pertenencia a una temática dada de manera exitosa. Se comprueba además que la parte más compleja reside en la afirmación de pertenencia. Dicho de otra manera, es más fácil deducir o inferir que el concepto no pertenece, que afirmar que sí se trata del tema del texto. En estos casos si que se ha deducido erróneamente. Se procede a analizar ambos.

```
[ [ { term: 'obama', probability: 0.065 },
    { term: 'president', probability: 0.043 },
    { term: 'barack', probability: 0.043 },
    { term: 'american', probability: 0.043 },
    { term: 'african', probability: 0.043 },
    { term: 'winning', probability: 0.022 },
    { term: 'united', probability: 0.022 },
    { term: 'strengthen', probability: 0.022 },
    { term: 'states', probability: 0.022 },
    { term: 'senate', probability: 0.022 } ] ]
```

```
[ [ { term: 'steve', probability: 0.059 },
    { term: 'jobs', probability: 0.059 },
    { term: 'iphone', probability: 0.047 },
    { term: 'years', probability: 0.035 },
    { term: 'products', probability: 0.035 },
    { term: 'moment', probability: 0.035 },
    { term: 'keynote', probability: 0.035 },
    { term: 'today', probability: 0.023 },
    { term: 'revolutionize', probability: 0.023 },
    { term: 'phone', probability: 0.023 } ] ]
```

En el caso de Obama, el fallo reside simplemente en un formato erróneo en la trata de los datos. El algoritmo LDA destaca palabras únicas como temas principales del texto, sin embargo, en los nodos de este trabajo, los conceptos a menudo van unidos y pueden contener más de una palabra. Se entiende entonces que según el trabajo, el concepto de Barack Obama va siempre junto, ya que perdería significado o daría lugar a ambiguaciones si fuese de otra manera. En el segundo caso se muestran las palabras más relevantes del texto de manera correcta. Quizás en el tercer caso el texto no haya sido escogido de la forma más apropiada, ya que realmente habla exclusivamente sobre el lanzamiento del iPhone y nunca de la empresa Apple en general. Se podría aumentar el rango de conceptos usados en el trabajo, o aumentar el rango de búsqueda de temas en el LDA, para intentar así encontrar un punto en común.

Capítulo 5:
Conclusiones y Trabajos Futuros

En este trabajo se ha presentado una **propuesta estratégica de etapas y sintonización de parámetros para la creación de un sistema interactivo dirigido al aprendizaje en usuarios y al uso de la Web Semántica en aplicaciones de la Inteligencia Artificial**. Los diferentes estudios llevados a cabo han confirmado la importancia de la creación de una red de conceptos para los dos tipos de enfoques de este trabajo, en términos de aprendizaje para los seres humanos, y como fuente de datos en aplicaciones inteligentes. En lo que se refiere al aprendizaje **se propone además un proceso de estrategias basado en tres fases**, desde la visualización y filtrado, a la interacción y el cierre. Estas fases, con sus propios objetivos, están diseñadas para apoyar y reforzar el desarrollo cognitivo de los usuarios y facilitan el acceso de los grandes repositorios 'escondidos' de datos a aquellos que no conozcan los lenguajes semánticos. Como trabajo futuro e implementación del sistema en casos reales, se entiende que este trabajo pudiese resultar de utilidad como herramienta dentro de un **Entorno Personal de Aprendizaje** o ser usado en **sistemas de autorías propios**, donde el usuario pretenda crear contenido y necesite de una guía o de recomendaciones de conceptos que deben ser incluidos en su tema. La aplicación cobra mayor importancia en el contexto de la UNED, pues podría ser usada en entornos virtuales tanto enfocado a alumnos como a profesores. Por otro lado, en el enfoque hacia la Inteligencia Artificial, el paquete de datos o dossier resultante del sistema resulta de utilidad para múltiples usos en **aplicaciones inteligentes donde la contextualización sea importante**, como son los sistemas de recomendación en frío, descubrimiento de información en textos, inferencia o pertenencia a temáticas, clusterización… El sistema podría ser usado como herramienta a estudiantes de la IA o de Ingeniería Informática en general, que requieran de **bolsas de palabras o conjunto de conceptos** para sus propias aplicaciones inteligentes o trabajos de investigación.

Se ha evaluado y confirmado la fiabilidad y potencial de las grandes bases de conocimiento disponibles y abiertas en la Web Semántica, centrándose en las bases de Wikidata y Wikipedia, siendo éstas lo suficientemente maduras y robustas como para usarse de base fundamental única en sistemas de Aprendizaje o Inteligentes. Además de presentar un breve sumario comparativo de las bases de conocimiento multidominio más importantes y de las tecnologías de la Web Semántica (Datos Enlazados, RDF, SPARQL…), que pudiese servir de **referencia a futuros trabajos** basados en esta temática. También como referencia se podrán usar los trabajos recopilatorios sobre las distintas formas de representación del conocimiento, tanto a nivel de programación como de visualizado.

Cabe destacar de éste trabajo el proceso de sintonización de parámetros, que apoyado por los experimentos, ha confirmado la necesidad de garantizar el balance entre la

cantidad de datos (el potencial enorme de la Web Semántica) y la eficacia (asegurando la relevancia de los datos incluso después del filtrado). El filtrado de datos resulta imprescindible al crecer de manera exponencial el número de conceptos y relaciones según se avanza en niveles de profundidad, los cuales van perdiendo relevancia a medida que se alejan del concepto principal semilla. Los experimentos llevados a cabo han situado el umbral de relevancia en los dos niveles, tanto generalistas hacia arriba, como específicos hacia abajo. Cabe apuntar que se haría necesario un testear con más usuarios, ya que quizás la muestra no ha sido lo suficientemente grande como para que las conclusiones resulten representativas, aunque se esté de acuerdo con los datos resultantes y se intuya la información correcta. Para el filtrado además se ha llevado un análisis exhaustivo de las propiedades de Wikidata, concluyendo con una **propuesta de tres factores que se deben tener en cuenta en la elección de propiedades**, ya sea por su frecuencia en el dataset, el tipo de objeto al que hagan referencia, o la categoría a la que pertenezca el sujeto de la declaración. Esta técnica, aunque llevada a cabo de manera específica para Wikidata, puede ser extrapolada y usada en cualquier otro conjunto de datos que siga una estructura en forma de tripletes.

Como trabajos futuros y para seguir mejorando el trabajo, se propone una mayor integración entre los datos estructurados y no estructurados mediante técnicas de descubrimiento de información en textos y aumentar el uso de la Inteligencia Artificial, sobre todo en la elección de las propiedades y atributos en la ontología (a la hora de la categorización del sujeto) y en la inferencia de relaciones entre conceptos. Se reitera además que los experimentos deberían hacerse con más usuarios para asegurar la representatividad de los resultados.

Bibliografía

@AngryLoki. (n.d.). Retrieved from Wikidata Graph Builder: https://angryloki.github.io/wikidata-graph-builder/

@Cotrino. (2016). *Knowledge Map*. Retrieved from Cotrino: http://www.cotrino.com/2016/03/knowledgemap/

A. Gazzola, N. V. (2017). *Learn Anything*. Retrieved from https://learn-anything.xyz/

Akman, V. (2002). *CONTEXT IN ARTIFICIAL INTELLIGENCE: A FLEETING OVERVIEW.* Ankara, Turkey: Department of Computer Engineering, Bilkent University.

