POLITÉCNICA

UNIVERSIDAD POLITÉCNICA DE MADRID

**ESCUELA TÉCNICA SUPERIOR DE
INGENIEROS INFORMÁTICOS**

Caracterización de enfermedades basada en su información fenotípica recuperada mediante la extracción de conocimiento biomédico de fuentes de información públicas

Tesis Doctoral

Autor: Gerardo Lagunes García,
MSc en Sistemas Computacionales
Director: Dr. Alejandro Rodríguez González
PhD in Computer Science

2020

UNIVERSIDAD POLITÉCNICA DE MADRID
ESCUELA TÉCNICA SUPERIOR DE
INGENIEROS INFORMÁTICOS

**Caracterización de enfermedades basada en su infor-
mación fenotípica recuperada mediante la extracción
de conocimiento biomédico de fuentes de información
públicas**

Tesis Doctoral

Autor: Gerardo Lagunes García,
MSc en Sistemas Computacionales
Director: Dr. Alejandro Rodríguez González
PhD in Computer Science

2020

Agradecimientos

No es fácil abarcar, en apenas unas líneas, a todas las personas a las que quisiera dar las gracias, de una u otra forma, por apoyarme tan incondicionalmente. Por ello, a todos los que habéis estado a mi lado, de una u otra forma. Gracias.

En primer lugar tengo que dar las gracias a mi padres. Gracias por acompañarme, no solo durante este doctorado, sino a lo largo de toda mi vida. Gracias por confiar en mi aún sabiendo que no siempre tomé las mejores decisiones. Gracias por ser mis figuras a alcanzar, por renunciar a tantas cosas por mi y por mis hermanos. Gracias por vuestro cariño y esfuerzo para que pudiéramos conseguir todo aquello que siempre quisimos, es, sin duda, la mejor de las herencias que nadie puede tener.

Gracias a mis hermanos Josué y Santiago, a pesar de la distancia y los distintos caminos que cada uno de nosotros ha tomado, se que siempre podré contar con vosotros.

Gracias a mis abuelitos por todo el amor que siempre nos habéis transmitido, espero que, desde el mejor sitio en el que podéis estar, estéis orgullosos de mi. Os quiero y extraño.

Gracias a los maestros que tanto me han enseñado a lo largo de toda mi vida, en especial al Maestro Rafael y a la Maestra Beatriz del Instituto Tecnológico de Orizaba, nunca podré agradecerlos todo lo que habéis hecho en mí, aunque vosotros ni siquiera seáis conscientes de ello.

Gracias a CONACYT, sin duda México tiene muchos puntos débiles, pero es extraordinario el apoyo que brindan a todos los que, como yo, quieren formarse a nivel académico en el extranjero. No tienen nada que envidiar a otras instituciones, sean del país que sean. Gracias por todo.

Gracias a mi inagotable director de tesis, Alejandro Rodríguez, por darme la oportunidad de ser tu doctorando. Mi más grande admiración y respeto. Gracias por tu infinita paciencia. Sin duda, no lo habría conseguido sin ti.

Gracias a Massimiliano, por formar parte en cada uno de mis artículos, por mejorar cada uno de mis trabajos, por resolver mis dudas y por ayudarme siempre.

Gracias a Ernestina, por acogerme en su grupo de trabajo y permitirme, junto con Alejandro y Consuelo, formar parte de un equipo de trabajo tan heterogéneo. La diversidad suma.

Gracias también a todos mis compañeros del CTB: Lucía, Eduardo, Luis, Diego, Jhonny, Roberto, Marjan y todos los que habéis compartido conocimiento, tiempo y espacio conmigo en estos cuatro años.

Y como no, gracias a mi familia, a mi esposa y a mis tres hijos, por sacarme de mi comodidad, por hacer siempre lo mejor, por ayudarme y apoyarme en todos mis proyectos, por quererme sin condición. Somos un buen equipo de cinco. Quiero decir, que sin mi esposa este doctorado no hubiera existido, su determinación siempre me ha empujado hacia la mejor dirección. La admiro, la respeto, pero sobre todo la amo. Es la mujer adecuada para mí.

Y por último, dar gracias a Dios, el precursor de mi historia, al que escribe cada renglón de mi vida, al que sabe darme lo que necesito siempre que lo necesito. Sin duda todas las cosas están bien hechas.

Resumen

El conocimiento diagnóstico es el primer elemento básico que se utiliza para descartar o identificar una enfermedad. Y estudiar enfermedades es una tarea compleja pero que es necesaria para mejorar la salud de la población mundial. Y para ello es importante tener información diagnóstica de calidad al alcance.

La pandemia del Coronavirus entre finales de 2019 y principios de 2020 es un claro ejemplo del problema que supone a la humanidad no tener conocimientos sólidos sobre una enfermedad.

El presente trabajo hace uso de diferentes análisis y experimentos para demostrar que la caracterización y definición de las enfermedades en base a sus elementos de conocimiento diagnóstico (DKE, por sus siglas en inglés) permite identificar parejas de enfermedad – enfermedad que pueden ser usadas para mejorar el conocimiento que tenemos de las enfermedades.

En este trabajo se realiza por lo tanto una revisión de la literatura de los trabajos más relevantes relativos a bases de datos biomédicas, enfoques de extracción y generación de conocimiento biomédico, enfoques novedosos de análisis de enfermedades como las redes de enfermedades y enfoques de similitud entre enfermedades. Además de este profundo análisis del estado del arte, se plantearon las hipótesis a verificar, que se han permitido enfocar los esfuerzos en preparar un conjunto de datos médico para caracterizar sus enfermedades usando el tiempo como un aspecto relevante para observar como el contenido médico diagnóstico ha ido cambiando.

El método que ha generado el conocimiento diagnóstico y que ha permitido caracterizar las enfermedades es capaz de ejecutarse periódicamente según defina el usuario para hacer capturas de las fuentes de información relevantes para esta tesis: Wikipedia, PubMed y Mayo Clinic. En el contexto de extracción de datos necesarios para el desarrollo de la tesis, esta periodicidad se estableció en 2 veces por mes para la fuente Wikipedia, pues es la fuente más susceptible de cambios de interés dado que puede ser modificada por cualquier usuario y puede presentar cambios incrementales (añadir contenido en artículos existentes y añadir nuevos artículos), decrementales (borrar contenido de artículos existentes y borrar artículos) y pseudoestacionarios (modificación de contenido existente, sin alterar significativamente en términos de incremento o decremento), mientras que la fuente de la clínica Mayo depende de la propia clínica y es altamente estática, y PubMed solo representaría cambios incrementales. Después, a través de herramientas de NLP MetaMap, que es ejecutada sobre los textos se obtiene la lista de conceptos médicos que posteriormente se procesará para obtener una mayor calidad de los datos. El resultado es un conjunto de datos médicos (enfermedad – DKE), por el cual se

ha desarrollado también un proceso de validación para verificar la precisión con la que el sistema obtiene los DKE y cuyo resultado se considera adecuado.

