Universitat Oberta de Catalunya

Máster en Ciencia de Datos

TRABAJO FINAL DE MÁSTER

ÁREA: PROCESAMIENTO DEL LENGUAJE NATURAL

Desarrollo de una taxonomía específica de dominio y aplicación a un motor de búsqueda

Autor: Víctor Cardoner Álvarez

Tutor: Nadjet Bouayad-Agha

Director: Jordi Casas Roma

5 de junio de 2022

FICHA DEL TRABAJO FINAL

Título del trabajo:	Desarrollo de una taxonomía específica de dominio y aplicación a un motor de búsqueda
Nombre del autor:	Víctor Cardoner Álvarez
Nombre del consultor/a:	Nadjet Bouayad-Agha
Nombre del PRA:	Jordi Casas Roma
Fecha de entrega:	06/2022
Titulación:	Máster Universitario en Ciencia de Datos
Área del Trabajo Final:	M2.881 - TFM - Área 4 aula 1
Idioma del trabajo:	Español
Palabras clave	Natural Language Processing, taxonomy, search engine

Resumen del Trabajo

El objetivo de este trabajo es el desarrollo de una taxonomía y su posterior aplicación a la optimización del motor de búsquedas de una app para madres y profesionales de la salud que facilita la lactancia materna.

La app permite crear un perfil personalizado con datos de la madre y sus bebés para realizar el seguimiento de la lactancia y contiene una base de datos con miles de artículos de consulta organizados por temáticas que las usuarias pueden consultar directamente para resolver sus dudas.

La app también dispone de un motor de búsqueda por palabra clave y de un chat para hablar con consultoras especializadas en lactancia cuando no encuentran las respuestas en los artículos.

La aplicación de modelos de NLP (*Natural Language Processing*) sobre las consultas y búsquedas apoyados en la taxonomía, permitirá mejorar la precisión en las

respuestas a las usuarias y reducir el tiempo que dedican las consultoras a responder de manera manual, lo que, en la actualidad, supone un importante cuello de botella y una limitación de la escalabilidad.

Para el desarrollo de la taxonomía se aprovechará toda la información presente en la app, todo el histórico de consultas realizado por las usuarias con sus respuestas, los conceptos más buscados y las temáticas más relevantes. Una vez definida esta taxonomía, se desarrollará una aplicación para validarla.

Abstract

The main goal of this project is to develop a taxonomy and an apply it to optimize the search engine of an app for mothers and health professionals that facilitates breastfeeding.

The app allows mothers to create a profile for them and their babies to keep track of breastfeeding and includes a knowledge database with thousands of articles organized by specific topics that they can check to clarify their doubts.

The app also has a search engine by keyword and a chatroom. Users can use the chatroom to send questions to the team of breastfeeding consultants when they don’t find the answer in the articles.

The use of NLP models combined with the taxonomy on the searches and questions posted in the chatroom, will improve the accuracy on the search results and answers, and will reduce the time consultants devote to manually answer questions through the chat, which currently is the main bottleneck for scalability.

For the definition of the taxonomy, all the knowledge present in the app, all the questions made by the users and the answers received, the most frequently searched concepts and the main topics identified will be used. Once created, an application to validate it will be developed.

Índice

Lista de figuras

Listado de tablas

Glosario

AI - Artificial Intelligence

EKG – Enterprise Knowledge Graph

IR – Information Retrieval

KO – Knowledge Organization

KOS – Knowledge Organization System

ML – Machine Learning

MRC - Machine Reading Comprehension

NLP – Natural Language Processing

NLU - Natural Language Understanding

OWL – Web Ontology Language

QA – Question-Answering

RDF - Resource Description Framework

RDFS - RDF Vocabulary Description language

SKOS - Simple Knowledge Organization System

SKOS-XL - Simple Knowledge Organization System eXtensions for Labels

URI – Uniform Resource Identifier

W3C – World Wide Web Consortium

XML - Extensible Markup Language

1 Introducción

1.1 Contexto y justificación del Trabajo

Lactapp Women's Health S.L. es una empresa del sector de la salud digital cuyo principal producto es Lactapp, la primera app dedicada a la lactancia y maternidad, con contenidos especializados y que, además, resuelve las dudas de sus usuarias de manera personalizada.

Desde su lanzamiento ha atendido más de 9 millones de consultas con las que ha construido un corpus de conocimiento organizado en multitud de árboles con más de 76.000 ramas.

Aunque muchas usuarias navegan por los diferentes árboles para resolver sus dudas, todavía se realizan miles de búsquedas directas a la app y consultas personalizadas que deben ser respondidas de manera manual por las consultoras especializadas en lactancia, lo que supone un importante cuello de botella y una limitación de cara a su expansión internacional.

Como parte de la estrategia de la compañía, se ha definido un plan para implementar modelos de aprendizaje automático para, por un lado, la optimización del motor de búsquedas y, por otro, la clasificación de preguntas y recomendación de respuestas y conseguir automatizar al máximo el trabajo de las consultoras. El objetivo final es el de escalar de manera rápida sin necesidad de incrementar el equipo de consultoras especializadas manteniendo la satisfacción de las usuarias.

1.2 Motivación personal

Cuando decidí empezar el Máster Universitario en Ciencia de Datos, mi motivación se basaba en una mezcla de inquietud intelectual y curiosidad. Como matemático de formación e inversor en startups tecnológicas de profesión, el avance del aprendizaje automático y las tecnologías alrededor del *big data* suponían campos muy atractivos en los que profundizar.

Al ir avanzando en el programa, fui observando un interés creciente por las técnicas de procesado del lenguaje natural en casi todos los sectores. Además, muchos expertos en IA lo señalaban (y lo señalan) como el campo que más demanda tendrá en los años venideros.

Decidirme a realizar el Trabajo de Fin de Máster (TFM) en Lactapp fue fruto de la casualidad. Conocí Lactapp hace años, después del nacimiento de mi segundo hijo, ya que mi mujer era una usuaria habitual de la app.

Más tarde tuve la suerte de conocer a los fundadores y pude profundizar más en su tecnología y modelo de negocio. Me pareció muy interesante lo que habían logrado con los recursos de que disponían.

He estado siguiendo la evolución de la compañía de cerca y, en cuanto me comentaron que estaban planeando el desarrollo de modelos de NLP para las consultas de las usuarias, les propuse colaborar con ellos en el marco del TFM.

Analizando las diferentes aplicaciones y modelos que querían desarrollar junto con mi tutora, Nadjet Bouayad-Agha, con gran experiencia en proyectos de NLP, vimos que el ámbito más claro para la colaboración podía ser el de la optimización del motor de búsquedas y consultas implementando una taxonomía en el dominio específico de la lactancia materna.

1.3 Objetivos del Trabajo

El principal objetivo de este trabajo ha sido desarrollar una taxonomía específica del dominio de la lactancia materna para mejorar tanto el motor de búsquedas como el de consultas y respuestas de Lactapp.

En la actualidad, las búsquedas se realizan por palabra clave y devuelven enlaces a artículos de la app. Por otro lado, las consultas que realizan las usuarias se derivan a consultoras de lactancia que responden de manera personalizada.

Respecto a las búsquedas, el objetivo final ha sido, por un lado, que éstas permitan el uso de lenguaje natural y no solamente palabras clave y, por otro, que los resultados sean lo más relevantes posible usando la taxonomía para realizar *query expansion*.

Respecto a las consultas, la creación de la taxonomía persigue automatizar al máximo las respuestas sin intervención de las consultoras de lactancia. La mayor dificultad radica en el tipo de consultas que realizan las usuarias. En la mayoría de las ocasiones, incluyen una contextualización de su situación o problema antes de realizar la pregunta. Un ejemplo del tipo de preguntas podría ser: *“mi bebé lleva tres días con fiebre, no está ganando peso y está adormecido, ¿podría esto significar que…?”*. Además, existen datos de perfil que pueden aportar información extra de contexto, como la edad del bebé, su peso, el tipo de alimentación, etc. Esto supone un reto, primero para entender la pregunta, segundo, para enriquecerla con datos de contexto y, tercero, para identificar una respuesta apropiada o derivarla a la consultora.

Para abordar ambas problemáticas y mejorar el motor de búsquedas y consultas se ha definido una taxonomía de conceptos de lactancia a partir del histórico de búsquedas y consultas realizado por las usuarias que permite identificar los conceptos clave en búsquedas y relacionarlas con las respuestas más adecuadas.

Una vez creada la taxonomía, se ha validado su efectividad con una aplicación que identifica los conceptos de la taxonomía en las búsquedas y las relaciona con los artículos y respuestas más apropiadas.

En resumen, los objetivos son los siguientes:

- Desarrollar una taxonomía que pueda ayudar a entender mejor las búsquedas y consultas y las relacione con los artículos y las respuestas más apropiados.
- Validar la taxonomía mediante la implementación de una aplicación de NLP de tipo *proof-of-concept* de mejora de la funcionalidad de búsqueda de la app.

1.4 Enfoque y método seguido

Para llevar a cabo el proyecto se seguirá un enfoque práctico basado en el histórico de consultas y búsquedas que ha recibido la app y las respuestas dadas. Asimismo, se buscará apoyo en el conocimiento de las consultoras de lactancia para determinar los conceptos más relevantes de cara a elaborar una taxonomía de la lactancia completa y exhaustiva.

A partir de todas las consultas y las búsquedas, usando métodos de NLP, se elaborará un listado de conceptos frecuentes que se filtrará hasta obtener los más esenciales.

Posteriormente se enriquecerán con sinónimos en los diferentes idiomas (en un inicio español, catalán e inglés) usando modelos multilenguaje pre-entrenados y de estrategias de *clustering*.

Seguidamente, se establecerán relaciones asociativas y jerárquicas entre ellos mediante identificación de cadenas de búsquedas (*query chains*) y modelos de *clustering*.

Todo esto proporcionará una taxonomía de la lactancia perfectamente adaptada al corpus de conocimiento de Lactapp.

Con esta taxonomía, se implementará un modelo de búsqueda mejorado que permita identificar, a partir de las consultas y búsquedas en lenguaje natural de las usuarias, los conceptos pertenecientes a la taxonomía y, poder así, recomendar respuestas o artículos relevantes.

1.5 Planificación del Trabajo

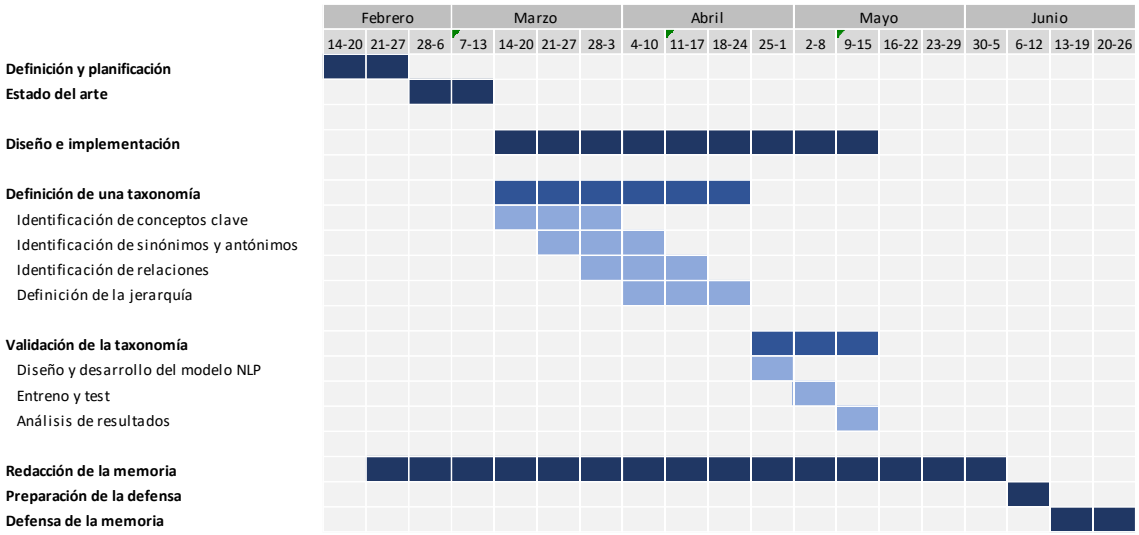


Figura 1 - Planificación del proyecto

2 Estado del arte

2.1 Introducción

La organización del conocimiento es un área de investigación que incluye tanto actividades de descripción, representación, archivo y organización de documentos y representaciones de documentos, como conceptos, llevados a cabo por humanos como por sistemas informáticos (Hjørland, 2008).