API:Categorymembers. (n.d.). Retrieved from www.mediawiki.org: https://www.mediawiki.org/wiki/API:Categorymembers

Basic membership properties. (2018, 4 3). Retrieved from www.wikidata.org: https://www.wikidata.org/wiki/Help:Basic_membership_properties

Cami, D. (2015). Retrieved from Wikidata Timeline: https://tools.wmflabs.org/wikidata-timeline/

Diefenbach, D., Tanon, T. P., Singh, K., & Maret, P. (2017). *Question Answering Benchmarks for Wikidata.* Lyon, France: Universite de Lyon, CNRS UMR 5516 Laboratoire Hubert Curien.

Dmitry Mouromtsev 1, G. W., Haase, P., Pavlov, D., Emelyanov, Y., & Morozov, A. (2018). *A Diagrammatic Approach for Visual Question Answering Over Knowledge Graphs.* t. Petersburg, Russia: Intern. Lab. of Information Science and Semantic Technologies, ITMO University.

Dudas, M., Zamazal, O., & Svate, V. (2014). Roadmapping and navigating in the ontology visualization landscape. *International Conference on Knowledge Engineering and Knowledge Management ,* (pp. 137 – 152). Springer.

Färber, M., Ell, B., Menne, C., & Rettinger, A. (2015). *A Comparative Survey of DBpedia, Freebase, OpenCyc, Wikidata, and YAGO.* Karlsruhe, Germany: Karlsruhe Institute of Technology (KIT), Institute AIFB.

Galería de Ejemplos con D3.js. (n.d.). Retrieved from https://d3js.org/: https://github.com/d3/d3/wiki/Gallery

Ghislain Auguste Atemezing, R. T. (2014). Towards a Linked-Data based Visualization Wizard. *5th International Workshop on Consuming Linked Data (COLD 2014).* Riva del Garda, Italy: EURECOM, Campus SophiaTech, France.

Graves, A., & Hendle, J. (2013). Visualization tools for open government data. *14th Annual International Conference on Digital Government Research* (pp. 136 - 145). ACM.

Gruber, T. (n.d.). Retrieved from What is an ontology?: https://web.njit.edu/~geller/what_is_an_ontology.html

Gruber, T. R. (1993). A translation approach to portable ontology specifications. *Knowledge acquisition, 5*(2), 199--220.

Jakub Klímek, J. H. (2014). Application of the Linked Data Visualization Model on Real World Data from the Czech LOD Cloud. *Linked Data on the Web (LDOW2014).* Seoul, Korea: Czech Technical University in Prague Faculty of Information Technology.

Janev, V., & Vranes, S. (2011). Applicability assessment of semantic web technologies. *Information Processing & Management*, pp. 507–517.

Jorn Hees, T. R.-B. (2010). Linked Data Games: Simulating Human Association with Linked Data. *LWA 2010 - Workshop-Woche: Lernen, Wissen & Adaptivität.*

Knowledge Management Department, German Research Center for Artificial Intelligence DFKI GmbH and Knowledge-Based Systems Group, University of Kaiserslautern.

Josep Maria Brunetti, S. A. (2012). The Linked Data Visualization Model. *International Semantic Web Conference*. Boston, USA: GRIHO, Universitat de Lleida Jaume II, 69. 25001 Lleida, Spain y AKSW, Computer Science University of Leipzig, Germany.

K. Ahmed, J. K. (2017). *Web Developer Roadmap and Mobile Developer Roadmap* . Retrieved from https://github.com/kamranahmedse/developer-roadmap and https://github.com/godrm/mobile-developer-roadmap

MediaWiki API. (n.d.). Retrieved from www.mediawiki.org: https://www.mediawiki.org/wiki/API:Main_page

Minsky, M. (1974). *A Framework for Representing Knowledge.* MIT-AI Laboratory Memo 306.

Moreira, M. A. (n.d.). *LA TEORÍA DE LOS CAMPOS CONCEPTUALES DE VERGNAUD, LA ENSEÑANZA DE LAS CIENCIAS Y LA INVESTIGACIÓN EN EL ÁREA.* UFRGS, Instituto de Física. Porto Alegre: Caixa Postal.

Newell, A., & Simon, H. (1956). The logic theory machine: A complex information processing system. In *IRE Transactions on Information Theory*.

Pavlov, D., Emelyanov, Y., Mouromtsev, D., Morozov, A., Razdyakonov, D., & Belyaeva, O. (2016). Ontodia.org - a simple cognitive service to fill the gap in linked open data management tools. *ISWC 2016.* Kobe, Japan: ITMO University,.

R. Davis, H. S. (1993). What is a Knowledge Representation? *AI Magazine*, 17-33.

Ramakrishnan, S., & Vijayan, A. (n.d.). study on development of cognitive support features in recent ontology visualization tools. *Artificial Intelligence Review.* 2014.

Ringler, D., & Paulheim, H. (2017). *One Knowledge Graph to Rule them All?. Analyzing the Differences between DBpedia, YAGO, Wikidata & co.* Mannheim, Germany: University of Mannheim, Data and Web Science Group.

Saphiro, C. (1992). *Sematinc Networks. Encyclopedia of Artificial Intelligence* . New York: John Wiley & Sons.

Schank, & Abelson. (1977). *Scripts, Plans, Goals, and Understanding.* Hillsdale , NJ: Earlbaum Assoc.

Semantic Web. (n.d.). Retrieved from www.w3.org: https://www.w3.org/standards/semanticweb/

Semantic Web Stack. (n.d.). Retrieved from http://dbpedia.org: http://dbpedia.org/describe/?uri=http%3A%2F%2Fdbpedia.org%2Fresource%2FSemantic_Web_Stack

Singhal, A. (2012, May). *Google Official blog*. Retrieved from https://googleblog.blogspot.com/2012/05/introducing-knowledge-graph-things-not.html

Stillings. (1987). Philosophy: The Foundations of Cognitive Science. In *Cognitive Science: An Introduction.* Cambridge: MIT Press.

team, W. S. (n.d.). *Wikimedia REST API*. Retrieved from en.wikipedia.org: https://en.wikipedia.org/api/rest_v1/

Vergnaud, G. (1990). La teoría de los campos conceptuales. *Recherches en Didáctique des Mathématiques, 10*(2, 3), 133-170.

Walliman, D. (2013). Retrieved from Domain of Science: https://www.flickr.com/people/95869671@N08/ y https://www.youtube.com/user/dominicwalliman/featured

Wikidata: Tools to Visualize Data. (n.d.). Retrieved from www.wikidata.org: https://www.wikidata.org/wiki/Wikidata:Tools/Visualize_data
Wikidata:Database reports/List of properties/Top100. (2018, 7 11). Retrieved from www.wikidata.org: https://www.wikidata.org/wiki/Wikidata:Database_reports/List_of_properties/Top100
Wikipedia. (n.d.). Retrieved from https://www.wikipedia.org/

Anexos

Ejemplos de esfuerzos en la creación de redes de conceptos

La visualización es un componente vital del proceso analítico y de aprendizaje, pero la información presentada sin contexto pierde parte de su idea. Los datos son más que un solo conjunto de números. Las ideas se pueden (y deben) tomar de varias partes de un conjunto o combinar diversas fuentes en una comprensión procesable. Es por ello que existen numerosos proyectos y ejemplos de intentos de agrupar, contextualizar y relacionar concocimiento:

Gráficos, mapas y roadmaps manuales: todos los ejemplos tienen algo en común, intentan aportar perspectiva, claridad y relación entre todos los conceptos de un campo de estudio. Sin embargo, estos mapas no muestran relaciones entre distintos campos del conocimiento humano, y están hechos a mano, de manera no autónoma (algunos incluso de manera colaborativa con el trabajo de cientos de personas, como es el caso de Learn Anything). El gran entusiasmo e interés que estos mapas despiertan en las personas demuestran la necesidad de este proyecto.