En esta tesis también presenta el sistema web que es capaz de integrar información de diversas fuentes médicas que permitirá aumentar el grueso de datos del conjunto postulado en este documento y así compartir cada dato generado para poder crear redes de enfermedades diversas y por consiguiente abordar el complejo estudio las relaciones entre enfermedades.

En esta tesis también postula una serie de experimentos y análisis que han permitido caracterizar y definir las enfermedades del conjunto de datos médicos ya validado. En primera lugar, se desarrollo un análisis descriptivo que ha servido para observar y comprender como esta compuesto el nuevo conocimiento generado: cantidad de enfermedades, cantidad de DKE, textos recuperados, códigos recuperado, las fuentes de conocimiento de donde provienen los códigos, entre otros elementos. Esto ha permitido definir las enfermedades según los DKE obtenido, así como también mostrar las distribuciones de los DKE según diferentes aspectos como, secciones, enfermedades, fuentes, *snapshots* y otros más.

Se desarrollo un análisis evolutivo sobre Wikipedia (fuente, que habilitada para este tipo de estudio) que ha permitido observar el estado del conocimiento a lo largo de dos años de capturar conocimiento. La observación sobre la evolución sugiere que tanto los artículos de enfermedad de Wikipedia como su contenido aumenta con el tiempo y se va mejorando. Así se ha observado como el tiempo ha caracterizado y definido las enfermedades en Wikipedia. En mismo contexto de la evolución se presenta otro análisis evolutivo pero ahora enfocado en como ha ido cambiando el contenido de los artículos de enfermedad. Esta evolución del contenido ha caracterizado las enfermedades mediante la observación del comportamiento del conocimiento médico en Wikipedia según, el tamaño total de caracteres del documento de enfermedad, la cantidad de referencias científicas PubMed y la cantidad de DKE. Algunos hallazgos muestran una relación fuerte entre la cantidad de texto y la de DKE, pero no así con las referencias.

Finalmente, se desarrollaron los cálculos de tres diferentes coeficientes de similitud con los que se han caracterizado las enfermedades en base a los DKE en común. Este análisis muestra el potencial del conjunto de datos para encontrar parejas enfermedad – enfermedad. Relaciones altamente interesantes porque sugieren que las enfermedades esas enfermedades pueden compartir no solo DKE sino genes, o alguna otra característica biológica. Este el primer paso del proyecto DISNET para conseguir apoyar el reposicionamiento de fármacos.

Abstract

Diagnostic knowledge is the first basic element that is used to discard or identify a disease. And studying diseases is a complex task but one that is necessary to improve the health of the world's population. And to do so, it is important to have quality diagnostic information at hand.

The Coronavirus pandemic between the end of 2019 and the beginning of 2020 is a clear example of the problem that humanity faces when there is no solid knowledge about a disease.

The present work makes use of different analyses and experiments to demonstrate that the characterisation and definition of diseases based on their diagnostic knowledge elements (DKE) allows us to identify disease - disease pairs that can be used to improve our knowledge of diseases.

In this work, therefore, a literature review of the most relevant works related to biomedical databases, approaches to the extraction and generation of biomedical knowledge, novel approaches to disease analysis such as disease networks and disease similarity approaches is carried out. In addition to this in-depth analysis of the state of the art, the hypotheses to be verified were raised, which have allowed to focus the efforts in preparing a set of medical data to have a characterisation of their diseases using time as a relevant aspect to observe how the medical diagnostic content has been changing.

The method that has generated the diagnostic knowledge and that has allowed to characterize the diseases is able to be executed periodically as defined by the user to make captures of the relevant information sources for this thesis: Wikipedia, PubMed and Mayo Clinic. In the context of extracting data necessary for the development of the thesis, this periodicity was established at twice a month for the source Wikipedia, as it is the source most susceptible to changes of interest given that it can be modified by any user and can present incremental changes (adding content in existing articles and adding new articles), The Mayo Clinic's source is highly static and PubMed would only represent incremental changes. Then, through MetaMap NLP tools, which is executed on the texts, the list of medical concepts is obtained which will later be processed to obtain a higher quality of data. The result is a set of medical data (disease - DKE), for which a validation process has also been developed to verify the accuracy with which the system obtains the DKE and whose result is considered adequate.

This thesis also presents the web system that is capable of integrating information from various medical sources that will increase the bulk of data from the set postulated in this document and thus share each data generated in order to

create networks of various diseases and therefore address the complex study relationships between diseases.

This thesis also postulates a series of experiments and analyses that have made it possible to characterise and define the diseases in the already validated medical data set. Firstly, a descriptive analysis was developed that has served to observe and understand how the new knowledge generated is composed: number of diseases, number of DKE, texts recovered, codes recovered, the sources of knowledge from which the codes come, among other elements. This has allowed to define the diseases according to the DKE obtained, as well as to show the distributions of the DKE according to different aspects like, sections, diseases, sources, snapshots and others.

An evolutionary analysis was developed on Wikipedia (source, which is enabled for this type of study) that has allowed us to observe the state of knowledge over two years of capturing knowledge. The observation on evolution suggests that both the Wikipedia disease articles and their content increase over time and are being improved. This is how time has characterised and defined diseases in Wikipedia. In the same context of evolution another evolutionary analysis is presented but now focused on how the content of the disease articles has changed. This evolution of content has characterised diseases by observing the behaviour of medical knowledge in Wikipedia according to, the total character size of the disease document, the amount of PubMed scientific references and the amount of DKE. Some findings show a strong relationship between the amount of text and the amount of DKE, but not with the references.

Finally, calculations were developed for three different similarity coefficients with which diseases have been characterised based on common DKE. This analysis shows the potential of the data set to find disease - disease pairs. These relationships are highly interesting because they suggest that diseases can share not only DKE but genes, or some other biological characteristic. This is the first step of the DISNET project to get support for drug repositioning.