Para llevar a cabo estas tareas, se han desarrollado métodos, reglas y estándares, como las taxonomías, los tesauros, las ontologías, los SKOS (*Simple Knowledge Organization System*) o los grafos de conocimiento.

2.2 KOS

Dentro de la organización del conocimiento, el término genérico *Knowledge Organization System* (KOS), o sistema de organización del conocimiento, agrupa todos aquellos esquemas para organizar conceptos que nos permitan organizar, clasificar, definir o recuperar información.

Existen multitud de KOS, siendo los más comunes los listados de términos, taxonomías, tesauros y ontologías. Cada uno de estos KOS tiene sus particularidades y nivel de complejidad, como se puede ver en la Figura 2.

Listado de términos	Taxonomías	Tesauros	Ontologías
Desambiguación	Desambiguación	Desambiguación	Atributos
	Sinónimos	Sinónimos	Clases
	Estructura jerárquica	Estructura jerárquica	Relaciones semánticas
		Relaciones asociativas	

Figura 2 - Diferencias entre KOS

Los usos genéricos de los KOS son variados, incluyendo búsquedas, filtrados, etiquetado, ordenación, navegación por conceptos, visualización de temáticas, etc. En resumen, hacen explícito el conocimiento de la organización.

2.2.1 Taxonomías y tesauros

Las taxonomías son sistemas de organización del conocimiento que agrupan conceptos, sus sinónimos y relaciones jerárquicas entre ellos. Estas relaciones jerárquicas nos permiten “subir” hacia conceptos más amplios (*broader concepts*) o “bajar” hacia

conceptos más específicos (*narrower concepts*). Asimismo, nos permiten definir términos (o etiquetas) preferidos para cada concepto y también sus sinónimos, además de soportar múltiples idiomas.

Las taxonomías pueden contener una o más jerarquías, lo que permite mayor flexibilidad a la hora de organizar la información. Un ejemplo de taxonomía con diferentes jerarquías podría darse al clasificar empresas por el sector en el que operan (Telecomunicaciones > Operadoras > Operadoras móviles > Movistar) o por la geografía en la que operan (Operadoras Europeas > Operadoras Españolas > Movistar).

En la Figura 3 vemos la representación de una taxonomía en un software comercial. La jerarquía se muestra en forma de árbol cuyas ramas se pueden ir ampliando para ver los conceptos contenidos en las ramas.

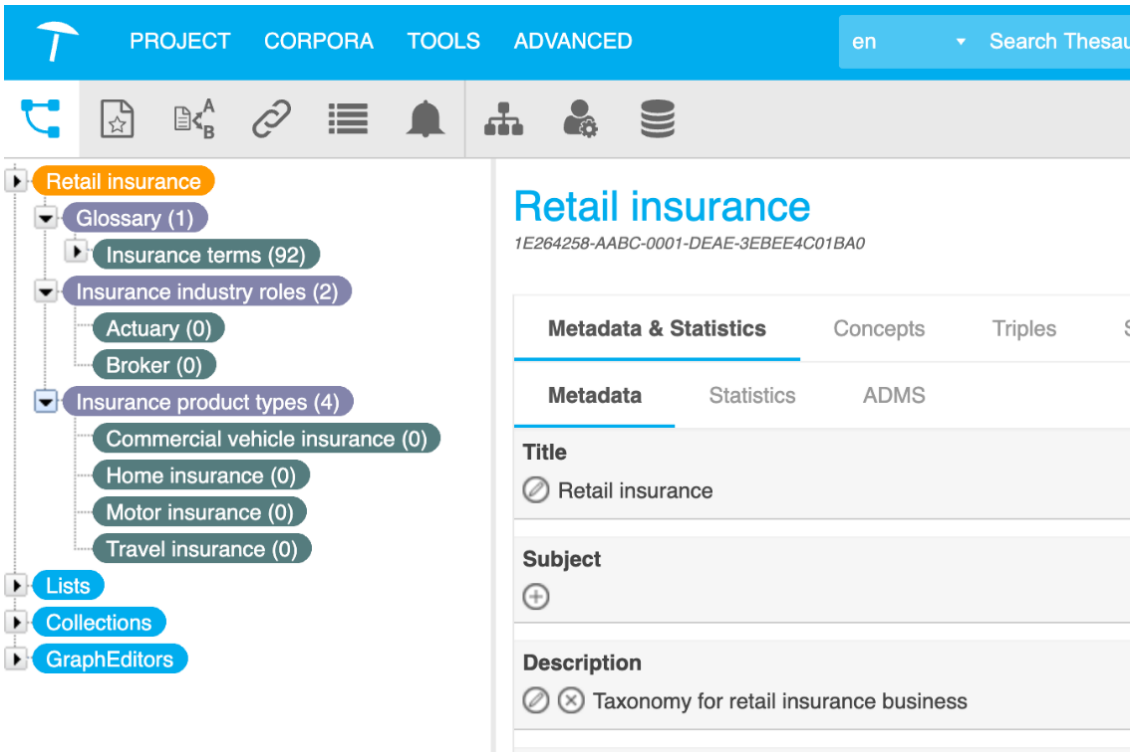


Figura 3 - Representación de una taxonomía en el software comercial Poolparty

Los tesauros son una extensión de las taxonomías, ya que se les añade un nivel mayor de complejidad y expresividad al incorporar relaciones de asociación entre conceptos que no están jerárquicamente relacionados. Esto permite realizar un mayor número de declaraciones de los conceptos.

En la Figura 4 vemos una representación de un tesauro en un software comercial en el que se relaciona la economía agraria con la industria agraria, la economía rural y las compañías del sector agrario.

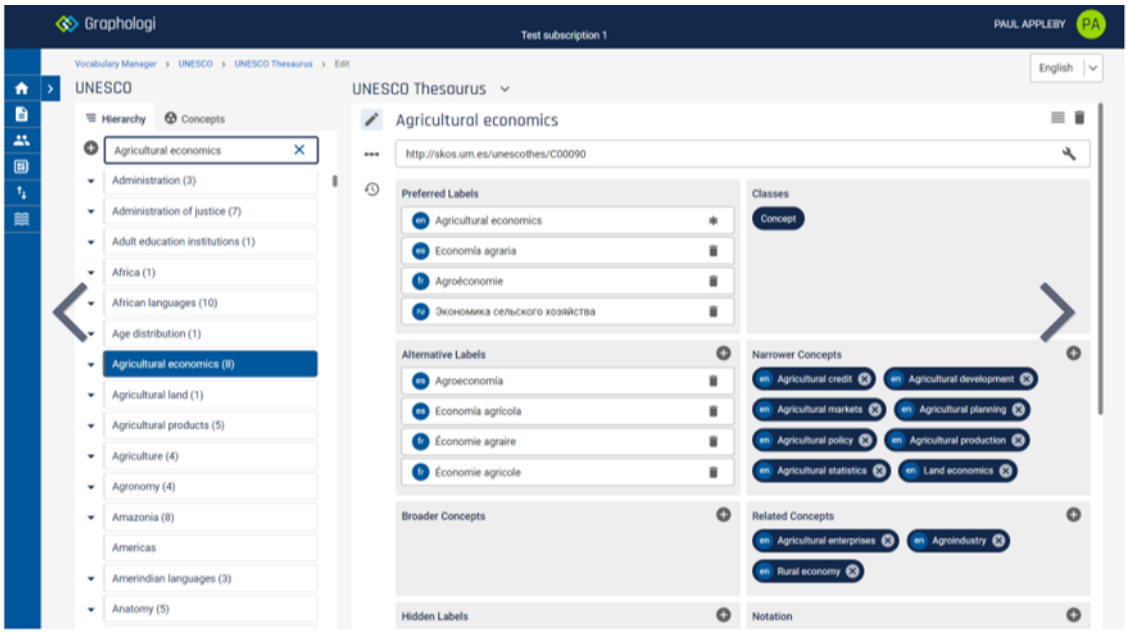


Figura 4 - Representación de un tesauro en el software comercial Graphologi

2.2.2 Ontologías

A diferencia de las taxonomías y los tesauros, una ontología es un conjunto de términos con sus atributos y las relaciones entre ellos, que se usa para representar un dominio. Estos términos, se conocen también como clases, y sus atributos definen las características de estas clases. Además, las relaciones tienen un sentido semántico ya que llevan un significado asociado.

La Figura 5 muestra un ejemplo sencillo de ontología, donde las clases son *Empresa*, *Empleado* y *Ciudad*, sus atributos son *Razón Social*, *CIF* y *Sector,* para las *Empresa*, *Cargo* y *E-mail,* para los *Empleados* y *Provincia* y *País* para *Ciudad*. Finalmente, las relaciones nos permiten afirmar, por ejemplo, que un determinado *Empleado* trabaja en una determinada *Empresa* o que un determinado *Empleado* vive en una determinada *Ciudad*.

Figura 5 - Ejemplo de ontología

De todos los sistemas de organización del conocimiento, las ontologías son los más ricos y complejos a nivel semántico (Biagetti, 2021).

2.2.3 Grafos de Conocimiento Empresarial (EKG)

Los *Enterprise Knowledge Graph* (EKG) son modelos de datos complejos que representan, de manera lógica, la estructura de datos de una empresa. Conectan el conocimiento de diferentes dominios y modelos de conocimiento sin alterar su estructura y sin tener que modificar su arquitectura.

2.3 La web semántica

En su artículo "*The Semantic Web*" (Berners-Lee *et al*., 2001), sus autores introdujeron el concepto de la web semántica como "una extensión de la web actual en la que la información está dotada de un significado bien definido, lo que permite que las computadoras y las personas puedan trabajar de manera cooperativa".

El objetivo principal de la Web Semántica es que la Web pase de ser una simple colección de documentos a convertirse en una base de conocimiento. La Web actual está, en su mayoría, desarrollada en HTML. Los navegadores permiten visualizar la información contenida en estas páginas, pero hacen falta lenguajes que permitan dotar de lógica y significado a los contenidos de la web y que permita a las máquinas entender la información que manejan y sacarle todo el partido. Estos lenguajes incluyen el *Extensible Markup Language* (XML), *Resource Description Framework* (RDF) o el *Web Ontology Language* (OWL).

El consorcio *World Wide Web Consortium* (W3C) está impulsando la web semántica creando estándares y tecnologías para publicar datos legibles por aplicaciones informáticas que permitan a los usuarios construir almacenes de datos en la Web, construir vocabularios y establecer reglas para manejar estos datos.

La Figura 6 muestra la arquitectura de la Web Semántica.