Learn Anything: tutoriales y fuentes de información para aprender cualquier cosa, mostradas en forma de gráfico de conceptos relacionados (A. Gazzola, 2017). Sitio web de código abierto creado por la comunidad para aprender cualquier cosa con mapas interactivos. El objetivo del sitio web es acelerar la velocidad a la que las personas aprenden y obtienen nuevos conocimientos al ofrecer los caminos más eficientes que se pueden tomar para obtener una comprensión completa de cualquier tema. Para hacer esto, se ha creando el primer sitio web comisariado y de código abierto de su clase, centrado en el aprendizaje de una manera lineal y progresiva. Busca en una base de datos abierta de mapas conectados e interactivos.

La guía gráfica del desarrollador web y del desarrollador móvil: un conjunto de gráficos que demuestran las rutas que se puede tomar y las tecnologías que desea adoptar para convertirse en desarrolladores frontend, back-end o devops, en web o móvil. El objetivo inicial de estas tablas se hicieron para un profesor que quería compartir algo con sus estudiantes universitarios que pudiera darles una perspectiva en sus carreras (K. Ahmed, 2017).

Los mapas de todo: Dominic Walliman ama la ciencia y realmente le gusta encontrar formas de explicarla a los demás. Sus mapas sobre todo y sus videos en youtube acumulan cientos de miles de visitas (Walliman, 2013).

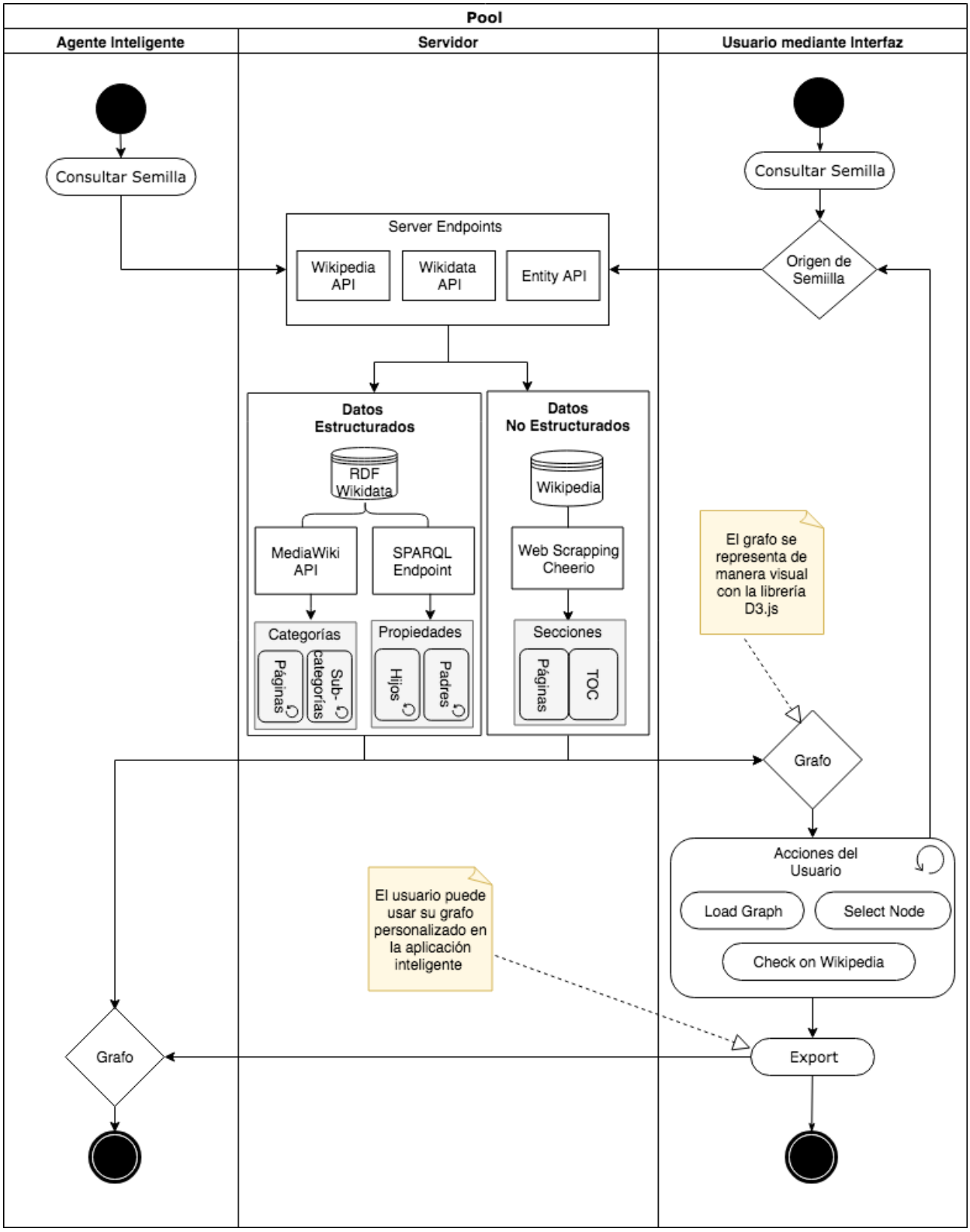
Enfoque en la visualización: el interés por crear una mejor visualización y representación del conocimiento y crear relaciones entre conceptos, materias y campos de estudio no es nuevo. La importancia de apoyar la visualización para

proyectos basados en grandes volúmenes de datos se puede demostrar por la gran cantidad de trabajos realizados que giran en torno a este tema. La propia existencia de herramientas y soporte desde la propia Wikidata (Wikidata: Tools to Visualize Data, n.d.) y de librerías como D3.js[11] hacen prueba de ello (Galería de Ejemplos con D3.js, n.d.). Con estos ejemplos de referencia se puede comprobar no solo la necesidad, pero también la problemática que surge al tratar con grandes cantidades de datos, es por eso que la mayoría de ejemplos no son soluciones generalistas, sino que se enfocan en un problema pequeño y específico (por ejemplo el proyecto Wikidata periodic table, que se centra en la navegación y visualización de los elementos químicos[12], o Wikidata Graph Builder: Represantación y visualización de datos de WikiData con d3.js (@AngryLoki, n.d.), WikiData Timeline: Web app for visualizing Wikidata items in the form of a timeline (Cami, 2015) y Knowledge Map, un proyecto implementado para tratar de responder las siguientes preguntas: ¿Qué sé? ¿Qué es lo que no sé? ¿Sé más sobre un tema que cualquier otra persona en particular? ¿Qué libros o fuentes pueden ampliar mi conocimiento? (@Cotrino, 2016)).

[11] D3.js (o solo D3 para documentos basados en datos) es una biblioteca de JavaScript para producir visualizaciones dinámicas e interactivas de datos en navegadores web. Hace uso de los estándares SVG, HTML5 y CSS ampliamente implementados.