Índice

CAPÍTULO 1 INTRODUCCIÓN.....	1
1.1 PREÁMBULO	1
1.2 INTRODUCCIÓN Y MOTIVACIÓN.....	2
1.3 HIPÓTESIS DE INVESTIGACIÓN.....	7
1.4 OBJETIVOS	8
1.5 ESTRUCTURA DEL DOCUMENTO	9
CAPÍTULO 2 ESTADO DEL ARTE.....	11
2.1 CONCEPTOS MÉDICOS	11
2.1.1 <i>Enfermedad</i>	11
2.1.2 <i>Síntomas y signos</i>	11
2.1.3 <i>Elemento(s) de diagnóstico</i>	12
2.2 REPOSITORIOS DE CONOCIMIENTO BIOMÉDICO.....	12
2.3 ENFOQUES DE EXTRACCIÓN Y GENERACIÓN DE CONOCIMIENTO BIOMÉDICO	21
2.4 REDES DE ENFERMEDADES.....	25
2.5 ENFOQUES DE SIMILITUD ENTRE ENFERMEDADES	34
2.6 CONCLUSIONES	37
CAPÍTULO 3 METODOLOGÍA	40
3.1 FUENTES DE INFORMACIÓN.....	41
3.1.1 <i>Wikipedia</i>	43
3.1.1.1 <i>Anatomía de un artículo de enfermedad</i>	44
3.1.1.2 <i>Enfoque de extracción</i>	48
3.1.2 <i>PubMed</i>	52
3.1.2.1 <i>Enfoque de extracción</i>	53
3.1.3 <i>MayoClinic</i>	55
3.1.3.1 <i>Enfoque de extracción</i>	56
3.2 ENFOQUE PRINCIPAL DE EXTRACCIÓN Y GENERACIÓN DE CONOCIMIENTO ...	59
3.2.1 <i>Flujo de trabajo</i>	59
3.2.1 <i>Compartición de la información</i>	63
3.2.2 <i>Conjunto de datos</i>	63
3.2.3 <i>Componentes técnicos</i>	67
3.2.4 <i>MetaMap</i>	72
3.2.4.1 <i>Configuración</i>	73
3.2.5 <i>Proyecto DISNET</i>	76
3.3 ENFOQUE DE EXTRACCIÓN DE INFORMACIÓN HISTÓRICA DE LOS ARTÍCULOS SOBRE ENFERMEDADES EN WIKIPEDIA	77
3.3.1 JUSTIFICACIÓN	77
3.3.2 PROCESO PARA RECUPERAR LA INFORMACIÓN HISTÓRICA	79
3.4 CONCLUSIONES	86
CAPÍTULO 4 RESULTADOS.....	89
4.1 CONSTRUCCIÓN DEL CONJUNTO DE DATOS	90
4.1.1 SNAPSHOT	90
4.1.2 ENFERMEDAD	90
4.1.3 LISTA DE ENFERMEDADES.....	92

4.2	VISTA GLOBAL	93
4.3	VALIDACIÓN DEL SISTEMA	95
4.3.1	DISEÑO DE LA EVALUACIÓN	96
4.4	ANÁLISIS DESCRIPTIVO DEL CONJUNTO DE DATOS MÉDICOS.....	103
4.4.1	EVOLUCIÓN DE LA INFORMACIÓN.....	107
4.5	ANÁLISIS DESCRIPTIVO DE IDENTIFICADORES DE ENFERMEDADES.....	114
4.6	ANÁLISIS DE LA EVOLUCIÓN HISTÓRICA DEL CONTENIDO DE ENFERMEDADES WIKIPEDIA	125
4.6.1	DISEÑO DEL ANÁLISIS.....	126
4.6.2	DESARROLLO DEL ANÁLISIS.....	128
4.7	SIMILITUD EN BASE A CARACTERÍSTICAS FENOTÍPICAS	136
4.8	CONCLUSIONES	145
CAPÍTULO 5 VERIFICACIÓN DE HIPÓTESIS.....		148
CAPÍTULO 6 CONCLUSIONES Y TRABAJOS FUTUROS		151
6.1	CONCLUSIONES	151
6.2	LÍNEAS FUTURAS.....	154
CAPÍTULO 7 BIBLIOGRAFÍA.....		155
PUBLICACIONES		170
7.1	PUBLICACIONES EN CONGRESOS.....	170
7.2	PUBLICACIONES EN REVISTAS DE ALTO IMPACTO.....	170

Índice de figuras

Figura 1. Nuevos registros de PubMed y PubMed Central (PMC) agregados por año (por 10,000s). Fuente: Exploring PubMed as a reliable resource for scholarly communications services (Williamson and Minter, 2019).....	5
Figura 2. Índice de contenido del documento, capturado a mediados de julio de 2019 del artículo Wikipedia referente a la “ <i>Gorham's disease</i> ”.....	46
Figura 3. a) “ <i>Infobox</i> ” ubicado al inicio del documento, capturado a finales de febrero de 2018 del artículo Wikipedia de la “ <i>Ebstein's anomaly</i> ”; b) <i>infobox</i> al final del documento, capturado a mediados de julio de 2019 del artículo Wikipedia de la “ <i>Influenza</i> ”.....	46
Figura 4. Contenido reestructurado de la casilla informativa ubicada al principio del documento Wikipedia referente al “ <i>Lung cancer</i> ”. Captura realizada a principios de marzo de 2020.	47
Figura 5. Consulta SPARQL para obtener la lista de enfermedades de Wikipedia mediante SPARQL.....	49
Figura 6. Textos médicos en secciones relevantes. Capturada la imagen de fondo a mediados de julio de 2019 del artículo Wikipedia de la “ <i>Acrodynia</i> ” y la imagen de frente capturada a principios de marzo de 2020 del artículo Wikipedia del “ <i>Lung cancer</i> ”.	51
Figura 7. Flujo de trabajo del proceso de extracción de conocimiento de Wikipedia....	52
Figura 8. Flujo de trabajo del enfoque propuesto de extracción de información de PubMed.	55
Figura 9. Lista principal de enfermedades en Mayo Clinic. Ubicada en su página web.	57
Figura 10. Lista de la categoría <i>A</i> relativa a todas las enfermedades cuyo nombre inicia con la primera letra del abecedario inglés.	58
Figura 11. Contenido del artículo Mayo Clinic relativo a la enfermedad “ <i>Acoustic neuroma</i> ”. Secciones relevantes: “ <i>Symptoms & causes</i> ” y “ <i>Diagnosis & treatment</i> ”.	58
Figura 12. Flujo de trabajo del proceso de extracción de conocimiento de Mayo Clinic.	59
Figura 13. Arquitectura y flujo de trabajo del enfoque de extracción y generación de conocimiento diagnóstico.....	60
Figura 14. Modelo entidad – relación del conjunto de datos.....	66
Figura 15. Estructura del sistema Web encargada de ejecutar el proceso de extracción y generación de conocimiento médico.	68
Figura 16. Comparación de dos revisiones del artículo de enfermedad Wikipedia sobre <i>Otic polyp</i> . a) Revisión del artículo que corresponde con la captura de conocimiento (<i>snapshot</i>) de Wikipedia del primero de febrero de 2019 contenida en el conjunto de datos propio; b) Revisión del artículo que corresponde con la captura de conocimiento (<i>snapshot</i>) de Wikipedia del quince de febrero de 2019 contenida en el conjunto de datos propio.....	78
Figura 17. Identificación gráfica de la limitación del contenido del conjunto de datos de cara al análisis de la evolución propuesto en esta sección. Se muestra un incremento anormal del número de caracteres de los textos de las secciones relevantes en el artículo Wikipedia de la enfermedad <i>Otic polyp</i> , entre las capturas de conocimiento (“ <i>snapshot</i> ”) del 1 de febrero de 2019 y del 15 de febrero de 2019.	79