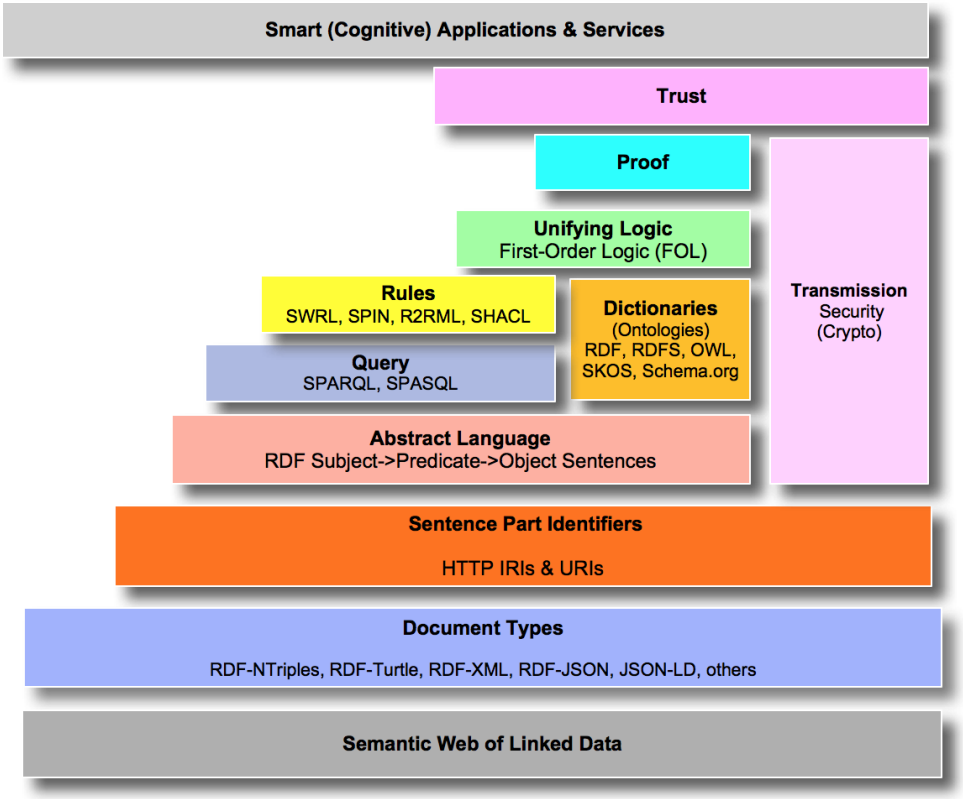


Figura 6 - Las capas de la Web Semántica - W3C's Semantic Web Layer Cake

2.3.1 SKOS y SKOS-XL

Los SKOS son modelos de datos para compartir y enlazar los diferentes sistemas de organización del conocimiento a través de la Web expresados en un formato entendible por las máquinas. Estos modelos permiten capturar la información presente, por ejemplo, en taxonomías, tesauros y ontologías, y compartirla entre diferentes aplicaciones.

Los conceptos provenientes de las taxonomías y los tesauros se identifican mediante *Uniform Resource Identifiers* (URI) en los SKOS. Estos conceptos pueden ser etiquetados y relacionados, tanto jerárquicamente como asociativamente, lo que permite también ser usados como modelos de datos para ontologías.

Cada concepto en el SKOS puede tener múltiples etiquetas (preferida, alternativa) y múltiples propiedades (las relaciones con otros conceptos), pero no se puede dotar de metadatos a las propias etiquetas. No podríamos, por ejemplo, indicar el origen de una etiqueta en concreto, su creador o la fecha en que fue creado. Para solventar esta limitación, el W3C publicó una extensión denominada SKOS-XL, que permite dotar de mayor flexibilidad a los SKOS.

En resumen, el SKOS y SKOS-XL son estándares propuestos por el W3C[1] para llevar las taxonomías y ontologías a la Web Semántica.

2.3.2 Estándares

El estándar SKOS se apoya en diferentes lenguajes, sintaxis y protocolos para su descripción. Los principales son RDF[2], que provee un lenguaje para la representación de la información de recursos en la web, RDFS[3], lenguaje para la descripción de vocabularios RDF, y OWL[4], lenguaje para la descripción de ontologías.

Ejemplos sencillos de SKOS serían:

Relaciones jerárquicas

```
ex:animales rdf:type skos:Concept;
 skos:prefLabel "animales"@es;
 skos:narrower ex:mamíferos.
ex:mamíferos rdf:type skos:Concept;
 skos:prefLabel "mamíferos"@es;
 skos:broader ex:animales.
```

Relaciones asociativas:

```
ex:pájaros rdf:type skos:Concept;
 skos:prefLabel "pájaros"@es;
 skos:related ex:ornitología
ex:ornitología rdf:type skos:Concept;
 skos:prefLabel "ornitología"@es.
```

La siguiente tabla contiene la lista de conceptos, etiquetas, relaciones semánticas, mapeos, documentación y colecciones:

Concept	Label& Notations	Documentation	Semantic Relations	Mapping Properties	Collections
Concept	prefLabel	note	broader	broadMatch	Collection
ConceptScheme	altLabel	changeNote	narrower	narrowMatch	orderedCollection
inScheme	hiddenLabel	definition	related	realtedMatch	member
hasTopConcept	notation	editorialNote	broaderTransitive	closeMatch	memberList
topConceptOf		example	narrowerTransitive	exactMatch	
		historyNote	semanticRelation	mappingRelation	
		scopeNote			

Tabla 1 - Elementos de los SKOS

[1] SKOS Simple Knowledge Organization System Reference - https://www.w3.org/TR/skos-reference/
[2] https://www.w3.org/TR/rdf-primer/
[3] https://www.w3.org/TR/rdf-schema/
[4] https://www.w3.org/OWL/

2.4 Desarrollo de KOS

El desarrollo de un KOS conlleva tomar una serie de decisiones en lo que respecta a las categorías, niveles, terminología, etc. En este proceso de toma de decisiones intervienen tanto el experto del dominio como el encargado del desarrollo del KOS. En muchas ocasiones, éste último debe tomar las decisiones en previsión de futuros problemas.

Los parámetros más habituales a definir son la estructura del KOS - habitualmente una jerarquía o conjunto de jerarquías -, el número de niveles - normalmente determinado por el dominio y el uso - y el número de conceptos por nivel - también muy dependiente del dominio y el uso.

El origen de los términos puede venir dado por un experto del dominio, se pueden extraer manualmente de un corpus de información o se pueden extraer de forma automática con técnicas de NLP.

2.5 Herramientas

Existen multitud de herramientas en el mercado para el diseño y almacenamiento de KOS, desde el uso de hojas de cálculo hasta completas aplicaciones comerciales con multitud de funcionalidades.

Las principales herramientas, con sus características y funcionalidades principales se listan a continuación

Herramienta	Usos	Características
MultiTes Pro	Taxonomías/Tesauros	
http://www.multites.com/productspro.htm		
Synaptica KMS	Taxonomías/Tesauros	Estándares RDF, OWL, SKOS
https://www.synaptica.com/synaptica-kms/		
Synaptica Graphite	Taxonomías, Tesauros, Ontologías, SKOS/RDF	Estándares RDF, OWL, SKOS
https://www.synaptica.com/synaptica-kms/		
Graphologi	Taxonomías, Tesauros, Ontologías, SKOS/RDF	Estándares OWL, SKOS, SKOS-XL
https://graphologi.com/		
Smartlogic Semaphore	Taxonomías, Tesauros, Ontologías, SKOS/RDF	Soporta SKOS, RDF, ISO 25964 Importa/exporta CSV, XML (RDF SKOS, Turtle, N Triple), SQL databases, y ficheros MultiTes
https://www.smartlogic.com/semaphore		
Data Harmony	Taxonomías/Tesauros	
https://www.accessinn.com/data-harmony/		

Mondeca ITM	Taxonomías, Tesauros, Ontologías, SKOS/RDF	Exporta a Excel, XML, RDF, SKOS y Topic Maps
https://mondeca.com/itmx		
PoolParty	Para taxonomías, tesauros, ontologías, SKOS/RDF	Construido sobre estándares W3C para Web Semántica: SKOS, RDF, OWL, SPARQL Importa/exporta: Excel, N3, N-Quads, Trix, Binary-RDF, MultiTes, RDF/XML, Turtle, N-Triples, RDF/JSON, Trig, JSON-LD y Zthes Conectores a SharePoint, Drupal, WordPress, Confluence, Alfresco, FontoXML
https://www.poolparty.biz/taxonomy-thesaurus-management/		
TopQuadrant TopBraid EDG	Para taxonomías, tesauros, ontologías, SKOS/RDF	Taxonomías SKOS o SKOS-XL; ontologías basadas en SHACL o OWL Importa/exporta: Excel/CSV, XML, RDF/OWL
https://www.topquadrant.com/products/topbraid-enterprise-data-governance/		
TopQuadrant TopBraid Composer	Para ontologías	
https://www.topquadrant.com/products/topbraid-composer/		
VocBench	Para taxonomías, tesauros, ontologías, SKOS/RDF	Open source
http://vocbench.uniroma2.it/		
TemaTres	Para taxonomías, tesauros, ontologías, SKOS/RDF	
https://vocabularyserver.com/web/		
Protégé	Para ontologías únicamente	Desarrollado por el Center for Biomedical Informatics Research de la Universidad de Medicina de Stanford
https://protege.stanford.edu/		
Infoneer SKOS Tool	Para taxonomías, tesauros, ontologías, SKOS/RDF	Desarrollado por el Engineering Informatics Research Group de la Texas State University
https://infoneer.wp.txstate.edu/software/		
iQvoc	Soporte para SKOS y SKOS-XL	Open Source
https://iqvoc.net/		
SKOSEd	Plugin para Protégé para crear y editar taxonomías y tesauros	

https://code.google.com/archive/p/skoseditor/	
ThManager	Herramienta Open Source para crear y mantener SKOS RDF
https://thmanager.sourceforge.io/	

Tabla 2 - Herramientas para KOS

2.6 NLP para el desarrollo de taxonomías y ontologías

El desarrollo de taxonomías y ontologías de un dominio específico es una tarea cada vez más importante debido al creciente volumen de información disponible y a la necesidad de procesar toda esa información y extraer conocimiento útil.

Identificar la información relevante de manera manual resulta inviable en la mayoría de dominios, lo que ha ayudado al crecimiento de un área de investigación en torno a la extracción de información usando técnicas de procesado del lenguaje natural.

Este proceso de extracción de información de un corpus de documentos o textos sigue un proceso que acostumbra a incluir, al menos, las siguientes fases:

1. Transformación de documentos a texto
2. Segmentación en frases
3. Tokenización
4. Lematización
5. Identificación y etiquetado de relaciones sintácticas
6. Análisis semántico
7. Detección de entidades:
    a. Detección basada en reglas
    b. Detección basada en aprendizaje automático
    c. Detección híbrida
8. Desambiguación de conceptos
9. Identificación de relaciones jerárquicas, asociativas y/o semánticas
10. Construcción de la taxonomía/ontología

Se pueden encontrar ejemplos de este proceso en (Strinyuk *et al*., 2021), usando un corpus de información sobre tránsito marítimo, o en (Medelyan *et al.*, 2013), sobre una colección de documentos.

2.7 Campos de aplicación de taxonomías y ontologías

Las taxonomías y ontologías tienen multitud de casos de uso relacionados con la organización de la información y su posterior explotación. Algunos de los más relevantes son:

- Etiquetado e indexado
- Nevegación entre temáticas y categorías

- Mejora en los resultados de las búsquedas al identificar cadenas de texto con conceptos de la taxonomía
- Identificación de la información relevante relacionada con la búsqueda
- Filtrado de resultados según conceptos de la taxonomía
- Consistencia en los metadatos
- Personalización de resultados o recomendaciones
- Mejora de los grafos de conocimiento

A continuación, se detallan algunos casos de uso reales de taxonomías, ontologías y grafos de conocimiento.

2.7.1 Categorización

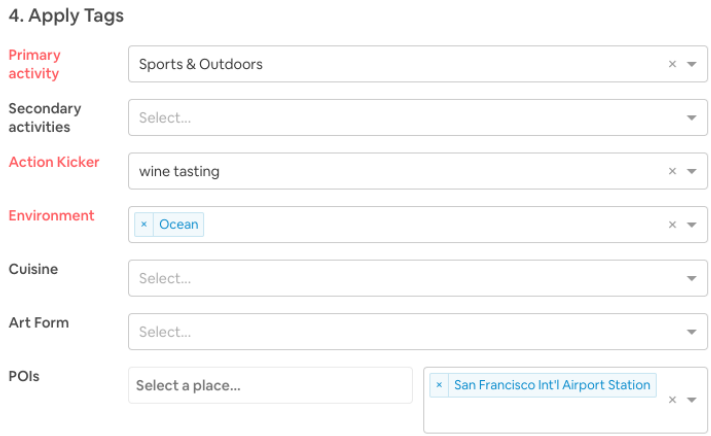
2.7.1.1 Airbnb

La plataforma de reserva de alojamientos Airbnb hace uso de una taxonomía y un grafo de conocimiento para categorizar las propiedades listadas y ofrecer información útil de viaje a los usuarios[5].

El desarrollo de una taxonomía (Figura 8) surge de la necesidad de categorizar el enorme stock de propiedades listadas, las experiencias y otros productos de la plataforma, y ofrecer la mayor cantidad de información posible a los usuarios, de manera consistente y precisa.