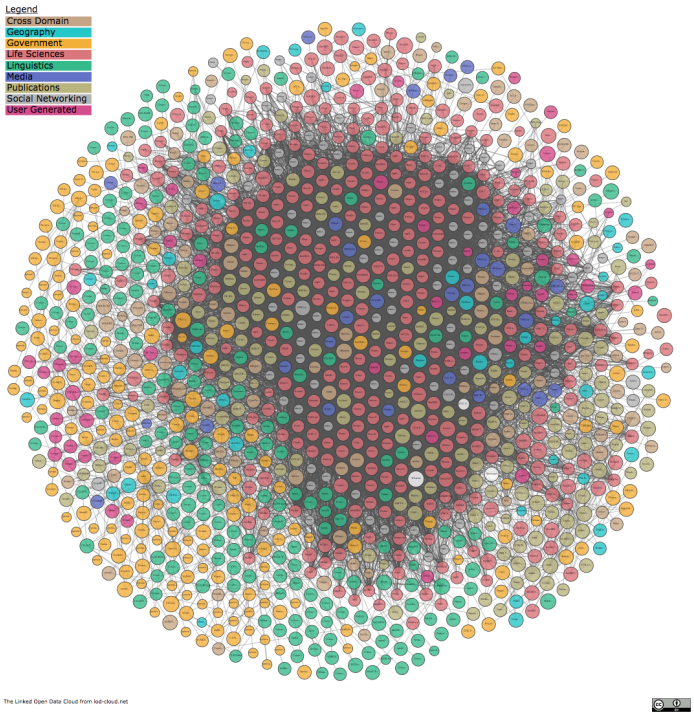
[12] https://tools.wmflabs.org/ptable/

Arquitectura de la Aplicación



Bases de Conocimiento estructuradas

El término Knowledge Graph es usado actualmente para indicar bases de conocimiento semánticas y estructuradas. Los Gráficos de Conocimiento públicos en la Web se consideran un activo valioso para el desarrollo de aplicaciones inteligentes. Contienen conocimientos generales que se pueden usar, por ejemplo, para mejorar las herramientas de análisis de datos, el procesamiento de texto o los sistemas de recomendación. El término "Gráfico de conocimiento" fue acuñado por Google cuando introdujeron su gráfico de conocimiento como columna vertebral de una nueva estrategia de búsqueda web en 2012, es decir, pasar del procesamiento de texto puro a una representación más simbólica del conocimiento (Singhal, 2012). Existen en la actualidad ingentes cantidades de bases de conocimiento con datos enlazados. Para poder hacerse una idea de la magnitud que estas bases alcanzan se recomienda visitar la página web lod-cloud[13], donde se representan en un solo diagrama los **1,224** datasets con sus **16,113** enlaces que conforman el conjunto de bases de conocimiento con datos enlazados disponibles y abiertas en la nube.

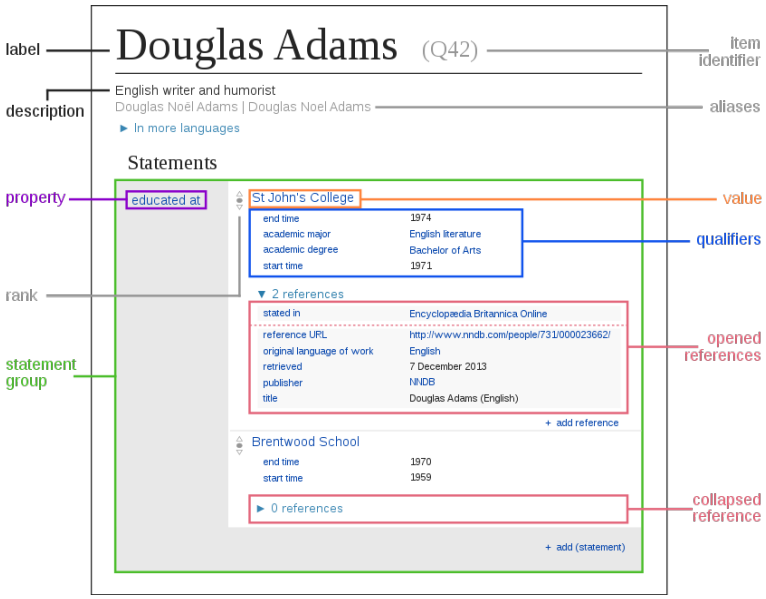


La Web Semántica no se trata solo de poner datos en la web en bases de conocimiento, se trata de hacer enlaces para que una persona o máquina pueda explorar la red de datos. Con los datos vinculados se pueden encontrar otros datos relacionados. Las bases de conocimiento pueden abarcar temáticas muy distintas, pero quizás las mas relevantes para este proyecto sean aquella de dominio cruzado o multidominio, es decir, bases generalizadas que pretenden recopilar y enlazar todo el conocimiento humano. A continuación se van a describir las bases de conocimiento multidominio más importantes y relevantes.

[13] https://lod-cloud.net

Wikidata:

Como se ha visto anteriormente, Wikipedia es la mayor base de conocimiento que existe, sin embargo todo ese conocimiento no está estructurado y los artículos e identificadores pueden cambiar constantemente. Es por eso que se creó Wikidata. Wikidata es una base de datos secundaria gratuita, colaborativa, multilingüe, que recopila datos **estructurados** para brindar soporte a Wikipedia, a Wikimedia Commons, a las otras wikis del movimiento Wikimedia y a cualquier persona en el mundo. La imposición de un alto grado de organización estructurada permite una fácil reutilización de datos por parte de los proyectos de Wikimedia y de terceros, y permite que las computadoras los procesen y "entiendan". Wikidata ayuda también a Wikipedia con cajas de información más fáciles de mantener y enlaces a otros idiomas, lo que reduce la carga de trabajo de edición y mejora la calidad.



El repositorio Wikidata consta principalmente de *items*, cada uno con una etiqueta (*label*), una descripción y cualquier cantidad de alias.

Los elementos se identifican de forma única con una Q seguida de un número, como por ejemplo Douglas Adams (Q42).

Las declaraciones describen características detalladas de un artículo y consisten en una propiedad y un valor. Las propiedades en Wikidata tienen una P seguida de un número, como con educado en (P69).

La mejor forma de hacer uso de datos de Wikidata es mediante SPARQL. SPARQL (acrónimo de Simple Protocol And RDF Query Language) es un lenguaje de consulta RDF, es decir, un lenguaje de consulta semántica para bases de datos. Con SPARQL se puede extraer cualquier clase de datos, con una consulta compuesta de combinaciones lógicas de tripletes (*triplets*), elemento – propiedad – valor, (*item – property - value*). En Wikidata hay un total de 5,800,000,000 tripletes[14].

DBpedia

DBpedia (de "DB" para "base de datos") es un proyecto que busca **extraer contenido estructurado** de la información creada en el proyecto de Wikipedia, y fue creado antes de que surgiese Wikidata. Esta información estructurada está disponible en la World Wide Web. DBpedia permite a los usuarios consultar semánticamente las relaciones y

[14] https://lod-cloud.net/dataset/wikidata

propiedades de los recursos de Wikipedia, incluidos los enlaces a otros conjuntos de datos relacionados.
El conjunto de datos de DBpedia describe 4.58 millones de entidades, de las cuales 4.22 millones se clasifican en una ontología consistente. Uno de los desafíos al extraer información de Wikipedia es que los mismos conceptos se pueden expresar usando diferentes parámetros en infobox y otras plantillas. DBpedia contiene un total de 9,500,000,000 tripletes.