Figura 18. Seguimiento para ver una revisión del artículo de Wikipedia sobre la enfermedad <i>Otic polyp</i> . a) Versión actual de este artículo Wikipedia, a septiembre de 2019; b) historial de revisiones del artículo en septiembre de 2019; y c) contenido de una de sus revisiones específicas (la del 10 de enero de 2019).....	81
Figura 19. Respuesta JSON de la consulta WQone usando como elementos de búsqueda: <i>titles = Otic polypy rvstart = 2019 - 01 - 15</i>	84
Figura 20. Respuesta JSON de la consulta WQtwo usando como elemento de búsqueda: <i>oldid = 877794912</i>	85
Figura 21. Flujo de trabajo del proceso de extracción de información histórica de artículos Wikipedia.....	86
Figura 22. Entidades del conjunto de datos médicos que permiten determinar enfermedades DISNET. Extracto de relaciones del DMDS (<i>DISNET Medical Dataset</i>).	91
Figura 23. Consulta SQL que obtiene la lista de enfermedades de la fuente PubMed. .	92
Figura 24. Comparativa de las métricas de validación tanto en Wikipedia como en PubMed.	100
Figura 25. Distribución de los DKE agrupados por las secciones relevantes (<i>snapshot</i> : 1 de marzo de 2020).	107
Figura 26. Comparación entre la primera <i>snapshot</i> (1 de febrero de 2018) y la última (1 de marzo de 2020).	110
Figura 27. Número de códigos de enfermedades recuperados según cada vocabulario biomédico registrado en el conjunto de datos (considerando todas las <i>snapshots</i> de todas las fuentes).....	115
Figura 28. Códigos de enfermedades recuperados mediante el proceso de extracción de conocimiento de PubMed.	116
Figura 29. Códigos de enfermedades recuperadas mediante el proceso de extracción de conocimiento de Wikipedia.....	117
Figura 30. Comparación entre el número de códigos de los vocabularios obtenidos por la extracción de conocimiento de PubMed y Wikipedia.	117
Figura 31. a) Se muestran todas las enfermedades que tienen al menos un identificador (5.868) dentro del conjunto de datos en contraste con el total de enfermedades, considerando en ambos todas las <i>snapshots</i> . b) Enfermedades con al menos un identificador en Wikipedia y en PubMed, considerando todas las <i>snapshots</i> para ambas fuentes.....	118
Figura 32. Comparación del número de enfermedades con al menos un identificador asociado en la primera y última <i>snapshot</i> de Wikipedia.....	119
Figura 33. Enfermedades según los vocabularios con mayor número de códigos, considerando todas las <i>snapshots</i>	120
Figura 34. Mapeo por nombre entre las enfermedades en Wikipedia y Mayo Clinic. ..	121
Figura 35. Frecuencia de mapeo de códigos entre DISNET y DisGeNET y provenientes de cada vocabulario.....	121
Figura 36. Frecuencia de los códigos procedentes de cada vocabulario en DisGeNET en su versión 6.	122
Figura 37. Distribuciones de (DISNET id – CUI) y (CUI - DISNET id) de acuerdo a los mapeos hechos con DisGeNET.	122
Figura 38. Evolución de las <i>snapshots</i> de Wikipedia. (Izquierda) Distribución de la probabilidad de la variación del número de caracteres (línea negra), DKE (línea roja) y referencias (línea azul) en dos <i>snapshots</i> consecutivas, expresada en	

porcentaje. (Derecha) Distribución de probabilidad acumulada de las mismas tres métricas.....	128
Figura 39. Evolución del número de caracteres, DKE y referencias, a lo largo de todas las <i>snapshots</i> . Cada parte de color representa un decil de la distribución (véase la leyenda de la derecha). Todos los valores están normalizados de tal manera que el valor inicial es cero.....	130
Figura 40. Relaciones entre las métricas consideradas. Los cuatro cuadros informan de los gráficos de densidad que representan las relaciones entre el número de términos, caracteres, referencias y referencias PMID para un mismo artículo de Wikipedia. Las sombras más oscuras indican un mayor número de snapshots. La línea punteada azul indica el mejor ajuste lineal, junto con la pendiente y r^2 correspondientes. La línea punteada roja del cuarto panel corresponde a la diagonal principal, es decir, un número de referencias del PMID igual al número de referencias.	131
Figura 41. Relación entre el tamaño de cada artículo y el número de veces que se ha modificado. Todos los artículos se han dividido en diez grupos, que corresponden a los diez deciles según el número de caracteres - la mediana de cada grupo se indica en el eje X. Cada cuadro representa la distribución del número de cambios para ese decil.	132
Figura 42. Comparación entre la evolución de los artículos sobre enfermedades en Wikipedia con una cantidad alta y baja de contenido textual. a),b).....	133
Figura 43. Relación entre el tamaño de cada artículo y el número de autores que han modificado ese artículo. Como se ha hecho en la Figura 41, los artículos se han dividido en diez deciles según el número de caracteres - la mediana de cada grupo se indica en el eje X. Cada recuadro representa la distribución del número de autores para ese decil.....	134
Figura 44. Evolución a través del tiempo del número de caracteres de los 12 artículos de enfermedad con mayor variabilidad.....	135
Figura 45. Distribución de los coeficientes de similitud Coseno, Jaccard, Dice. La figura ha sido creada considerando la información de la <i>snapshot</i> de Wikipedia del 1 de marzo de 2020.	139
Figura 46. Grafo de enfermedades relacionadas según el alto número de DKE que tienen en común. El rango de DKE en común es entre 20 y 42 elementos. El grosor de las líneas de relación depende del número de DKE así como el color del mismo. El grafo se desarrolló usando información de la <i>snapshot</i> de Wikipedia del 1 de marzo de 2020.	142
Figura 47. Extracto de la Figura 46 en el que se aprecia la única relación (“ <i>Head injury</i> ”– “ <i>Brain damage</i> ”) que se hace con más de 40 DKE. También se aprecia una de las enfermedades más relacionadas, la “ <i>Dementia</i> ”.	142
Figura 48. Distribución de los coeficientes de similitud Coseno, Jaccard, Dice. La figura ha sido creada considerando la información de la <i>snapshot</i> de PubMed del 3 de abril de 2018.....	143
Figura 49. Distribución de los coeficientes de similitud Coseno, Jaccard, Dice. La figura ha sido creada considerando la información de la <i>snapshot</i> de Mayo Clinic del 1 de marzo de 2020.	143
Figura 50. Distribución de los coeficientes de similitud Coseno, Jaccard, Dice considerando solo los DKE pertenecientes a los <i>semantic types</i> : “ <i>sosy</i> ” y “ <i>fndg</i> ”. La figura ha sido creada considerando la información de la <i>snapshot</i> de Wikipedia del 1 de marzo de 2020.....	144