Esta categorización, inicialmente manual (Figura 7), se mostró poco escalable rápidamente y forzó la introducción de un enfoque híbrido, en el que se inferían categorías de la descripción de la propiedad, su entorno y el perfil del propietario, para después validarla con el propietario.

Esta taxonomía alimenta también al grafo de conocimiento, la herramienta que permite a Airbnb ofrecer información relevante para sus usuarios, relacionando la propiedad reservada con experiencias en la zona, información del destino, eventos y otros datos relevantes (Figura 9).



[5] https://medium.com/airbnb-engineering/contextualizing-airbnb-by-building-knowledge-graph-b7077e268d5a

Figura 7 - Herramienta para la categorización de experiencias

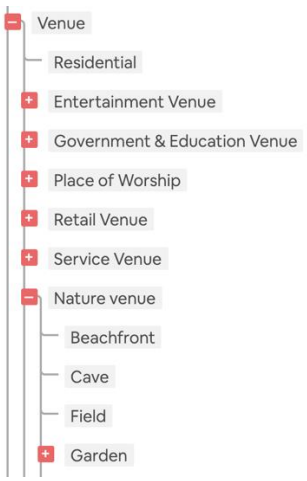


Figura 8 - Detalle de la taxonomía

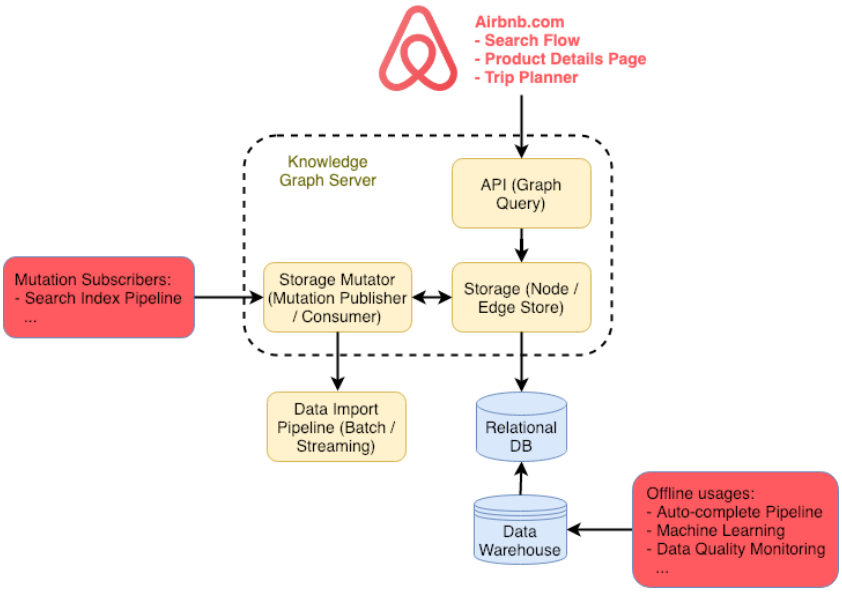


Figura 9 - Arquitectura del grafo de conocimiento de Airbnb

2.7.1.2 LinkedIn

Igual que en el caso de Airbnb, LinkedIn[6] dispone de una taxonomía de las entidades presentes en la red social y la usa para alimentar un grafo de conocimiento.

La taxonomía contiene la identidad de todas las entidades y sus atributos. Estas entidades son creadas de dos formas:

1. Creadas por los usuarios de la red: usuarios, empresas, eventos, ofertas de empleo…

[6] https://engineering.linkedin.com/blog/2016/10/building-the-linkedin-knowledge-graph

2. Creadas por LinkedIn: habilidades, idiomas, títulos, certificados, ubicaciones

Estas entidades se convertirán en los nodos del grafo de conocimiento, pero hay un proceso previo de limpieza de las entidades generadas por los usuarios para evitar que haya nombres incorrectos, inválidos o sin sentido. Este proceso consta de 4 fases:

1. Generación de candidatos a entidades
2. Desambiguación mediante *clustering* para contextualizar la entidad
3. Eliminación de duplicados, mediante un proceso híbrido de *clustering* y validación manual
4. Traducción a otros idiomas

El grafo de conocimiento se alimenta de estas entidades de la taxonomía y de las relaciones entre ellas. Gracias a esta herramienta, LinkedIn es capaz de mejorar su motor de recomendaciones, búsquedas, monetización y analítica.

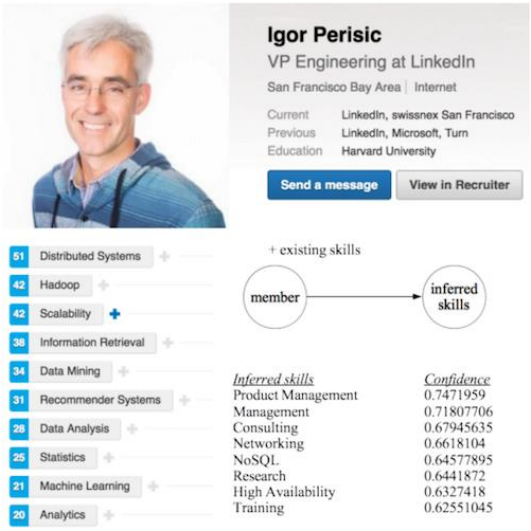


Figura 10 - Ejemplo de inferencia en habilidades

2.7.2 Búsquedas

Otro caso de uso claro de taxonomías y ontologías es el de las búsquedas.

Como hemos visto con anterioridad, la Web Semántica ofrece un enfoque prometedor de la Web actual en que, tanto máquinas como personas, pueden interpretar la información y hacer un uso inteligente de ella. Esto tiene un impacto directo en las búsquedas que, con este nuevo enfoque semántico, se apoyarán, por un lado, en el uso de ontologías y, por otro, en técnicas de recuperación de la información (IR) para ofrecer resultados más relevantes.

En (Oyebisi Oyefolahan *et al*., 2018), los autores exploran diversas técnicas de IR basadas en ontologías para superar las limitaciones actuales de las búsquedas basadas

en palabras clave. Las técnicas más relevantes son *Query Expansion* (QE) y *Semantic Annotation* (SA).

En (Aguilar-Lopez *et al*., 2009) los autores proponen un modelo para aumentar la relevancia de los resultados de motores de búsqueda basándose en el uso de ontologías de dominio, en su caso WordNet, y otros vocabularios informales, midiendo el aumento de relevancia como la disminución del tiempo necesario en validar los resultados.

2.7.3 Recomendaciones

Los motores de recomendación se ven también beneficiados al combinar taxonomías y ontologías en los procesos de generación de contenidos relevantes para los usuarios.

Por ejemplo, en (Tarus *et al*., 2018), los autores muestran que el uso de ontologías como representación del conocimiento en el ámbito de los sistemas de recomendación de e-learning mejora la calidad y las recomendaciones.

En el sector del e-commerce, (Bagherifard *et al.*, 2017) muestra un enfoque alternativo a los sistemas de recomendación del sector, usando ontologías junto a las técnicas de *Collaborative Filtering* y *Content-Based Filtering* para mejorar la precisión de las recomendaciones.

3 Construcción de la taxonomía

3.1 Introducción

Para la construcción de la taxonomía específica de dominio se ha seguido un enfoque basado en la combinación de estrategias y modelos de NLP con un filtrado y etiquetado manual.

Se ha perseguido limitar los procesos manuales al máximo, aunque, para ciertas tareas, se ha requerido la aplicación de un criterio experto para el filtrado, la desambiguación y el enriquecimiento de conceptos.

3.2 Fuentes de datos

La construcción de la taxonomía se ha centrado en dos fuentes principales de datos. La primera fuente la componen las consultas realizadas por las usuarias a las expertas en lactancia en el chat. La segunda, la forma la lista de conceptos consultados por las usuarias usando el buscador de la app.

3.2.1 Consultas realizadas en el chat

Desde el lanzamiento de la app, las usuarias han realizado más de 9 millones de consultas, con volúmenes superiores a 500.000 consultas al mes puntualmente durante el último año.

Las consultas de las usuarias en el chat acostumbran a ir acompañadas de información contextual, como la edad del bebé, peso o evolución, para plantear la duda a partir de esa base.

Aunque las consultas son muy variadas y cubren un amplio espectro de temáticas, hay una gran concentración de consultas alrededor de las problemáticas propias de la lactancia, el crecimiento y la salud.

3.2.2 Búsquedas

Para tratar los datos provenientes de las consultas de conceptos usando el buscador, se han usado dos enfoques.

El primero ha sido tratar los conceptos buscados de manera individual, agregándolos para ver su frecuencia. De esta manera, se ha conseguido identificar más de 74.000 conceptos buscados. Las frecuencias oscilan entre las 12.000 apariciones del concepto más buscado (crisis) y los conceptos con una única búsqueda (habitualmente frases más elaboradas y errores).

El segundo enfoque ha consistido en analizar cadenas de consultas o *query chains*. Una *query chain* es una secuencia de consultas realizadas por la misma usuaria en el mismo día, habitualmente muy cercanas en el tiempo.

Se dan casos de cadenas de consultas de conceptos relacionados entre sí (búsquedas refinadas en función de los resultados obtenidos) y otras ocasiones en que los conceptos no tienen relación entre sí (diferentes dudas en busca de información).

El primer enfoque (búsquedas aisladas) ofrece un listado de conceptos a añadir a la taxonomía, el segundo (*query chains*) permite enriquecer los conceptos con sinónimos, encontrar relaciones semánticas, jerárquicas y asociativas.

3.3 Extracción inicial de conceptos

3.3.1 Conceptos extraídos del chat

Para obtener un primer conjunto de conceptos para la taxonomía, se han usado las consultas de las usuarias en el chat de expertas.

Un ejemplo de consulta sería el siguiente:

> *"hola, soy mama primeriza de un bebé de seis semanas y querría saber si hay algún alimento (brócoli, coliflor, espárragos verdes...) que afecte al sabor de la leche o a la lactancia... muchas gracias"*

Si usamos esta consulta de ejemplo, se observa que se pueden extraer muchos conceptos relacionados con la lactancia como "mamá primeriza", "bebé de seis semanas", "alimento" o "sabor de la leche".

Como un procesado manual es completamente inviable dado el volumen de consultas, se ha seguido una metodología estándar de procesado del lenguaje natural y posterior proceso de anotación. El esquema seguido ha sido el siguiente:

1. Transformación de las consultas en frases
2. Tokenización de las palabras
3. Lematización: sustantivos en género neutro, verbos en infinitivo
4. Generación de n-gramas hasta orden 3 - unigramas, bigramas y trigramas. A notar que los *stopwords* no cuentan en este cálculo, así "sabor de la leche" es un bigrama y "bebé de seis semanas" es un trigrama.
5. Agregación y cálculo de frecuencia de aparición.

Tras filtrar los n-gramas con una frecuencia de aparición mayor a 10, se obtienen unos 24.000 n-gramas para su posterior proceso de anotación. En la tabla 3 se puede ver la distribución desglosada de los n-gramas con ejemplos.

n-grama	Número	Frecuencia agregada	Ejemplos
Unigramas	14.467	5.182.848	“vientre”, “trabajo”, “regurgitar”
Bigramas	6.101	1.258.335	“sospecha de infección”, “agarre profundo”, “alimentación infantil2
Trigramas	3.861	1.810.805	“proteína de la leche de vaca”, “atraganta cuando mama”, “coger más parte de la areola”

Tabla 3 - Distribución de n-gramas

3.3.2 Conceptos extraídos de las búsquedas

Al conjunto de n-gramas obtenidos a partir de las consultas en el chat, se añaden también los términos más frecuentes de las búsquedas. Por consistencia, se añaden únicamente aquellos términos que han sido buscados más de 10 veces, que son unos 4.000 en total.

3.3.3 Query chains

Como se ha explicado en el punto 3.2.2, las *query chains* son búsquedas realizadas por la misma usuaria en el mismo día, relacionadas o no entre sí.

De los 9 millones de búsquedas presentes en el histórico de la app, eliminando duplicados, se obtienen 470.000 búsquedas.

Seleccionando únicamente aquellas búsquedas de al menos dos conceptos diferentes realizadas por la misma usuaria el mismo día, se consiguen 179.950 *query chains*.