Freebase

Freebase era una gran base de conocimiento colaborativo que consistía en datos compuestos principalmente por miembros de su comunidad. Era una colección en línea de datos estructurados recopilados de muchas fuentes (mayoritariamente Wikipedia), incluidas contribuciones individuales y presentadas por los usuarios. [3] Freebase tenía como objetivo crear un recurso global que permitiera a las personas (y máquinas) acceder a información común de manera más efectiva. Fue desarrollado por la empresa estadounidense de software Metaweb y empezó públicamente en marzo de 2007. Metaweb fue adquirido por Google en 2010, con el propósito de impulsar con Freebase el proyecto Knowledge Graph de Google. En 2014, Knowledge Graph anunció que cerraría Freebase durante los siguientes seis meses y ayudaría a mover los datos de Freebase a Wikidata. Freebase contiene un total de 337,203,427 tripletes.

OpenCyc

Cyc es el proyecto de Inteligencia Artificial más longevo del mundo, que intenta integrar una ontología y una base de conocimiento que abarquen los conceptos básicos y las "reglas de oro" sobre cómo funciona el mundo, con el objetivo de permitir que las aplicaciones de Inteligencia Artificial hagan un razonamiento humano y sean menos "frágiles" cuando se enfrentan con situaciones nuevas para las que no fueron preconcebidas. OpenCyc contiene alrededor de 1,600,000 tripletes.

YAGO

YAGO (Yet Another Great Ontology) es una base de conocimiento de código abierto desarrollada en el Instituto Max Planck de Ciencias de la Computación en Saarbrücken. Se extrae automáticamente de Wikipedia, WordNet y GeoNames.
Cabe destacar de YAGO que cada relación está anotada con su valor de confianza, concede una dimensión temporal y una dimensión espacial a muchos de sus hechos y entidades y además extrae y combina entidades y hechos de 10 idiomas diferentes en la Wikipedia. Contiene alrededor de 120,000,000 tripletes.

Knowledge Graph de Google

El Knowledge Graph es una base de conocimiento utilizada por Google y sus servicios para mejorar los resultados de su motor de búsqueda con información recopilada de una variedad de fuentes, entre las que se incluye Wikidata y Wikipedia. El Knowledge Graph fue en parte potenciado por Freebase. En 2014, se lanzó también un proyecto de investigación llamado Knowledge Vault, una nueva iniciativa para suplantar las capacidades del Knowledge Graph, diferenciándose en que se debía tratar con hechos seguros (un término para información que se considera que tiene más del 90% de posibilidades de ser cierto), reuniendo y fusionando automáticamente información de Internet en una base de conocimiento capaz de responder preguntas directas.

Unas comparativas más extensas y cuantitativas de las bases de conocimiento estructuradas más importantes las podemos encontrar en (Ringler & Paulheim, 2017) y en (Färber, Ell, Menne, & Rettinger, 2015).

Representación del conocimiento a nivel de programación en la Inteligencia Artificial.

Listas: las listas enlazadas se utilizan para representar el conocimiento jerárquico. Desarrolladas por primera vez por Cliff Shaw y Herbert Simon como la principal estructura de datos para su Lenguaje de Procesamiento de la Información (IPL). Dicho lenguaje fue de los primeros en usarse para problemas de Inteligencia Artificial.

Árboles: estructuras de datos que representan el conocimiento jerárquico. LISP, el lenguaje de programación principal de AI, fue desarrollado para procesar listas y árboles. Los árboles, y más concretamente los árboles de decisiones, son comúnmente utilizados en la búsqueda de operaciones, específicamente en el análisis de decisiones, para ayudar a identificar una estrategia que pueda alcanzar un objetivo, pero también son una herramienta popular en el aprendizaje automático y por lo tanto en el área de IA. Los árboles tienen la ventaja de que son muy fáciles de interpretar y entender.

Redes semánticas - nodos y enlaces - almacenados como proposiciones. Ejemplos en (Stillings, 1987). Las Redes Semánticas son estructuras que representan el conocimiento en forma de patrones de nodos y arcos interconectados, éstos arcos representan relaciones semánticas entre conceptos en dicha red o estructura gráfica. Las implementaciones informáticas de redes semánticas se desarrollaron primero para la Inteligencia Artificial y la traducción automática, pero las versiones anteriores se han utilizado durante mucho tiempo en filosofía, psicología y lingüística (Saphiro, 1992).

Esquemas: utilizados para representar conocimiento de sentido común o estereotipado.

- **Marcos, o en inglés, Frames** (Minsky, 1974)- Describen objetos. Consiste en un grupo de nodos y enlaces manipulados como un todo. El conocimiento está organizado "slots". Los marcos están organizados jerárquicamente, derivándose originalmente de redes semánticas y, por lo tanto, siendo parte de representaciones de conocimiento basadas en la estructura. La suposición subyacente de la teoría de marcos es que sus terminales ya están llenos de valores predeterminados o por defecto. Es decir, que cuando uno encuentra una nueva situación (o se produce un cambio sustancial en la visión de la situación), uno selecciona de su memoria el marco que representa un concepto dado y lo cambia para reflejar la nueva realidad.

- **Scripts** (Schank & Abelson, 1977) describen eventos en lugar de objetos. Consisten en cadenas de eventos causales o temporales ordenados estereotípicamente. Los scripts se parecen mucho a los marcos o frames, excepto que los valores que llenan los espacios o slots se deben ordenar. Un script es una representación estructurada que describe una secuencia estereotipada de eventos en un contexto particular. Los scripts se usan en los

sistemas de comprensión del lenguaje natural para organizar una base de conocimiento en términos de las situaciones que el sistema debería comprender. Ejemplos en (Stillings, 1987).

Representaciones basadas en reglas (Newell & Simon, 1956) se usan en contextos específicos de resolución de problemas. Implica reglas de producción que contienen pares si-entonces o acción-situación. Contienen:

- Estado inicial
- Estado objetivo
- Operadores legales, es decir, cosas que se le permite hacer
- Restricciones del operador, es decir, factores que limitan la aplicación de operadores

Representaciones basadas en lógica: pueden usar razonamiento deductivo o inductivo. Contiene:

- Hechos y premisas
- Reglas de lógica proposicional (booleana: trata de declaraciones completas)
- Reglas de cálculo de predicados (permite el uso de información adicional sobre objetos en la proposición, el uso de variables y funciones de variables).

Medidas de certeza: pueden implicar factores de certeza (por ejemplo, si el diagnóstico es un síntoma (CF)) que podría derivarse de la estimación de un experto o de datos estadísticos; Probabilidad bayesiana; o lógica difusa (en la que los conceptos o la información tienen algún valor de certeza asociado).

Propiedades de Wikidata

Según la propia definición que aporta Wikidata en su documentación, las propiedades de Wikidata describen el valor de datos de una declaración y pueden considerarse como una categoría de datos (por ejemplo, "color" para el valor de datos "azul"). Las propiedades, cuando se combinan con valores (o con calificadores), forman una declaración en Wikidata. Las propiedades tienen sus propias páginas en Wikidata y están conectadas a los elementos, lo que da como resultado una estructura de datos vinculada y enlazada.

Las propiedades son similares a los elementos, *'items'* de varias maneras, cada propiedad tiene una etiqueta, una descripción e incluso uno o varios alias que se pueden agregar en varios idiomas. También constan de instrucciones que ayudan a describir más completamente la propiedad, incluidas las restricciones sobre cómo se puede usar. Sin embargo, las propiedades no tienen una sección en sus páginas para enlaces a otros proyectos de wikimedia, y tampoco tienen identificadores externos.
Cada propiedad tiene lo que se conoce como *Ayuda:* Tipo de datos que define el tipo de valores permitidos en las declaraciones con esa propiedad.
Al igual que los elementos, las propiedades también tienen un identificador único que se diferencia en que mientras que los elementos comienzan con un prefijo Q y se encuentran en el espacio de nombre principal de Wikidata, las propiedades tienen un prefijo P y se almacenan en el espacio de nombres de la propiedad. Para ver el conjunto de propiedades de los que cuenta Wikidata se recomienda visitar el informe actualizado sobre las listas de propiedades del dataset completo[15].