Índice de tablas

Tabla 1. Lista de <i>semantic types</i> relevantes para esta tesis.	75
Tabla 2. Números de enfermedades DISNET, identificadores (códigos de enfermedades), términos médicos y vocabularios.....	93
Tabla 3. Hoja de validación para el artículo Wikipedia referente a la enfermedad " <i>Acute decompensated heart failure</i> ". Información recuperada el 1 de febrero de 2018. Abreviaciones de columnas: WIKIPEDIA (W), METAMAP (M), TVP (T), RELEVANT (R).....	96
Tabla 4. Valores totales de la <i>snapshot</i> del 1 de febrero de 2018 de Wikipedia y de la <i>snapshot</i> del 3 de abril de 2018 de PubMed.....	100
Tabla 5. Valores de los parámetros clave que conforman la <i>snapshot</i> del 1 de marzo de 2020.	104
Tabla 6. Artículos Wikipedia de enfermedad con mayor y menor cantidad DKE en enfermedades (<i>snapshot</i> : 1 de marzo de 2020).....	105
Tabla 7. Frecuencia mayor y menor de DKE (total de apariciones del concepto; total de artículos de enfermedad en los que aparece el concepto) (<i>snapshot</i> : 1 de marzo de 2020).....	105
Tabla 8. Distribución de los DKE agrupados según los <i>semantic types</i> (<i>snapshot</i> : 1 de marzo de 2020).....	106
Tabla 9. Evolución de la información contenida en cada una de las <i>snapshots</i> en Wikipedia, del 1 de febrero de 2018 al 1 de marzo de 2020.....	108
Tabla 10. Comparación entre todas las <i>snapshots</i> de enfermedades en Wikipedia. Considerando las <i>snapshots</i> que comprenden del 1 de febrero de 2018 al 1 de marzo de 2020.....	111
Tabla 11. Comparación entre todas las <i>snapshots</i> de enfermedades en Mayo Clinic. Considerando las <i>snapshots</i> que comprenden del 15 de agosto de 2018 al 1 de marzo de 2020.....	113
Tabla 12. Enfermedades con mayor y menor número de códigos en PubMed y Wikipedia.	119
Tabla 13. Mapeo del CUI C0023794 en DisGeNet con identificadores DISNET.	123
Tabla 14. Mapeos del identificador DISNET DIS006978 con los CUI de DisGeNet....	124
Tabla 15. Artículos de enfermedad con el mayor y el menor promedio de caracteres..	131
Tabla 16. Número de DKE de los artículos sobre enfermedades con mayor contenido textual (número de caracteres).....	136
Tabla 17. Artículos de enfermedades con el mayor número de DKE.	136
Tabla 18. Ejemplo del resultado de consultar los elementos diagnósticos de cada enfermedad. Los datos presentados en esta tabla corresponden a la <i>snapshot</i> de Wikipedia del 1 de marzo de 2020.	137
Tabla 19. Ejemplo del conjunto de datos resultantes de sumar el número de DKE para cada una de las enfermedades. Los datos presentados en esta tabla corresponden a la <i>snapshot</i> de Wikipedia del 1 de marzo de 2020.	138
Tabla 20. Ejemplo del conjunto de datos resultante de realizar todas las posibles asociaciones enfermedad–enfermedad. Los datos presentados en esta tabla corresponden a la <i>snapshot</i> de Wikipedia del 1 de marzo de 2020.....	138
Tabla 21. Ejemplo del conjunto de datos resultante tras calcular tres índices de similitud para las relaciones enfermedad–enfermedad. Los datos presentados en esta tabla corresponden a la <i>snapshot</i> de Wikipedia del 1 de marzo de 2020.....	138

Tabla 22. Medias los valores de los coeficientes de similitud calculados. Considerando la <i>snapshot</i> de Wikipedia del 1 de marzo de 2020.....	140
Tabla 23. Información detallada de las once relaciones enfermedad–enfermedad cuyos coeficientes de similitud Coseno son muy altos. Considerando la <i>snapshot</i> de Wikipedia del 1 de marzo de 2020.....	140

Capítulo 1 Introducción

1.1 Preámbulo

“Future nations will know by history only that the loathsome small-pox has existed and by you has been extirpated”

Thomas Jefferson

En pleno siglo XVIII la viruela era una enfermedad que asediaba a la población mundial; la padecía el 60% de los seres humanos de los cuales casi un 30% de los adultos y un 50% de los niños fallecía.

Se propagaba con rapidez debido al gran aumento demográfico y por la facilidad de su transmisión, pues no solo se contagiaba por contacto directo con los enfermos, sino también a través de los objetos contaminados.

Esta enfermedad producía fiebre, malestar, vómitos y fuertes dolores, pero lo que verdaderamente distinguía a la viruela del resto de enfermedades eran las ampollas y pústulas que esta producía, primero en la lengua y manos, después en el resto del cuerpo.

La gran mayoría de los especialistas asegura que la viruela mató a más personas por sí sola que todas las demás enfermedades infecciosas juntas.

Fue en 1795, aproximadamente, cuando un joven médico rural, Edward Jenner (E.J.), amante de la observación, pudo darse cuenta de tres cosas: La primera de ellas era que las vacas padecían y transmitían una clase diferente de viruela (viruela bovina) que llenaba sus ubres de pústulas y ampollas muy similares a las de la viruela y que contagiaban a las personas que tenían contacto directo con ellas (en especial, a las ordeñadoras de vacas); La segunda, que dicha viruela era muy leve, produciendo una rápida sanación tanto para los animales como para las personas que se contagiaban; pero lo que realmente importaba era la tercera: las personas infectadas de viruela bovina, jamás enfermaban del virus de la viruela humana.

Edward Jenner fue más allá, en 1796 inoculó pus de las yagas de una ordeñadora infectada de viruela bovina a un niño de 8 años. Pudo comprobar que le produjo malestar general, algo de fiebre y ampollas y pústulas en el lugar de la

1.2 Introducción y motivación

inoculación, pero en menos de una semana el niño se encontraba plenamente sano.

Unos días después de su curación, Jenner inoculó en el mismo niño, mediante dos cortes transversales, el virus de la viruela humana, resultando en que aquel niño no desarrolló la enfermedad.

Gracias a la capacidad de observación de Edward Jenner se dio el primer paso de un largo proceso en el que la viruela, el azote de la humanidad, sería totalmente erradicada (Riedel, 2005).

¿Qué hubiera pasado con la raza humana si E. Jenner no hubiera estudiado la relación entre estas dos enfermedades (viruela bovina y viruela humana)?

1.2 Introducción y motivación

La complejidad que conlleva el estudio de la gran cantidad de enfermedades de las que se tienen conocimiento y que aquejan al ser humano es un hecho; son muchos los factores que intervienen en cada una de ellas, desde factores genéticos, entendiendo por tales aquellos que acompañan al ser humano desde el mismo momento de su concepción; hasta biológicos, entendiendo por tales aquellos que tienen que ver con cambios hormonales, ambientales; e incluso físicos, tales como la complexión física del individuo, la edad, su sexo, el lugar geográfico en el que se encuentra, etc.