En estas cadenas de búsquedas se dan muchas casuísticas. Generalizando, se han observado los tipos especificados en la siguiente tabla:

Tipo de consulta	Ejemplo
Específico a general	"Destete respetuoso" - "Destete" "Crisis de los 12 meses" - "Crisis" "amoxicilina" - "antibióticos"
Sinónimos	"lactancia en gemelos" - "lactancia tandem"
Frases cortas a largas	"conservación" - "conservación leche"
Reformulación	“ya no tengo leche" - "pérdida de producción de leche"
Búsquedas relacionadas	"baby-led weaning" - "recetas"
Errores ortográficos	"empacho" - "empacha" - "empachado"
General a específico	"biberón" - "rechaza el biberón"

Tabla 4 - Tipos de consultas en las "query chains"

Hay cadenas de 2 búsquedas y cadenas de más de 10 conceptos. El máximo observado son 26 conceptos en una misma cadena.

Estas cadenas ofrecen mucha información a nivel semántico, jerárquico y asociativo de los conceptos por lo que se han usado en diversos procesos del enriquecimiento de la taxonomía.

Para la mayoría de los casos de aplicación de las *query chains*, se ha considerado que las cadenas son transitivas y simétricas. Por ejemplo, en una cadena formada por 3 conceptos A – B – C, se considerarán igualmente *query chains* las parejas A – B, A – C y B – C, así como B – A, C – A y C – B.

De hecho, para extraer sinónimos y relaciones, se han extraído todos los pares de conceptos que se han observado en alguna *query chain* y se ha calculado su frecuencia. Existen 139.641 pares de conceptos diferentes extraídos de *query chains*, de los cuales 1.211 tienen una frecuencia de aparición superior o igual a 10. Por contrastar esta cifra, hay 122.000 pares de conceptos que únicamente se dan una vez.

3.4 Proceso de anotación

Con el listado de n-gramas de las consultas y los conceptos de las búsquedas, se ha construido el siguiente esqueleto de la taxonomía, basado en la estructura estándar de un SKOS.

Columna	Descripción
n-grama	Término detectado en la consulta. Puede ser un unigrama (un solo token), un bigrama (2 tokens) o un trigrama (3 tokens). Los *stopwords* no se tienen en cuenta (como se comenta en el punto 3.3.1).
Frecuencia	El número de veces que aparece el n-grama en el dataset. Sólo se han guardado los n-gramas con una frecuencia mínima de 10 apariciones.
n>1	Es igual a 0 si se trata de un unigrama (un token) y 1 si no.
Guardar	Etiqueta que indica si el n-grama es válido (1) o no (0). Un n-grama es válido si expresa un concepto válido dentro del dominio, como por ejemplo "lactancia materna" o "biberón". Algunos ejemplos de n-gramas que no expresan conceptos válidos dentro de nuestro dominio podrían ser "convertir" o "email".
Etiqueta	Esta columna contiene el n-grama correcto, con sus acentos, género y conjugación de verbo.
Concepto	Concepto al que se refiere el n-grama. Si el n-grama es “dar el pecho”, el concepto podría ser “lactancia materna”. Según esta definición, diferentes n-gramas/etiquetas pueden referirse al mismo concepto.
Concepto más amplio (*broader concept*)	Concepto más general al del concepto al que se refiere el n-grama. Si el concepto fuera “lactancia diferida” el concepto más amplio (*broader concept*) sería “lactancia materna”, ya que la

	lactancia diferida es un concepto más concreto de la lactancia materna.
Categoría	Esta columna contiene la categoría raíz a la que pertenece el concepto. Las categorías se han definido partiendo de la categorización de los temas en la app. Estas categorías están descritas más adelante.
Idioma	El lenguaje del n-grama en ISO standard. Por defecto es español (“es”). Si el n-grama está en catalán, entonces el idioma es "ca", si el n-grama está en inglés, el idioma es "en". Conceptos en otros idiomas se han descartado por carecer de relevancia.
Término preferido	Indica si el n-grama es el término preferido del concepto (valor 1) o no (valor 0). Hay un término preferido para cada concepto por idioma. En los casos en que existen varios n-gramas para un mismo concepto, se usa el criterio de frecuencia para determinarlo y el criterio de las expertas cuando la frecuencia no discrimina.
Término alternativo	Indica si el n-grama es una alternativa a un concepto existente (1) o no (0).
Término oculto	Indica si el n-grama es una alternativa a un concepto existente pero no visible (1) o no (0). Estos conceptos suelen responder a acepciones informales o incorrectas de conceptos de la taxonomía, como “bibe” por “biberón” o “guatita” por “barriga”.
Texto 1 **Texto 2** **Texto 3**	Cada una de esas columnas presenta una oración donde aparece el n-grama. Útil para contextualizar el n-grama cuando su significado no es obvio. Estos textos no existen cuando el origen son búsquedas o *query ch*ains ya que no hay información de contexto.
Origen	Indica si el concepto proviene de la búsquedas, las consultas del chat, las *query chains*, las anotaciones manuales o cualquier otro origen.

Tabla 5 - Estructura inicial de la taxonomía

El proceso de anotación no se puede automatizar completamente y requiere de una combinación de procesos automáticos con otros manuales de selección y clasificación.

Procesos como eliminar duplicados o determinar el idioma del concepto se pueden automatizar, pero determinar qué conceptos son relevantes para la taxonomía, por ejemplo, requiere de un proceso manual apoyado en conocimiento del dominio.

Por simplicidad, se ha partido de las siguientes premisas:

- Las etiquetas se añadirán en minúscula y podrán contener errores ortográficos, siempre que su frecuencia lo justifique.
- El concepto estará escrito correctamente, sin errores ortográficos.
- El concepto tendrá siempre una etiqueta idéntica a su término preferido.

- Los *broader concepts* tendrán también un concepto idéntico a ellos. Si no existe de partida, se creará.
- Todos los conceptos que no tenga una etiqueta idéntica a ellos serán etiquetados como términos alternativos u ocultos.

Para ilustrar estas premisas, a continuación, se muestra un ejemplo de conceptos y etiquetas de la taxonomía:

Etiqueta	Concepto	*Broader Concept*	Categoría	Término preferido	Término oculto	Término alternativo	Idioma
areola	areola	pecho	madre	1			es
aureola	areola	pecho	madre		1		es
parte de la areola	areola	pecho	madre			1	es
areolar	areola	pecho	madre			1	es
zona de la areola	areola	pecho	madre			1	es
aureola	areola	pecho	madre		1		es
pecho	pecho		madre	1			es

Tabla 6 - Ejemplo de conceptos y etiquetas de la taxonomía

Este proceso ha generado 2.122 etiquetas en la taxonomía.

3.4.1 Definición de categorías

La elección de las categorías generales para agrupar los conceptos está inspirada en las usadas en la app, pero con un mayor nivel de abstracción y generalización.

Categoría	Descripción
Madre	Agrupa todos los conceptos relacionados con la madre, incluyendo el embarazo, el parto, la alimentación, el descanso, etc.
Bebé	Agrupa los conceptos relacionados con el bebé, incluyendo su crecimiento y desarrollo, descanso, transporte, etc.
Lactancia	Todo lo relacionado con la lactancia materna, lactancia mixta, el destete, la transición a sólidos, extracción y conservación de la leche y problemas relacionados con la lactancia.
Alimentación	Todos los conceptos de alimentación no relacionados con la leche materna.
Salud	Todos los conceptos relacionados con la salud tanto de la madre como del bebé, medicación, tratamientos, dolencias y prevención.
Lactapp	Todos los conceptos relacionados con la app, incluyendo su blog, el chat, las consultas, las expertas en lactancia, webinars y formación.

Tabla 7 - Categorías de la taxonomía

Cada concepto puede pertenecer a una única categoría. Cuando se dan casos en que el concepto puede pertenecer a dos categorías, se desambigua. Un ejemplo de desambiguación sería “alcohol”, que podría pertenecer a la categoría “alimentación” o, según el contexto en que se usa, podría pertenecer a la categoría “lactancia”, en el contexto del impacto que puede tener el alcohol en la lactancia materna. En este caso concreto, la desambiguación se realiza distinguiendo el concepto “alcohol” como “bebida” dentro de la categoría “alimentación” y el concepto “alcohol y lactancia” dentro de la categoría “lactancia”.

3.4.2 Relaciones semánticas

Una de las tareas más críticas en la construcción de la taxonomía es la de agrupar términos semánticamente cercanos, de manera que se puedan agregar conceptos a la taxonomía, identificar sinónimos (términos alternativos) y conceptos erróneos pero frecuentes (términos ocultos).

Esta tarea se ha realizado en 3 fases. La primera ha consistido en usar un modelo pre-entrenado para representar nuestro listado inicial de conceptos en un espacio vectorial, calcular las distancias semánticas entre ellos e identificar otros conceptos cercanos del propio modelo. Este proceso se puede visualizar mediante *TensorBoard* al proyectar los vectores en un espacio tridimensional usando sus componentes principales (PCA).

En la segunda fase se han usado los vectores generados en la fase anterior para realizar un proceso de *clustering* mediante un modelo de *Affinity Propagation*. Este modelo no requiere definir el número de clústers a los que ajustar los *datapoints*, con lo que se obtienen diferentes distribuciones de clústeres en función de la distancia usada. Tras esta fase se pueden identificar agrupaciones de conceptos claras con las que iniciar el proceso de agregación de conceptos y generación de sinónimos.

Por último, una vez agrupados los conceptos, se pueden obtener más relaciones semánticas usando las *query chains*. Conceptos en la misma *query chain* que son sinónimos según la agrupación de la fase anterior o que están dentro de la misma categoría según el proceso inicial de anotación, se agregan y se añaden a la lista de sinónimos.

Estas tres fases se detallan a continuación.

3.4.2.1 Uso de un modelo de Sentence Transformers y visualización mediante TensorBoard

Para obtener sinónimos de los conceptos, se ha usado un modelo pre-entrenado de *Sentence Transformers* multilingüe, en concreto el denominado “distiluse-base-multilingual-cased-v1”[7]. El modelo genera *embeddings* de los conceptos que se han

[7] https://www.sbert.net/docs/pretrained_models.html#sentence-embedding-models/

identificado en el proceso de anotación y que provienen de las búsquedas y las consultas previamente trabajadas.

Estos *embeddings* son representaciones en forma de vector de los conceptos que incluyen su significado codificado. Así, dos vectores cercanos usando una métrica o distancia determinada dentro de este espacio vectorial, son también cercanos desde un punto de vista semántico.

Este procedimiento es muy útil a la hora de encontrar sinónimos de los conceptos ya anotados, permitiendo enriquecer la taxonomía mediante etiquetas alternativas.

Para visualizar los resultados se ha utilizado *TensorBoard*[8], una herramienta que permite visualizar el espacio vectorial generado y verificar la coherencia de los resultados.

Figura 11 – Visualización de los embeddings usando TensorBoard

Para cada concepto se obtiene una lista de puntos cercanos dentro del mismo espacio. Se puede definir la cercanía usando dos distancias, coseno y Euclídea. Tras un análisis detallado de los resultados, se observa que la distancia Euclídea ofrece mejores resultados que la del coseno. Se observa que dentro de una distancia inferior a 0.2, los conceptos son semánticamente similares. Se fija esta distancia como el umbral de tolerancia.

Para cada concepto de entrada al modelo, se extraen un máximo de 5 *embeddings* cuya cercanía está dentro del umbral de tolerancia definido. Estos conceptos se añaden a la taxonomía como conceptos sinónimos.

Este proceso, tras eliminar duplicados y conceptos no relevantes para el dominio, ha añadido 534 etiquetas nuevas a la taxonomía.

[8] https://www.tensorflow.org/tensorboard

3.4.2.2 Clustering

También se ha usado un método de *clustering* con los conceptos de las búsquedas y *embeddings* obtenidos con el modelo pre-entrenado de *Sentence Transformers*. El método en concreto es el *Affinity Propagation* que, a diferencia de otros métodos de *clustering* como *K-Means* o *K-medoids*, no requiere determinar el número de clústeres a priori, sino que el propio modelo identifica elementos "ejemplares" de cada clúster en los elementos de entrada del modelo.