Las propiedades escogidas para este proyecto han sido cuidadosamente escogidas por aportar y reflejar mayor sentido contextual de acuerdo a ciertos factores:

- **Frecuencia de uso:** es sencillo encontrar en Wikidata, mediante consultas a SPARQL o mediante los datos estadísticos que se muestran en tiempo real gracias a los continuos dumps del dataset, las propiedades más frecuentes y de más uso en Wikidata. Por ejemplo, instancia de (P31), es la propiedad más usada y aparece en un total de 46,421,407 artículos; o la propiedad subclase de (P279), en un total de 1,444,443. Se han escogido pues aquellas propiedades que tienen mayor posibilidad de aparecer en la definición de un concepto. Entre las propiedades más usadas se encuentran también las propiedades de membresía básica[16], que son aquellas que muestran pertenencia o clase (instancia de (P31), subclase de (P279) y

[15] https://www.wikidata.org/wiki/Wikidata:Database_reports/List_of_properties/all

[16] https://www.wikidata.org/wiki/Help:Basic_membership_properties

forma parte de (P361)) y que por lo tanto cuentan con especial relevancia en el contexto de este trabajo. Estas propiedades son las más indicadas para mostrar conceptos de pertenencia y jerarquía, ayudando a una visualización contextual y justificando su inclusión en el filtrado. Si prácticamente todos los conceptos de Wikidata cuentan con una declaración con la propiedad P31, querrá decir que su definición resulta imprescindible para entender el concepto.

- **El tipo de datos al que hagan referencia:**[17] Los tipos de datos definen cómo se comportará una declaración y qué tipo de datos tomará. Los diferentes tipos de declaraciones utilizan diferentes tipos de propiedades y también usan diferentes tipos de datos (durante la creación de las propiedades, una de las opciones es establecer el tipo de datos, y una vez establecida, la propiedad se bloquea con este tipo de datos). Esto tiene implicaciones sobre cómo se pueden usar y debe haber una cierta cantidad de planificación antes de poder definir la utilidad de una propiedad. Se han evitado las propiedades cuyo valor suele ser cuantitativo (date of birth (P569), height (P2048), population (P1082)…) y que no aportan en la creación de una red de conceptos. Se ha centrado el trabajo en escoger propiedades que hagan referencia a otros elementos de Wikidata (tipo de datos ***wikibase-item***), ya que son las relaciones entre conceptos las que aportan contextualización. Los tipos de datos cuantitativos que se han evitado en el filtrado son por ejemplo *quantity, string, time, url…* También se han evitado aquellos que hacen referencia al léxico o la semántica, como *wikibase-lexeme* o *monolingualtext*, o tipos de datos que hacen referencias a conjunto de datos externos, como *external-id*, o referencias a otras propiedades, *wikibase-property*.
- **Según el sujeto de la declaración:** Las propiedades en Wikidata están agrupadas según el sujeto al que hagan referencia, así por ejemplo se permitirán distintas propiedades si el sujeto es una instancia de la clase humano (propiedades como father (P22) o date of death (P570)), que si el sujeto es de la clase transporte (connecting line (P81) o airline hub (P113)). En este trabajo no se ha llevado a cabo una clasificación por éste método ya que no se ha inferido el tipo de sujeto, sin embargo se ha hecho un esfuerzo por mostrar la categorización del sujeto al usuario. Para ello se ha añadido la propiedad categoría en Commons (P373) que resulta muy interesante para esquematizar, y que revela datos importantes que de alguna manera se podrían haber perdido, por ejemplo, aunque se ha omitido la propiedad capital (P36), usado en sujetos de tipo geográfico, existe la categoría Capitales, que aparecerá si ese elemento es una capital importante.

Como proyecto futuro, esta elección de propiedades se podría llevar a cabo metódica y automáticamente, mediante algoritmos de Inteligencia Artificial, en este trabajo sin embargo, se considera que los objetivos son específicos al área de conceptualización y contextualización, por lo que merece la pena el esfuerzo de búsqueda y elección manual de estas propiedades.

[17] https://www.wikidata.org/wiki/Category:Properties_by_datatype

World Wide Web Consortium (W3C) y la Web Semántica

La Web Semántica es una extensión de la World Wide Web a través de estándares definidos en el World Wide Web Consortium, W3C, (Semantic Web, n.d.). Estos estándares promueven formatos de datos comunes y protocolos de intercambio en la Web. Según el W3C, "La web semántica proporciona un marco común que permite que los datos se compartan y reutilicen a través de los límites de la aplicación, la empresa y la comunidad". Por lo tanto, la Web Semántica se considera como un integrador de diferentes contenidos, aplicaciones de información y sistemas. Es importante destacar que no solo la Web Semántica necesita acceso a los datos, sino que las relaciones entre la información también deberían estar disponibles para crear una Red de Datos (en lugar de una colección pura de conjuntos de datos). Esta colección de conjuntos de datos interrelacionados en la Web también puede denominarse Datos Vinculados, Datos Enlazados o Linked Data en inglés. Los Datos Enlazados están potenciados por tecnologías como RDF, SPARQL, OWL y SKOS. La Pila de la Web Semántica ilustra la arquitectura de la Web Semántica. Las funciones y relaciones de los componentes se pueden resumir de la siguiente manera:

XML: proporciona una sintaxis elemental para la estructura de contenido dentro de los documentos, sin embargo, no asocia ninguna semántica con el significado del contenido que contiene. Actualmente, XML no es un componente necesario de las tecnologías de la Web Semántica

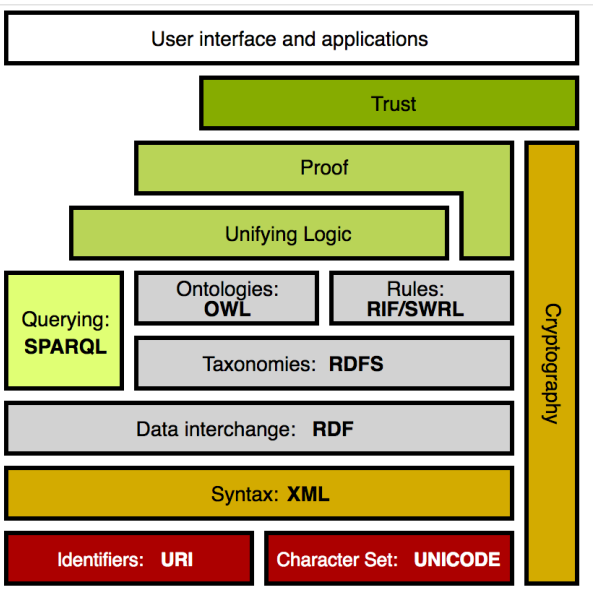


Ilustración 9: Pila de la Web Semántica (Semantic Web Stack, n.d.)

XML Schema: es un lenguaje para proporcionar y restringir la estructura y el contenido de los elementos contenidos en los documentos XML.

RDF: es un lenguaje simple para expresar modelos de datos. En el siguiente apartado se estudia con más detalle.

Esquema RDF: amplía RDF y es un vocabulario para describir propiedades y clases de recursos basados en RDF, con semántica para jerarquías generalizadas de dichas propiedades y clases.