Todos estos factores influyen de forma decisiva en el desarrollo de una misma enfermedad, de forma que se desarrollará de una forma diferente y se exteriorizará de forma distinta dependiendo del sujeto concreto en el que ésta se produzca.

Son importantes, también, otros factores tales como el económico, que afecta directamente a la calidad de los alimentos consumidos o el acceso a los diferentes tipos de tratamientos; o los aspectos sociales, que se reflejan en la forma en la que nos relacionamos unos con otros.

El conjunto de todos estos factores, de los que apenas se han nombrado algunos, influyen en el comportamiento fenotípico de las enfermedades, y es sumamente importante su estudio para la mejora de la calidad de vida de las personas.

Las manifestaciones fenotípicas, es decir, las distintas maneras en las que se manifiesta una enfermedad, son uno de los primeros elementos a tener en cuenta a la hora de diagnosticar, categorizar y tratar enfermedades, permitiéndonos también observar las posibles formas de interacción entre las diferentes enfermedades y encontrar elementos comunes entre las mismas.

1.2 Introducción y motivación

Estas manifestaciones fenotípicas, elemento clave en el desarrollo de la presente tesis, tienen una importancia primaria en varios campos de investigación, siendo especialmente relevantes en tareas como la creación de bases de conocimiento fenotípico de enfermedades (Xu, Li & Wang, 2013) o de sistemas de diagnóstico automatizados (Ku, Li & Chang, 2013; Rodríguez-González & Alor-Hernández, 2013).

Así, con el estudio de las manifestaciones fenotípicas, se proporcionan medios que permiten mejorar el conocimiento médico existente sobre las diversas dolencias que aqueja el cuerpo humano.

La RAE define *salud* como el “*estado en que el ser orgánico ejerce normalmente todas sus funciones*”, por lo tanto, cuando hablamos de mejora de la salud, la OMS engloba, no solamente la curación de afecciones o enfermedades, sino todos los elementos que afectan a este buen funcionamiento del ser humano incluyendo, por lo tanto, el bienestar mental y social de la población (WHO, 1946).

Por ello la RAE no define curar como la desaparición de los síntomas de una enfermedad (signos visibles), sino que de las diversas acepciones que la RAE recoge, se entiende por curar como algo más profundo, algo que va más allá que eliminar una enfermedad de un cuerpo, estableciendo que se entiende por curar:

tr. Hacer que un enfermo o lesionado, o una parte de su cuerpo enferma o dañada recupere la salud.

tr. Hacer que una lesión, dolencia, herida o enfermedad remita o desaparezca.

tr. Hacer que una persona que sufre anímicamente se recobre, o hacer que ese sufrimiento desaparezca.

tr. Remediar un mal.

Recuperar la salud supone, por lo tanto, la eliminación de todos y cada uno de los síntomas y secuelas que una enfermedad produce (Pizzorno, 2016; OXFORD, 2019; RAE, 2019).

Así, una persona que supera un cáncer no se cura cuando ese cáncer desaparece, cuando sus marcadores tumorales descienden, cuando la persona es dada de alta; el proceso de curación es, en muchas ocasiones, mucho más largo.

La curación se produce cuando ese ser humano puede volver a llevar la vida que llevaba antes, cuando psicológicamente supera esa enfermedad, cuando vuelve a relacionarse con su entorno de la misma forma en la que lo hacía antes de padecerla. Solo en ese momento la enfermedad es erradicada, solo en ese momento *el ser orgánico ejerce normalmente todas sus funciones*.

1.2 Introducción y motivación

Actualmente existen alrededor de 10.000 enfermedades conocidas de las cuales solo 500 de ellas cuentan con una cura (un 5%) (Upton & DeGette, 2015; Kessler, 2016).

Entre el grupo principal de enfermedades que no tienen cura se encuentran aquellas enfermedades catalogadas como enfermedades raras, entendiendo la OMS que son enfermedades raras aquellas que afectan a menos de 5 de cada 10,000 personas (0,05%) (Rath et al., 2012; Humphreys, 2012).

Estas enfermedades son comúnmente hereditarias y crónicas, y deben su bajo conocimiento de las mismas a la alta complejidad etiológica, diagnóstica y evolutiva (Huete García et al., 2009), no existiendo cura para ellas y contando, en muy pocos casos, con tratamientos y cuidados médicos dirigidos a mejorar la calidad de vida de las personas que las padecen.

Es imprescindible para los profesionales de la salud contar con información relevante, de calidad y actualizada, pues solo de esta forma es posible realizar una buena toma de decisiones, reducir el tiempo de diagnóstico, dirigir nuevas investigaciones y crear nuevos enfoques tendentes a desarrollar curas y tratamientos.

A mediados del siglo pasado se ha producido un gran aumento en el conocimiento médico debido a los diversos avances tecnológicos y médicos; cada vez se genera mayor contenido médico en menor tiempo. Algunos estudios han estimado que la frecuencia con la que se incrementará o se duplicará el conocimiento llegará a ser hasta de 73 días para el presente año (Densen, 2011).

Esta mejora en el aumento de conocimiento en menor tiempo se debe, en gran medida, a Internet. Cada vez se crean más bases de datos públicas accesibles a través de Internet, y disponibles para cualquier persona. Esto permite incrementar el conocimiento que podemos tener de un tema determinado, pues no solo contamos con la información y conocimiento que nosotros poseemos, sino con la información y conocimiento que el resto también comparte.

Como muestra de este avance en el conocimiento, la Figura 1 muestra el incremento de estudios e investigaciones científicas en el campo de la medicina entre el año 2000 y el 2018 en PubMed (Williamson & Minter, 2019), uno de los recursos Web con mayor información biomédica concentrada, que cuenta ya en 2019 con más de 28 millones de artículos científicos en su acervo (Yang & Lee, 2018), rastro indiscutible de la inmensa cantidad de datos médicos a los que se puede tener acceso.

1.2 Introducción y motivación

Figura 1. Nuevos registros de PubMed y PubMed Central (PMC) agregados por año (por 10,000s). Fuente: Exploring PubMed as a reliable resource for scholarly communications services (Williamson and Minter, 2019).

Con el fin de lograr un mejor conocimiento de las enfermedades los científicos han venido utilizando diversos recursos bioinformáticos mediante los cuales han logrado caracterizarlas, es decir, analizar y establecer cuáles son sus principales características con el fin de poder relacionarlas y distinguirlas de las demás.

En este contexto nace la denominada red de enfermedades, concepto introducido por K.-I. Goh et al. en el año 2007 en su publicación titulada *The human disease network* (HDN) (Goh et al., 2007), donde se publica la construcción de una red que incluye 1.284 enfermedades catalogadas en base a sus elementos genéticos.