Usando dos distancias diferentes, coseno y Euclídea, se han podido obtener dos clusterizaciones diferentes, la primera de 45 clústers y la segunda de 864.

En un análisis en detalle de los resultados obtenidos con cada distancia, se observa una mejor clasificación usando la distancia Euclídea, como en el apartado anterior.

Estos clústeres permiten, por una parte, validar la clasificación realizada en el proceso de anotación y, por otro, identificar otras relaciones semánticas entre los conceptos de la taxonomía.

Con estos clústeres se realiza un proceso de agregación de conceptos similares y se expande la lista de sinónimos.

El proceso de *clustering*, tras la eliminación de duplicados y filtrado, ha añadido 1.171 etiquetas nuevas a la taxonomía.

3.4.2.3 Extracción de relaciones semánticas de las *query chains*

Gracias al proceso de anotación y al de agregación de conceptos, se pueden usar parejas de conceptos que aparecen en las *query chains* para identificar similitud semántica. Analizando los conceptos, sus etiquetas *broader*/*narrower* y las categorías a las que pertenecen, se puede determinar que parejas de términos de una misma *query chain* que pertenezcan a la misma categoría, tienen cercanía semántica. Se obtiene un resultado similar al de las fases anteriores, en el que esta cercanía semántica nos permite agregar conceptos e identificar sinónimos.

Por último, las *query chains*, han añadido 2.303 etiquetas a la taxonomía.

3.4.3 Relaciones jerárquicas

Dentro de la taxonomía existen relaciones jerárquicas entre conceptos de varios niveles. Cada concepto puede tener, dentro de una taxonomía, una relación jerárquica más amplia (*broader*) y una más específica (*narrower*). Tal y como están definidas en el estándar SKOS, estas relaciones jerárquicas pueden ser transitivas o no. Así, son igualmente válidas relaciones jerárquicas transitivas como no transitivas.

```
ex:medicamento rdf:type skos:Concept;
  skos:prefLabel "medicamento"@es;
```

```
    skos:narrower ex:paracetamol.

ex:crecimiento rdf:type skos:Concept;
    skos:prefLabel "crecimiento"@es;
    skos:narrower ex:peso.
```

En el primer ejemplo, “paracetamol” es un tipo de “medicamento”, con lo cual es una relación transitiva. En el segundo ejemplo, “peso” no es un tipo de “crecimiento”, pero claramente es un concepto relacionado con el crecimiento, junto a la “talla”, “percentil” y otros conceptos *narrower* de crecimiento. Decimos que ambas relaciones son jerárquicas pero la primera es transitiva y la segunda no.

Para identificar las relaciones jerárquicas de los tipos definidos en el estándar SKOS, se han usado diferentes estrategias.

3.4.3.1 Uso de *query chains* y representación en grafo

Para explotar el potencial de las *query chains*, se ha desarrollado un grafo ponderado en el que cada nodo es un concepto presente en alguna de las *query chain* y la arista entre dos nodos existe si esos dos conceptos representados en el nodo han aparecido juntos en una *query chain*, independientemente del orden en el que han aparecido. A cada arista se le ha asociado un peso que equivale al número de ocasiones en que esos dos conceptos han aparecido juntos en una *query chain.*

Se ha usado Neo4j para construir el grafo y simplificar su consulta.

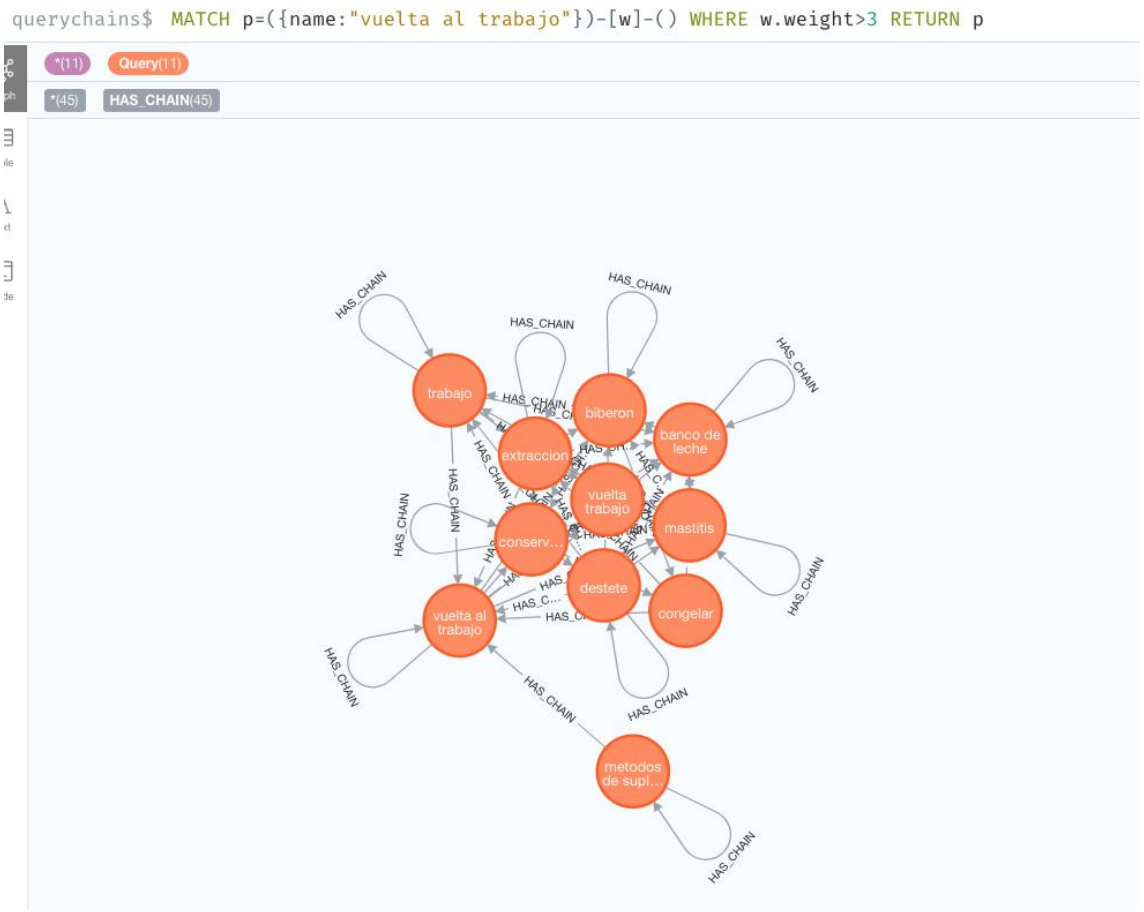


Figura 12 – Visualización del grafo de query chains usando Neo4j

Para explotar el grafo, se ha realizado un análisis usando el algoritmo de modularidad de Louvain, que mide la calidad de los grupos presentes en un grafo comparando la densidad de sus relaciones contra una red de conexiones aleatoria.

Este algoritmo ofrece una clusterización de los nodos que también se ha usado para definir relaciones jerárquicas a los conceptos.

Se obtienen 183 clústeres sobre un grafo con 6.985 nodos.

3.4.4 Relaciones asociativas

En la definición de los SKOS, una relación asociativa permite relacionar conceptos que no están jerárquicamente relacionados, es decir, que no están enlazados vía relaciones del tipo *narrower*/*broader*. En este caso, se usa una relación denominada *related* en el estándar SKOS. Las relaciones asociativas son simétricas, es decir, si el concepto A está relacionado con el concepto B, el concepto B estará relacionado con el concepto A. Un ejemplo de relación asociativa no jerárquica sería la que existe entre el concepto “destete”, que pertenece a la categoría “lactancia” con el medicamento “cabergoline”, de la categoría “salud”, usado para cortar la producción de leche materna. Tanto “cabergoline” está relacionado con “destete” como “destete” con “cabergoline”.

Las *query chains* ofrecen, de nuevo, una manera de identificar relaciones asociativas entre conceptos de la taxonomía. Al contrario que en las relaciones semánticas, se van a seleccionar aquellos pares de conceptos que, perteneciendo a la misma *query chain*, no pertenezcan a la misma categoría. Con esto se garantiza que no existe ninguna relación jerárquica entre ellos.

3.5 Construcción del SKOS

3.5.1 Normalización

Para la construcción de la taxonomía usando los estándares SKOS, lo primero que se ha realizado es una normalización. Se ha realizado una consolidación de todos los conceptos idénticos y agregado las etiquetas *alternate*/*hidden* y *narrower*/*broader*, de manera que se ha obtenido un registro por cada concepto.

3.5.2 Exportación a OWL

Usando Owlready[9], la librería de Python para gestionar taxonomías, ontologías y grafos de conocimiento optimizada para OWL/RDF, se ha trabajado el conjunto de conceptos obtenido para generar una taxonomía que cumple con los estándares SKOS y pueda ser explotada tanto dentro del marco de la web semántica, como por herramientas de software especializado, como PoolParty, Graphologi o Protégé, listadas en el apartado 2.5.

3.5.3 Carga en Protégé

Para la gestión y visualización de la taxonomía, se ha escogido Protégé, software open-source de la Universidad de Stanford, con una gran comunidad que lo soporta y que cumple con todos los estándares W3C. Dispone de versión Web y de escritorio.

La ingesta es sencilla y la navegación entre las distintas categorías y conceptos es muy intuitiva.

[9] https://pypi.org/project/Owlready2/

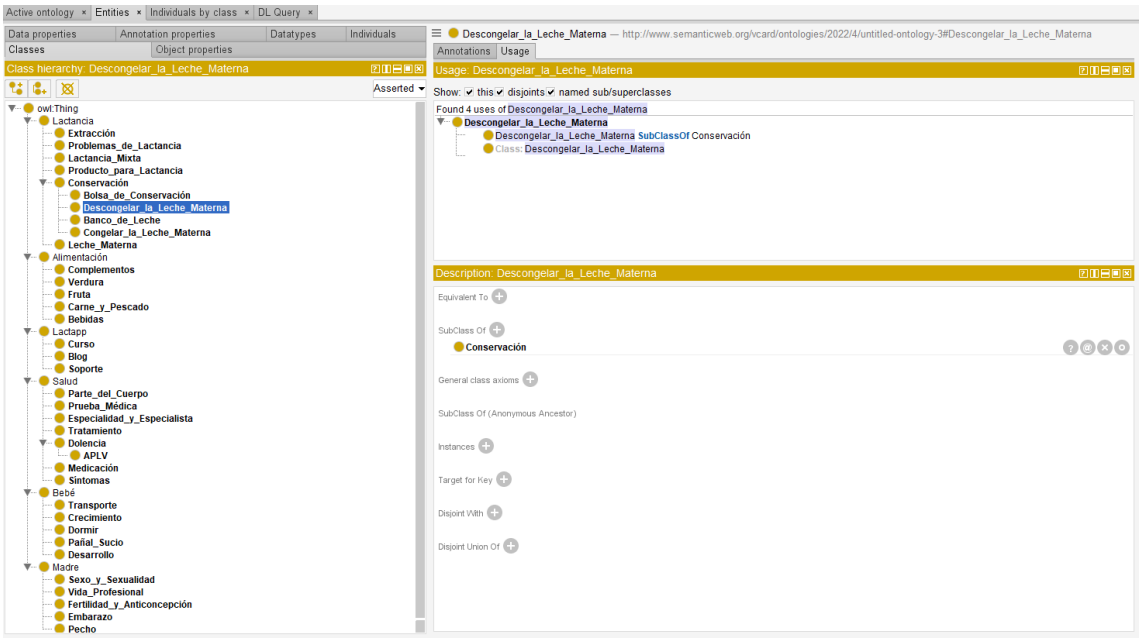


Figura 13 - Taxonomía cargada en el software open-source Protégé

3.5.4 Resultado

La taxonomía desarrollada tiene las siguientes métricas:

Objeto	Número
Conceptos	2566
Sinónimos	3564
Conceptos *broader*	1916
Conceptos *narrower*	147
Conceptos *related*	486
Porcentaje de conceptos en Español	97,6%

Tabla 8 - Métricas de la taxonomía desarrollada

4 Aplicación de la taxonomía

4.1 Uso de la taxonomía junto a un motor de búsqueda

Para explotar la taxonomía en un caso de uso concreto, se han desplegado dos motores de búsqueda de ElasticSearch[10], el primero con parámetros estándar y el segundo usando la opción de sinónimos.