OWL: agrega más vocabulario para describir propiedades y clases: entre otras, las relaciones entre clases (por ejemplo, disjuntos), cardinalidad (por ejemplo, "exactamente una"), igualdad, tipado de propiedades más rico, características de propiedades (por ejemplo, simetría) y clases enumeradas.

SPARQL: es un protocolo y lenguaje de consulta para fuentes de datos web semánticos. En el siguiente apartado se estudia con más detalle.
RIF: es el formato de intercambio de reglas del W3C. Es un lenguaje XML para expresar reglas web que las computadoras pueden ejecutar. RIF proporciona múltiples versiones, llamadas dialectos. Incluye un Dialecto Lógico Básico RIF (RIF-BLD) y un Dialecto de Reglas de Producción RIF (RIF PRD).

Tecnologías de la Web Semántica: RDF y SPARQL

El sistema de descripción de recursos conocido como Resource Description Framework (RDF) es una familia de especificaciones del World Wide Web Consortium, W3C, (Semantic Web, n.d.) originalmente diseñado como un modelo de datos de metadatos. Se ha utilizado como un método general para la descripción conceptual o el **modelado de información** que se implementa en recursos web, utilizando una variedad de notaciones de sintaxis y formatos de serialización de datos (por ejemplo para representar, entre otras cosas, información personal, redes sociales, metadatos sobre artefactos digitales, así como para proporcionar un medio de integración sobre fuentes de información dispares y en aplicaciones de gestión del conocimiento). Se basa en la idea de hacer afirmaciones sobre recursos web en expresiones que sigan la fórmula sujeto-predicado-objeto, conocido como triples. El sujeto denota el recurso, y el predicado denota rasgos o aspectos del recurso, y expresa una relación entre el sujeto y el objeto. Por ejemplo, una forma de representar la noción "El cielo es de color azul" en RDF es como el siguiente triplete: un sujeto que denota "el cielo", un predicado que denota "es de color" y un objeto que denota "azul". Por lo tanto, RDF usa sujeto en lugar de objeto (o entidad) en contraste con el enfoque típico de un modelo entidad-valor-atributo en diseño orientado a objetos: entidad (cielo), atributo (color) y valor (azul). Para decirlo de otra manera, RDF es una estructura de datos del tipo grafo dirigido y etiquetado para representar información en la Web. Esta especificación define la sintaxis y la semántica del lenguaje de consulta SPARQL para RDF.

SPARQL (acrónimo de Simple Protocol And RDF Query Language) es un lenguaje de consulta RDF, es decir, un lenguaje de consulta semántica para bases de datos, capaz de recuperar y manipular datos almacenados en formato RDF. Técnicamente, las consultas SPARQL se basan en patrones (triples) y proporcionan uno o más patrones contra tales relaciones. Estos patrones triples son similares a los triples de RDF, excepto que una o más de las referencias de recursos constituyentes son variables. Un motor SPARQL devolvería los recursos para todas los tripletes que coinciden con estos patrones.

Textos usados en el test 3

Barack Obama Q76

I. Texto sobre el concepto

Barack Obama, in full Barack Hussein Obama II, (born August 4, 1961, Honolulu, Hawaii, U.S.), 44th president of the United States (2009–17) and the first African American to hold the office. Before winning the presidency, Barack Obama represented Illinois in the U.S. Senate (2005–08). He was the third African American to be elected to that body since the end of Reconstruction (1877). In 2009 he was awarded the Nobel Peace Prize "for his extraordinary efforts to strengthen international diplomacy and cooperation between peoples."

II. Texto que no es sobre el concepto

Ken Jennings, the 74-time winner of the popular trivia quiz, and Brad Rutter, a 20-time champion, have gone head-to-hard-drive with an IBM supercomputer called Watson three times in the past three days. Unlike in The Terminator, they lost each time. The supercomputer, named after former International Business Machines corporation president Thomas Watson, is a showcase of the company's expertise in advanced science and computing. Watson showed off its encyclopedic knowledge of topics ranging from ancient languages to fashion design, along with a few glitches.

III. Texto sobre el concepto sin incluir el término

The Norwegian Nobel Committee has decided that the Nobel Peace Prize for 2009 is to be awarded to the 44th President of the United States of America for his extraordinary efforts to strengthen international diplomacy and cooperation between peoples. The Committee has attached special importance to his vision of and work for a world without nuclear weapons. As President , he has created a new climate in international politics. Multilateral diplomacy has regained a central position, with emphasis on the role that the United Nations and other international institutions can play. Dialogue and negotiations are preferred as instruments for resolving even the most difficult international conflicts. The vision of a world free from nuclear arms has powerfully stimulated disarmament and arms control negotiations. Thanks to his initiative, the USA is now playing a more constructive role in meeting the great climatic challenges the world is confronting. Democracy and human rights are to be strengthened.

IV. Texto que no es sobre el concepto pero incluye el término

Obama is a small city located about 35 km west-southwest of Tsuruga city.
It faces the Cove with the same name, which is surrounded by two small peninsulas like breakwater. The city had been one of the port town on the Sea of Japan side since

ancient times, and it had been the entrance from the continent to Japan since the 7th century.
Wakasa area has good fishing grounds of Wakasa Bay. So various marine foods from Wakasa Bay had been sent from the city to Kyoto, the ancient Japanese capital. The city had been the very important supply center of marine foods for Emperor in Kyoto. The carriers in the city had always run through the mountain road about 80 km long to Kyoto within a day. Especially fresh mackerels were salted quickly, and when the carriers reached Kyoto the next day, the mackerels were more tasty. So the route from the city to Kyoto has been called "Mackerel Road" ("Saba Kaidou" in Japanese). Of course, various cultures in Kyoto had come to the city through this road. In the result, about 130 Buddhist temples had been built in the 8-14th centuries in this small city. They has many national treasure and national important cultural properties. "Obama" means "small beach" in Japanese.

Paris Q90

I. Texto sobre el concepto

This was my first time to Paris, but certainly not my last. I had a great time and I wanted to share some information for future travelers to Paris. First off I walked everywhere in Paris so yes I never used a Taxi except for getting from the airport to my hotel and also taking a taxi to the train station. One thing you will find out about Paris is you can walk everywhere and see so much. The food is great, the vibe of the city and the culture I found real interesting and the amount to see was fantastic. I had five whole days in Paris so that is why I decided to walk everywhere. If you don't have that much time, then grab the Metro, since it is quicker then walking. Once you get to Paris buy a Museum pass. I can not explain to you how many times when I got to a museum the line was so long, but since I had this pass I went by all these people and right in. If you don't speak french don't worry about it. I learned the basic phrases and then after that spoke english and I had no problem communicating with anyone. Paris is a great city and I really enjoyed it so book a flight and get there and enjoy.

II. Texto que no es sobre el concepto

Ken Jennings, the 74-time winner of the popular trivia quiz, and Brad Rutter, a 20-time champion, have gone head-to-hard-drive with an IBM supercomputer called Watson three times in the past three days. Unlike in The Terminator, they lost each time. The supercomputer, named after former International Business Machines corporation president Thomas Watson, is a showcase of the company's expertise in advanced science and computing. Watson showed off its encyclopedic knowledge of topics ranging from ancient languages to fashion design, along with a few glitches.

III. Texto sobre el concepto sin incluir el término

The capital of France has many nicknames, but its most famous is "La Ville-Lumière" ("The City of Light"), a name it owes first to its reputation as a centre of education. In

addition to being the capital it is also the largest city of France. It is situated along the Seine River, in northern France, at the heart of the Île-de-France region. The city within its administrative limits (the 20 arrondissements), has a population of about 2, 230, 000. Its metropolitan area is one of the largest population centres in Europe, with more than 12 million in habitants. Considered as green and highly liveable, the city and its region are the world's leading tourism destination.