Posteriormente, en el año 2014, (Zhou et al., 2014b), publica una investigación en la cual se construye una nueva red llamada *Human symptoms-disease network* (HSDN) un nuevo enfoque de la HDN en la que la catalogación de las enfermedades ya no se realiza en base a sus elementos genéticos sino en base a sus manifestaciones fenotípicas (Zhou et al., 2014b).

Gracias al estudio de la HSDN se ha podido afirmar que, en muchos casos, aquellas enfermedades que comparten elementos fenotípicos comparten no solamente eso, sino también elementos biológicos, como genes o interacciones (proteínas/proteínas; rutas moleculares, etc.).

1.2 Introducción y motivación

Yendo más allá, sugieren que el estudio de esta red podría abordar cuestiones tan importantes como las asociaciones inesperadas entre distintas enfermedades, es decir, el descubrimiento de otras similitudes (distintas a los elementos fenotípicos), entre las diferentes enfermedades que forman dicha red.

En última instancia, el conocimiento de estas interrelaciones podría ser tremendamente útil para el descubrimiento de nuevos usos de fármacos en otras enfermedades diferentes a las cuales se utilizan actualmente, lo que es conocido como reposicionamiento de fármacos.

Todos estos enfoques han abierto un amplio horizonte de conocimiento, ya que la hipótesis de que diferentes enfermedades, causadas por diferentes organismos y con distintos genes involucrados comparten características biológicas y genéticas, permiten formular otras grandes hipótesis relativas al diagnóstico, tratamiento y pronóstico de una enfermedad.

La investigación objeto de esta tesis pretende ir más allá en lo que al estudio de las enfermedades se refiere. Mientras las dos redes anteriormente mencionadas están formadas por la información médica contenida en OMIN y PubMed exclusivamente, mediante este proyecto se pretende la creación de un conjunto de datos de contenido médico relevante extraídos de la mayor cantidad de fuentes de información pública posible.

Por ello, durante dos años se ha minado el contenido médico de algunas de estas fuentes de información públicas (Wikipedia, PubMed y Mayo Clinic), y con la información obtenida se ha creado un conjunto de datos propio.

Este conjunto de datos no se ha creado de una única vez y con una única extracción, sino que se ha ido generando y actualizando mediante el mismo proceso de extracción y generación de conocimiento dos veces por mes, con el fin de poseer la mayor cantidad de información médica posible y, de cara al futuro, poder analizar la evolución de dicha información en las fuentes de información públicas utilizadas. No obstante, se debe mencionar que dicho proceso de generación quincenal solo ha sido aplicado sobre Wikipedia, debido a que el motivo de investigación principal de esta tesis es, precisamente, el análisis de la evolución de la información almacenadas en fuentes con gran variabilidad de información; así, la información contenida en Mayo Clinic es bastante estática, ya que son definiciones o descripciones de enfermedades proporcionadas por dicha entidad, y que carecen de grandes modificaciones. De igual manera ocurre con PubMed, donde si bien es cierto que la información puede aumentar con la inclusión de nuevas publicaciones científicas, no se modifica.

A modo conclusivo en este punto recalcar que, a pesar de que se ha utilizado la información médica contenida en tres fuentes de información pública (Wikipedia, PubMed y Mayo Clinic), solo Wikipedia cumple los requisitos necesarios para

1.3 Hipótesis de investigación

que su información pueda ser analizada a lo largo del tiempo, debido a la gran variabilidad de la misma. Así:

- Solo Wikipedia (dentro de las tres fuentes analizadas) es susceptible de ser modificada por el público en general, de ahí que la información se encuentre constantemente actualizada.
- Con respecto a la evolución del contenido médico de PubMed tiene dos limitaciones: la primera de ellas es que solo han podido analizarse los abstracts de los artículos de libre acceso al público (no centrándose en el artículo completo, que proporciona más información); y la segunda de ellas es que los artículos contenidos en PubMed no se modifican a lo largo del tiempo, solo aumenta el número de ellos, pero no aumenta o disminuye el contenido médico contenido en los artículos ya publicados.
- Con respecto a Mayo Clinic, la información que contiene tampoco es objeto de variaciones relevantes, debido a que únicamente puede ser modificada por el equipo de Mayo Clinic y por lo tanto la información permanece estática durante largos periodos, no siendo por lo tanto susceptible del análisis objetivo en esta tesis.

Por todo ello, todas las hipótesis referidas en el presente trabajo de tesis se refieren únicamente a Wikipedia, pues es la única fuente de información que cumple con los requisitos necesarios para ser objeto de análisis a lo largo del tiempo: fuente de información médica, de libre acceso, con alto grado de variabilidad y con un gran nivel de fiabilidad.

El objetivo último de esta tesis es el análisis del conjunto de datos creado y actualizado periódicamente para tratar de determinar las similitudes fenotípicas existentes entre las diferentes enfermedades que la forman, para así poder catalogar las distintas enfermedades en base a dichas similitudes y poder tener un mejor entendimiento sobre la forma en que se relacionan entre ellas.

1.3 Hipótesis de investigación

Las hipótesis que pretende estudiar la presente tesis doctoral son:

- **Hipótesis 1 (H1):** Es posible extraer y generar información fenotípica proveniente de diversas fuentes de información pública, por medio de la implementación de un enfoque basado en el procesamiento de lenguaje natural (NLP *Natural-Language Processing*, por sus siglas en inglés).
- **Hipótesis 2 (H2):** La información fenotípica contenida en los artículos de Wikipedia clasificados como enfermedades es de calidad.

1.4 Objetivos

- **Hipótesis 3 (H3):** A medida que pasa el tiempo estas fuentes contienen más y mejor información que permiten enriquecer el conocimiento sobre las enfermedades.
- **Hipótesis 4 (H4):** La cantidad de elementos de conocimiento diagnóstico (DKE, por sus siglas en inglés) obtenidos de textos con contenido diagnóstico de los artículos de enfermedades en la versión inglesa de Wikipedia están directamente relacionados con el número de referencias científicas y el número de caracteres totales.
- **Hipótesis 5 (H5):** Los elementos de conocimiento de diagnóstico (DKE, por sus siglas en inglés) obtenidos de fuentes textuales de conocimiento públicas son suficientemente relevantes para caracterizar enfermedades en base a sus características fenotípicas.

1.4 Objetivos

El objetivo principal de esta tesis es:

- Analizar y caracterizar las enfermedades mediante sus elementos conocimiento diagnóstico contenidos en un conjunto de datos médicos construido a partir de la extracción de información de fuentes biomédicas para mejorar el conocimiento que se tiene sobre las enfermedades, y analizar la forma en la que se relacionan.