Las búsquedas se realizan sobre un conjunto de artículos del blog y los árboles que contienen las consultas, sus respuestas y una serie de temáticas relevantes. Toda esta información se almacena en ficheros json y csv, que contienen los campos título, palabra clave y texto. No todos los documentos contienen todos estos campos.

Los índices se generan a partir de la información contenida en estos ficheros, en su lenguaje correspondiente (en este caso el español).

Ambos motores de búsqueda comparten una configuración similar: lógica difusa (los conceptos no deben ser exactos, pueden tener ligeras variaciones como en “psatillas” por “pastillas”), *multi match* (la búsqueda se realiza en diversas fuentes de información con ponderaciones diferentes), *boost* (los resultados “exactos” ponderan mejor que los que tienen ligeras variaciones) y otros parámetros secundarios.

Además, el motor de búsqueda con sinónimos recalcula los índices para incluir todos los sinónimos (etiquetas *preferred* y *alternate* de cada concepto de la taxonomía) para enriquecer las búsquedas mediante un procedimiento denominado *query expansion,* que maximiza las posibilidades de encontrar el resultado que mejor se adapte a la búsqueda realizada. Así, una búsqueda que incluya los términos “dar el pecho”, también buscará “lactancia materna” y “mamar”.

Por último, el cálculo de la puntuación de la búsqueda se basa en una ponderación en función de dónde se ha encontrado el concepto buscado. Así, si el concepto se encuentra en la lista de palabras clave pondera con un peso de 5, si aparece en el título, el peso es 3 y si es en el texto, el peso es 2.

4.2 Ejecución

Para comparar los resultados de ambos motores de búsqueda se han escogido dos conjuntos diferentes de cadenas de búsqueda. Por un lado, las 50 búsquedas más frecuentes y, por otro, 50 búsquedas escogidas aleatoriamente entre toda la lista de consultas realizadas por las usuarias. Ambos conjuntos de búsquedas se han lanzado contra cada uno de los motores de búsqueda y se han almacenado los 20 primeros resultados de cada una en un fichero CSV para posterior análisis.

[10] https://www.elastic.co/

Junto con el concepto buscado, se han almacenado el número de resultados obtenidos, los campos en los que se ha habido coincidencia y la puntuación del resultado.

4.3 Comparación de resultados

Los resultados de las búsquedas usando ambos enfoques, con y sin sinónimos, son buenos. El motor de ElasticSearch ya dispone de funcionalidades muy avanzadas, como la búsqueda difusa, búsqueda en múltiples campos u opcionalidad en la coincidencia. Optimizar estos resultados ligeramente supone una gran mejora de prestaciones.

4.3.1 Resultados para las 50 búsquedas más frecuentes

Los resultados recogidos para las 50 búsquedas más frecuentes muestran una mejora apreciable de los resultados en la mayoría de los casos, no sólo en cuanto al número de resultados sino también en cuanto a la puntuación de éstos.

El resumen de los resultados se puede observar en la siguiente tabla:

	CON SINÓNIMOS		SIN SINÓNIMOS	
	# de resultados	Puntuación	# de resultados	Puntuación
Media	349,71 (+89%)	36.220,37 (+31%)	184,67	27.577,26
Mediana	143,50 (+58%)	23.892,84 (-11%)	90	26.640,82

Tabla 9 - Resultados comparativos de la búsqueda con y sin sinónimos para las 50 búsquedas más frecuentes

En las gráficas siguientes se puede apreciar la comparación de número de resultados y puntuación usando sinónimos y sin usarlos.

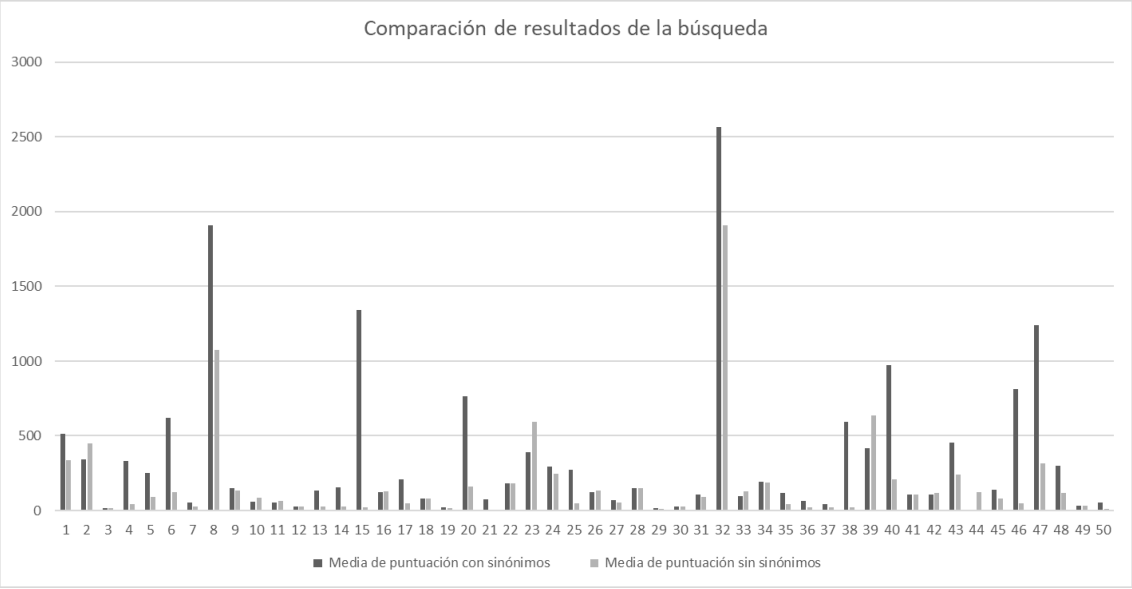


Figura 14 - Comparación de la puntuación de la búsqueda con y sin sinónimos

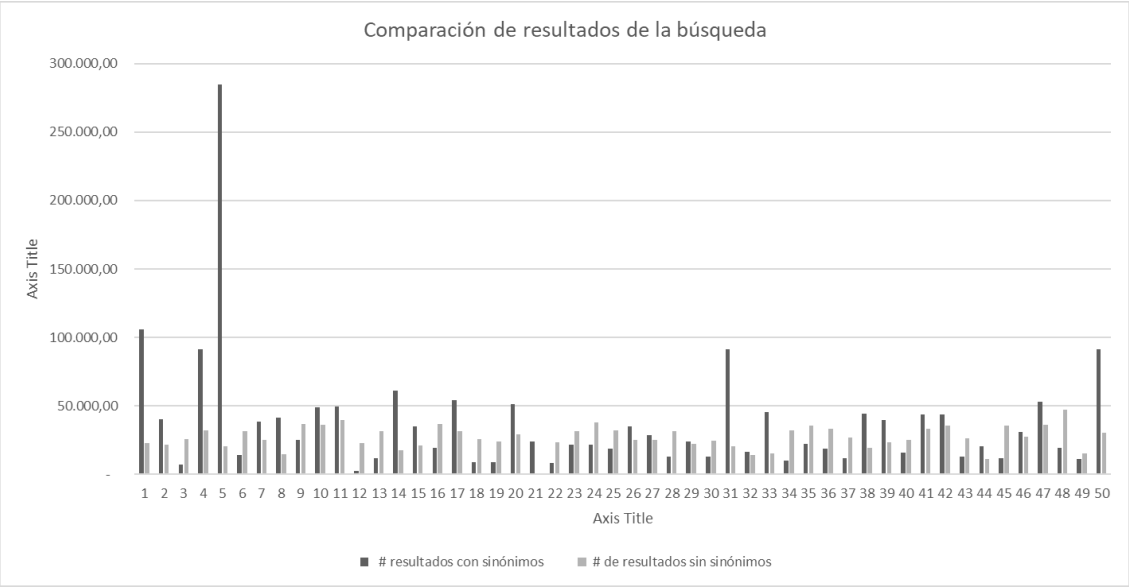


Figura 15 - Comparación del número de resultados de la búsqueda con y sin sinónimos

4.3.2 Resultados para 50 búsquedas aleatorias

Para comparar ambos motores de búsqueda en un contexto más cercano a la realidad de la app, se han analizado los resultados obtenidos usando conceptos escogidos de manera aleatoria de entre todo el histórico de búsquedas para comprobar si añadir sinónimos realmente ayuda en la búsqueda de conceptos menos frecuentes.

Los resultados recogidos para el conjunto aleatorio de cadenas de búsqueda muestran una mejora mucho más apreciable de los resultados en la mayoría de los casos, no sólo en cuanto al número de resultados sino también y, sobre todo, en cuanto a la puntuación de éstos.

El resumen de los resultados se puede observar en la siguiente tabla:

	CON SINÓNIMOS		SIN SINÓNIMOS	
	# de resultados	Puntuación	# de resultados	Puntuación
Media	179,27 (+24%)	18.937,08 (+479%)	144,96	3.269,71
Mediana	37 (+68%)	7.980,02 (+392%)	22	1.620,09

Tabla 10 - Resultados comparativos de la búsqueda con y sin sinónimos para 50 conceptos escogidos aleatoriamente

En las gráficas siguientes se puede apreciar la comparación de número de resultados y puntuación usando sinónimos y sin usarlos. Es relevante observar que un 10% de las consultas sin sinónimos no ofrecen ningún resultado. En cambio, con sinónimos todas las consultas devuelven algún resultado.

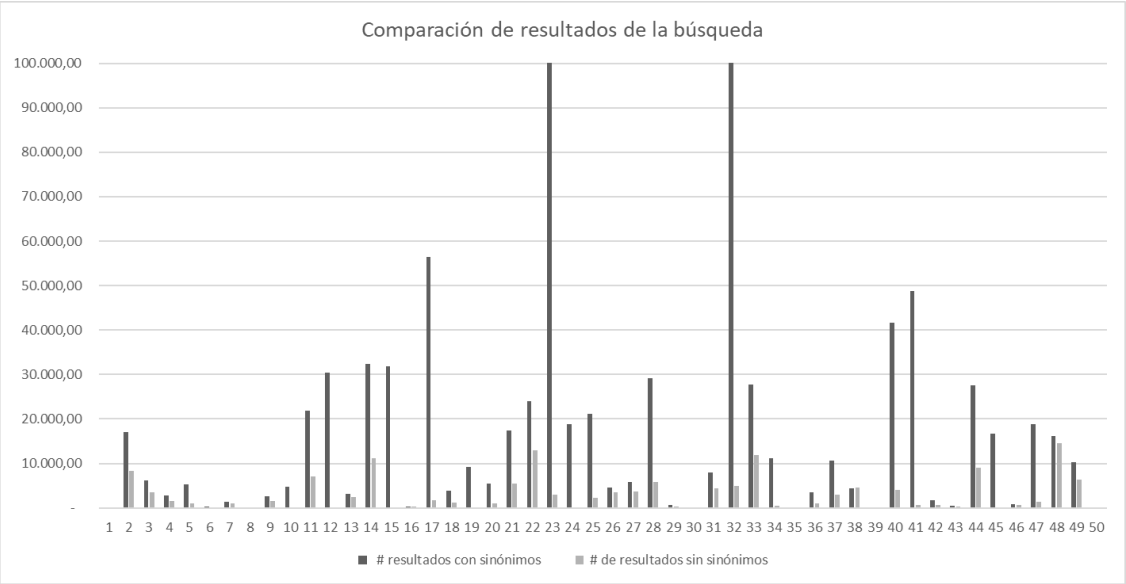


Figura 16 - Comparación de la puntuación de la búsqueda con y sin sinónimos

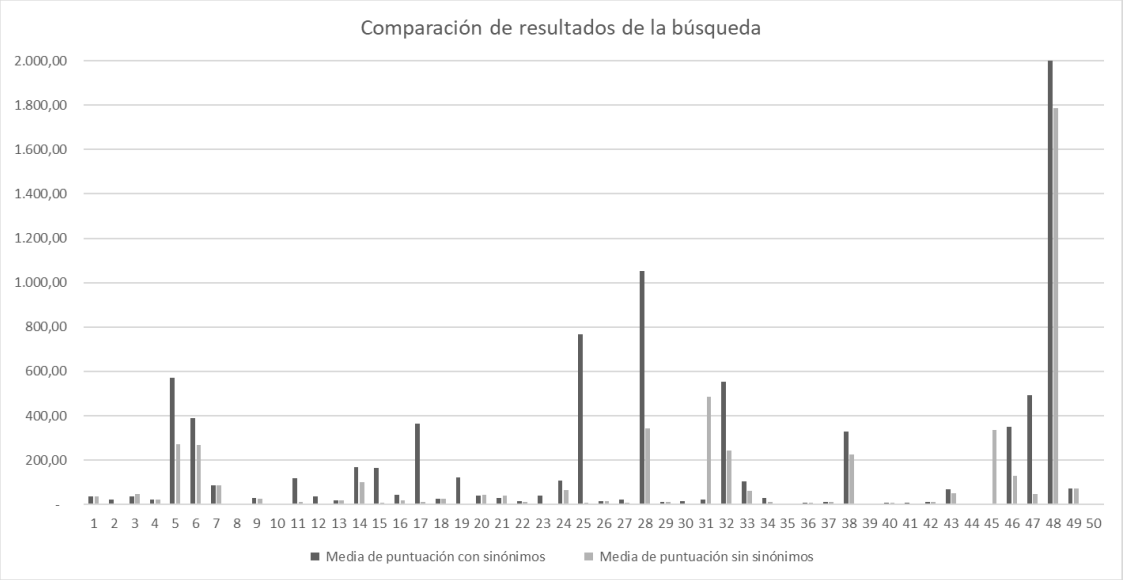


Figura 17 - Comparación del número de resultados de la búsqueda con y sin sinónimos