IV. Texto que no es sobre el concepto pero incluye el término

Hotel heiress and socialite P. Hilton rose to fame via the reality TV series 'The Simple Life,' and continues to court media attention through her books, businesses, music and screen appearances. After starring in *The Simple Life*, she earned a Teen Choice Award in her portrayal in the film *House of Wax* and her first book *Confessions of An Heiress* made its way onto the *New York Times* bestseller list. She has since made her way in and out of the headlines through her romances, music ventures and television shows like *The World According to Paris* and *Hollywood Love Story*. Given the childhood nickname "Star" by her mother and grandmother, Hilton, who began modeling as a child at charitable functions, was born into acting stock. Her maternal grandmother was actress "Big" Kathy Dugan, and aunts Kim and Kyle Richards continue to earn film and television roles.

Apple (empresa) Q312

I. Texto sobre el concepto

It's official: Apple has become the first $1 trillion company in history. The milestone is even more significant when you consider that Apple almost didn't even get the chance to make it this far. When Steve Jobs took over as CEO of Apple in 1997, the company had been struggling to find its legs in a market increasingly dominated by Microsoft and its partners. Indeed, Michael Dell himself once quipped that if he were in Jobs' shoes, he'd shut Apple down and return the money to the shareholders. Here's a look into the history of Apple in photos, from its inception, through its hard times, and through to the triumphant return of Jobs.

II. Texto que no es sobre el concepto

Deserts are areas where the rainfall is too low to sustain any vegetation at all, or only very scanty scrub. The rainfall in desert areas is less than 250 mm or 10 inches per year, and some years may experience no rainfall at all. The hot deserts are situated in the subtropical climate zone where there is unbroken sunshine for the whole year due to the stable descending air and high pressure. Such areas include the Sahara, Saudi Arabia, large parts of Iran and Iraq, northwest India, California, South Africa and much of Australia. Here, maximum temperatures of 40 to 45 degrees C are common, although during colder periods of the year, night-time temperatures can drop to freezing or below due to the exceptional radiation loss under the clear skies. The Gobi desert in Mongolia is an example of a cool desert. Though hot in summer, it shares the

very cold winters of central Asia. The Arctic and Antarctic regions, too, receive very little precipitation during the year, owing to the exceptionally cold dry air, but are more usually classified as types of polar climate. Semi-desert areas include the Steppes of southern Russia and central Asia, and the Parries of Canada.

III. Texto sobre el concepto sin incluir el término

11 years ago today, Steve Jobs introduced the iPhone. Eleven years ago today Steve Jobs announced a wide-screen iPod with touch controls, a revolutionary mobile phone, and a breakthrough internet device. But it wasn't three products. It was one product. And we got it, Steve. We got iPhone. on January 9, 2007, Steve Jobs put sneaker to stage for what was the most incredible keynote presentations of his life—a life filled with incredible keynote presentations—and in the history of consumer electronics. The company had been working for over two years on the Purple Experience Project. It had gone from a tablet to a phone. From a dream to reality. And just before he stepped out in front of the crowd, Jobs assembled his team and told them to remember the moment: The moment before iPhone. Because, in the next moment, everything would change. During the keynote, Steve Jobs said it was rare enough for a company to revolutionize even one product category, but they had already revolutionized two: Computers with the Mac and personal music players with the iPod. With the iPhone they'd be going for three.

IV. Texto que no es sobre el concepto pero incluye el término

Conquer The Big Apple: 30 Things To Do In New York City. Home to over 8 million people, New York is the most populous city in the United States. Having been depicted in numerous films like Breakfast and Tiffany's and Goodfellas, New York is now often associated with Wall Street's soaring skyscrapers and monuments, the neon signs of Times Square and the greenery of Central Park, all contributing to the unfading energy of the city. From Broadway shows, to world-class museums like the Museum of Modern Art, to the iconic Statue of Liberty, this exciting city is just brimming with activities to entertain every traveler. It's hard to cover all that New York has to offer with so many attracting options to choose from, and it can make planning your trip a bit baffling. That's why this list will come in handy. Here are the 30 must-dos, ranging from the iconic landmarks to local favorites.

Aluminium Q663

I. Texto sobre el concepto

Aluminum and aluminium are two names for element 13 on the periodic table. In both cases, the element symbol is Al, although Americans and Canadians spell and pronounce the name aluminum, while the British (and most of the rest of the world) use the spelling and pronunciation of aluminium. Why Are There Two Names? You can blame the element's discoverer, Sir Humphry Davy, Webster's Dictionary, or the International Union of Pure and Applied Chemistry (IUPAC). Sir Humphry Davy

proposed the name aluminum when referring to the element in his 1812 book Elements of Chemical Philosophy, even though he had used the name alumium for the element (1808). Despite Davy's two names, the official name "aluminium" was adopted to conform with the -ium names of most other elements. The 1828 Webster's Dictionary used the "aluminum" spelling, which it maintained in later editions. In 1925, the American Chemical Society (ACS) decided to go from aluminium back to the original aluminum, putting the United States in the "aluminum" group. In recent years, the IUPAC had identified "aluminium" as the proper spelling

II. Texto que no es sobre el concepto

In Andean belief, Titicaca is the birthplace of the sun. In addition, it's the largest lake in South America and the highest navigable body of water in the world. Banner blue skies contrast with bitterly cold nights. Enthralling and in many ways singular, the shimmering deep blue Lake Titicaca is the longtime home of highland cultures steeped in the old ways. Pre-Inca Pukara, Tiwanaku and Collas all left a mark on the landscape. Today the region is a mix of crumbling cathedrals, desolate altiplano and checkerboard fields backed by rolling hills and high Andean peaks. In this world, crops are still planted and harvested by hand. Campesinos (peasants) wear sandals recycled from truck tires, women work in petticoats and bowler hats, and llamas are tame as pets.

III. Texto sobre el concepto sin incluir el término

The element 13 is the third most abundant metal in the Earth's crust, and the third most abundant element overall. No other metal can compare to it when it comes to its variety of uses. The metal is incredibly popular because it is: Lightweight, Strong, Resistant to corrosion, Durable, Ductile, Malleable, Conductive, Odorless. The element is also theoretically 100% recyclable with no loss of its natural properties. It also takes 5% of the energy to recycle for reuse The most common uses include: transportation, construction, electrical, consumer goods or even explosives, the metal powder is highly flammable and so one of its most common uses is in pyrotechnic displays. The powder burns very brightly and is used to create different flash effects in fireworks displays by using different grades of powder. It is also used in a similar capacity as an ingredient in blasting agents used in commercial mining (the metal itself is mined from bauxite and then isolated using a chemical process known as the Bayer process.). In the past, when photography was in its infancy, the powder was also used to create camera flashes.

IV. Texto que no es sobre el concepto pero incluye el término

The first actual production can for Coke was a test market can which was produced out of the Hayward, CA plant for export to American Troups overseas in late 1955. A second can from the New Bedford Mass plant for export to the American troops in the far east was produced in early 1956. The Hayward can is quite a bit more difficult to locate however. There is one tell tale identifier on this can which separates it from the

rest. On the side of the can above the seam, the sentence "Prepared for export only" exists. This is an extremely tough can to find and even tougher to find in very good shape. The other somewhat unique feature is in the lids that were used. The original experimental lids did not have any production information, but rather had very plain & somewhat familiar Coke logo's.