5 Conclusiones

En este trabajo se ha desarrollado una taxonomía específica en el dominio de la lactancia materna, usando las consultas y las búsquedas de las usuarias de la app líder en el sector.

El desarrollo ha involucrado múltiples técnicas de NLP, modelos pre-entrenados, grafos, algoritmos de *clustering* y un filtrado y proceso de anotación semi-manual para completar el proceso.

El resultado ha sido una taxonomía con más de 2.500 conceptos, más de 3.500 sinónimos y multitud de relaciones semánticas, jerárquicas y asociativas.

Para validar la taxonomía, se ha realizado una prueba de concepto con el buscador de la app, basado en ElasticSearch, usando los sinónimos de la taxonomía para realizar *query-extension* de las búsquedas y validar que los resultados contienen más respuestas y tienen mejor puntuación.

La comparación de ambos motores de búsqueda muestra unos resultados muy prometedores con la opción de añadir sinónimos, con un aumento medio en el número de resultados del 89% y un aumento medio de la puntuación de los resultados del 31% cuando comparamos conceptos frecuentes, y mejoras muy ostensibles, sobre todo en puntuación (+479%) para conceptos no tan frecuentes, en los que los sinónimos ayudan a identificar los conceptos relacionados y mejorar los resultados. Estos resultados invitan a seguir expandiendo y refinando la taxonomía para mejorar más aún estos resultados.

Además, estos resultados invitan a explorar otros campos de aplicación de la taxonomía más allá de las búsquedas apoyadas en sinónimos.

6 Trabajo futuro

6.1 Evolución y enriquecimiento de la taxonomía

La taxonomía desarrollada para Lactapp en el marco de este TFM es una versión reducida y limitada del proyecto más amplio en el que se enmarca.

Se ha construido en base al conjunto de consultas y búsquedas de las usuarias de la app y se ha ampliado con conceptos y relaciones extraídos tras la aplicación de diferentes modelos y técnicas de NLP.

En el *roadmap* de desarrollo de la taxonomía se contemplan una serie de evoluciones y mejoras para dotarla de mayor granularidad y alcance. Entre estas evoluciones se incluyen las siguientes:

1. Uso de nuevas fuentes de datos
    a. Respuestas de las consultoras de lactancia a las consultas de las usuarias
    b. Artículos del blog
    c. Lactapp Medical, la versión de la app para profesionales de la lactancia
    d. Fuentes externas, como Lactapedia[11]
2. Soporte multilenguaje con todos los conceptos y sinónimos traducidos a inglés y catalán en una primera fase, y a otros idiomas en el medio y largo plazo
3. Uso de las extensiones de SKOS-XL para enriquecer la taxonomía con información extraída de los diferentes procesos y modelos, como los pesos de las aristas en las relaciones extraídas del grafo, los orígenes de los datos (Lactapp Medical, Blog, consultas)

6.2 Aplicación a otros casos de uso

Además de la evolución y enriquecimiento de la taxonomía en sí misma, otro ámbito de trabajo futuro es el de la identificación de nuevos casos de uso.

En un análisis inicial, se han identificado diversos usos prácticos para la taxonomía:

1. Autocompletado en el buscador
2. Aplicación al motor de *question-answering* de las consultas
3. Aplicación en un motor de recomendación
4. Refinado de los *keywords* en artículos del blog y árboles de decisión de la app

[11] https://lactapedia.com/lactapedia/all

7 Bibliografía

A. H. Osman and O. M. Barukub, "Graph-Based Text Representation and Matching: A Review of the State of the Art and Future Challenges," in IEEE Access, vol. 8, pp. 87562-87583, 2020, doi: 10.1109/ACCESS.2020.2993191.

Aguilar-Lopez, D., Lopez-Arevalo, I. and Sosa-Sosa, V. (2009) “Uso de ontologías para la mejora de resultados de motores de búsqueda web,” Profesional de la Informacion, 18(1), pp. 34–40. doi:10.3145/EPI.2009.ENE.05.

Anna Rogers, Matt Gardner, and Isabelle Augenstein. 2021. QA Dataset Explosion: A Taxonomy of NLP Resources for Question Answering and Reading Comprehension. 1, 1 (July 2021), 38 pages. arXiv.2107.12708

Bagherifard, K. et al. (2017) “Performance improvement for recommender systems using ontology,” Telematics and Informatics, 34(8), pp. 1772–1792. doi:10.1016/J.TELE.2017.08.008.

Ben Abacha, A., Demner-Fushman, D. A question-entailment approach to question answering. BMC Bioinformatics 20, 511 (2019). https://doi.org/10.1186/s12859-019-3119-4

Berners-Lee, T., Hendler, J. and Lassila, O. (2001) “The semantic web,” Scientific American, 284(5), pp. 34–43. doi:10.1038/SCIENTIFICAMERICAN0501-34.

Biagetti, M.T. (2021) Ontologies as KOSs (IEKO), Knowledge Organization 48, no. 2. Available at: https://www.isko.org/cyclo/ontologies (Accessed: March 7, 2022).

Bin Fu, Yunqi Qiu, Chengguang Tang, Yang Li, Haiyang Yu, Jian Sun: “A Survey on Complex Question Answering over Knowledge Base: Recent Advances and Challenges”, 2020; [http://arxiv.org/abs/2007.13069 arXiv:2007.13069].

Broder, Andrei. (2002). A Taxonomy of Web Search. SIGIR Forum. 36. 3-10. 10.1145/792550.792552.

Cortes, E.G., Woloszyn, V., Barone, D. et al. A systematic review of question answering systems for non-factoid questions. J Intell Inf Syst (2021). https://doi.org/10.1007/s10844-021-00655-8

Fromm, Hansjörg; Wambsganss, Thiemo & Söllner, Matthias: Towards a Taxonomy of Text Mining Features. 2019. - Twenty-Seventh European Conference on Information Systems (ECIS2019). - Stockholm, Sweden.

H. Kahaduwa, D. Pathirana, P. L. Arachchi, V. Dias, S. Ranathunga and U. Kohomban, "Question Answering system for the travel domain," 2017 Moratuwa Engineering

Research Conference (MERCon), 2017, pp. 449-454, doi: 10.1109/MERCon.2017.7980526.

Hjørland, B. (2008) “What is Knowledge Organization (KO)?,” Knowledge Organization, 35(2–3), pp. 86–101. doi:10.5771/0943-7444-2008-2-3-86.

Jurafsky, Martin. (2021). Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition.

LactaResearch Group. LactaPedia [Internet]. Boss M, Hartmann P, editors. Frauenfeld (CH): Family Larsson-Rosenquist Foundation; 2018 [cited yyyy mm dd]. Available from: https://www.lactapedia.com/

Landolt, Severin; Wambsganss, Thiemo & Söllner, Matthias (2021) A Taxonomy for Deep Learning in Natural Language Processing.

Màrquez. (2018). Automatic Question Answering. Problem Solved?. Amazon, Core AI, Barcelona. IberSPEECH 2018. Barcelona, November 23

Medelyan, O. et al. (2013) “Constructing a Focused Taxonomy from a Document Collection,” Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 7882 LNCS, pp. 367–381. doi:10.1007/978-3-642-38288-8_25.

Nickerson, R., Varshney, U. & Muntermann, J. A method for taxonomy development and its application in information systems. Eur J Inf Syst 22, 336–359 (2013). https://doi.org/10.1057/ejis.2012.26

Oyebisi Oyefolahan, I., Femi Aminu, E. and Bashir Abdullahi, M. (2018) “A Review of Ontology-based Information Retrieval Techniques on Generic Domains,” International Journal of Applied Information Systems (IJAIS), 12(13). Available at: www.ijais.org (Accessed: March 12, 2022).

Sharma, Ravi S. and Foo, Schubert and Morales-Arroyo, Miguel, Developing Corporate Taxonomies for Knowledge Auditability - A Framework for Good Practices (August, 26 2008). Journal of Knowledge Organization, Vol. 35, No. 1, 2008, Available at SSRN: https://ssrn.com/abstract=1258162

Strinyuk, S., Scherbakova, I. and Lanin, V. (2021) “Corpus Based Information Extraction Approach for Marine Ontology Development,” in 15th IEEE International Conference on Application of Information and Communication Technologies, AICT 2021. Institute of Electrical and Electronics Engineers Inc. doi:10.1109/AICT52784.2021.9620410.

Szopinski, D., Schoormann, T., & Kundisch, D. (2019): Because Your Taxonomy is Worth It: Towards a Framework for Taxonomy Evaluation. In: Proceedings of the European Conference on Information Systems (ECIS), Stockholm-Uppsala, Sweden.

T. Baltrušaitis, C. Ahuja and L. Morency, "Multimodal Machine Learning: A Survey and Taxonomy," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 41, no. 2, pp. 423-443, 1 Feb. 2019, doi: 10.1109/TPAMI.2018.2798607.

Tarus, J.K., Niu, Z. and Mustafa, G. (2018) “Knowledge-based recommendation: a review of ontology-based recommender systems for e-learning,” Artificial Intelligence Review, 50(1), pp. 21–48. doi:10.1007/S10462-017-9539-5/TABLES/2.

Wanyun Cui, Yanghua Xiao, Haixun Wang, Yangqiu Song, Seung-won Hwang, Wei Wang: “KBQA: Learning Question Answering over QA Corpora and Knowledge Bases”, 2019, Proceedings of the VLDB Endowment, Volume 10 Issue 5, January 2017; arXiv:1903.02419. DOI: 10.14778/3055540.3055549.

Weiguo Zheng, Hong Cheng, Lei Zou, Jeffrey Xu Yu, and Kangfei Zhao. 2017. Natural Language Question/Answering: Let Users Talk With The Knowledge Graph. In Proceedings of the 2017 ACM on Conference on Information and Knowledge Management (CIKM '17). Association for Computing Machinery, New York, NY, USA, 217–226. DOI:https://doi.org/10.1145/3132847.3132977

Xiao Huang, Jingyuan Zhang, Dingcheng Li, and Ping Li. 2019. Knowledge Graph Embedding Based Question Answering. In Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining (WSDM '19). Association for Computing Machinery, New York, NY, USA, 105–113. DOI:https://doi.org/10.1145/3289600.3290956

Ziwei Xu, Mounira Harzallah, Fabrice Guillet, Ryutaro Ichise, Modular Ontology Learning with Topic Modelling over Core Ontology, Procedia Computer Science, Volume 159, 2019, Pages 562-571, ISSN 1877-0509, https://doi.org/10.1016/j.procs.2019.09.211.