Para la consecución del objetivo principal debemos alcanzar, previamente, los siguientes objetivos:

- **Objetivo 1 (O1).** Identificar las fuentes de información Web que contengan información médica de calidad relativa a los elementos fenotípicos, y evaluar si estas fuentes son aptas para dicho fin, es decir, si son de acceso público, si su formato permite dicha extracción, etc.
- **Objetivo 2 (O2).** Diseñar un método de extracción y generación de conocimiento fenotípico que nos permita recuperar la información médica relevante para este trabajo y que, además, tenga la capacidad de ejecutarse de forma periódica automáticamente.
- **Objetivo 3 (O3).** Desarrollar una plataforma Web que permita la ejecución del método de extracción y generación de conocimiento propuesto en el O2.
- **Objetivo 4 (O4).** Evaluar manualmente el proceso de extracción y generación de conocimiento implementado en la plataforma Web, calculando

1.5 Estructura del documento

la precisión con la que el sistema detecta los elementos de diagnóstico de cada enfermedad.

- Objetivo 5 (O5). Desarrollar una API (Application Programming Interface) para proveer el conocimiento fenotípico generado por el enfoque propuesto a cualquier usuario que se encuentre registrado en la base de datos de usuarios oficiales.
- Objetivo 6 (O6). Relacionar las enfermedades contenidas en el conjunto de datos propio mediante diferentes coeficientes de similitud (Coseno, Jaccard, Dice, entre otros).
- Objetivo 7 (O7). Caracterizar y evaluar las relaciones entre enfermedades identificadas en O6 para ser agrupadas y clasificadas por medio de diversas técnicas de análisis de datos.
- Objetivo 8 (O8). Interpretación de los resultados obtenidos en el O7.

1.5 Estructura del documento

Dejando atrás lo anterior, la presente tesis se encuentra organizada de la siguiente forma:

- El Capítulo 1 , denominado “*Estado del arte*”, donde se lleva a cabo el análisis de la literatura de los trabajos relacionados referentes a repositorios de conocimiento biomédico, enfoques de extracción y generación de conocimiento biomédico, análisis que caracterizan y definen enfermedades, redes de enfermedades y enfoques de similitud entre enfermedades.
- El Capítulo 3 , denominado “*Metodología*”, se lleva a cabo la descripción del enfoque metodológico que ha permitido crear el conjunto de datos de enfermedades y de elementos de conocimiento diagnóstico. Que son la base para desarrollar los análisis de caracterización de enfermedades.
- El Capítulo 4 , denominado “*Resultados*”, se lleva a cabo la descripción de los resultados de los análisis que han permitido caracterizar y definir las enfermedades contenidas en el conjunto de datos médicos en base a los elementos de conocimiento diagnóstico que tienen en común.
- El Capítulo 5 , denominado “*Verificación de hipótesis*”, se lleva a cabo la descripción de como los experimentos de caracterización de enfermedades validan cada una de las hipótesis planteadas en este documento de tesis.

1.5 Estructura del documento

- Finalmente, el Capítulo 6 , denominado “*Conclusiones y trabajos futuros*”, se lleva a cabo la descripción de cada una de las conclusiones a las que ha conducido esta tesis, así como las líneas de trabajos futuros que permitirán desarrollar con más amplitud este proyecto.

Capítulo 2 Estado del arte

Como ya se ha anunciado, en el presente capítulo se lleva a cabo el análisis de la literatura de los trabajos relacionados referentes a los repositorios de conocimiento biomédico, enfoques de extracción y generación de conocimiento biomédico, análisis referentes a caracterización de enfermedades, redes de enfermedades y enfoques de similitud entre enfermedades. Pero antes es importante presentar la descripción de algunos conceptos médicos relevantes para este documento de tesis.

2.1 Conceptos médicos

En esta sección se definen aquellos conceptos médicos que son fundamentales para la correcta comprensión del contenido de esta tesis.

2.1.1 Enfermedad

Hay varias acepciones de lo que se entiende por enfermedad, destacamos las más relevantes:

La Organización Mundial de la Salud (OMS) define enfermedad como la “*Alteración o desviación del estado fisiológico en una o varias partes del cuerpo, por causas en general conocidas, manifestada por síntomas y signos característicos, y cuya evolución es más o menos previsible*” (Elsevier; Herrero Jaén, 2016).

En otro término, el diccionario de Oxford la define como “*Un desorden en la estructura o función en un humano, animal o planta, especialmente uno que produce síntomas específicos o que afecta a un lugar específico y no es simplemente un resultado directo de una lesión física*”.

En ambas definiciones destaca, como una de las características principales de una enfermedad, su manifestación mediante síntomas y signos, cuyo significado se desarrolla a continuación.

2.1.2 Síntomas y signos

Se entiende por signo como aquella manifestación objetiva de una enfermedad ya que tiene la cualidad de poder ser medido y valorado mediante la exploración médica del paciente (fiebre, hipo, tos, etc.).

Se entiende por síntoma como la manifestación subjetiva de una enfermedad; es subjetivo porque son percibidas únicamente por el paciente, de forma que el

2.1.3 Elemento(s) de diagnóstico

diagnóstico médico dependerá de la información que la persona que padezca dicha alteración transmita (dolor, debilidad o mareo).

El conjunto de síntomas y signos que se manifiestan en un paciente a causa de una enfermedad se denomina cuadro clínico (Herrero Jaén, 2016).

El conjunto de signos y síntomas que caracterizan a una enfermedad conforman las manifestaciones fenotípicas, siendo tratados en este trabajo como equivalentes conceptualmente hablando.

2.1.3 Elemento(s) de diagnóstico

Como su propio nombre indica, son elementos de diagnóstico aquellos de los que se sirven los profesionales de la salud para diagnosticar correctamente una enfermedad manifestada en una persona (signos, síntomas, procedimientos médicos, procedimientos de laboratorio, etc.).

Estos elementos son considerados como el primer punto de partida para determinar el diagnóstico de una enfermedad, siendo de vital importancia su estudio.

2.2 Repositorios de conocimiento biomédico

En esta sección se describirán los trabajos relacionados con las fuentes de información pública relevantes para la presente tesis.

La totalidad de los repositorios escogidos para la creación de este trabajo se encuentran en la Web, ya que son los que han experimentado un mayor crecimiento en las últimas décadas, tanto en lo referente a su contenido como en lo referente a los usuarios que hacen uso de ellas.

Esta gran expansión se debe a que la mayor parte de sus recursos son de libre y fácil acceso para todos, no solo para los profesionales e investigadores sino para el público en general, de forma que la *World Wide Web* se ha convertido en uno de los mejores medios para la transferencia de conocimiento del mundo.

Una de las características que marca el conjunto de datos creado en esta tesis son los elementos relativos al diagnóstico de las enfermedades. Estos términos médicos son extraídos mediante la herramienta MetaMap (Aronson, 2006), que obtiene conceptos del vocabulario del *Unified Medical Language System* encontrado en textos. En el 3.2.4 se describirá MetaMap con detalle.

2.2 Repositorios de conocimiento biomédico

El Sistema unificado de lenguaje médico o UMLS¹ por sus siglas en inglés *Unified Medical Language System* (Bodenreider, 2004) es un conjunto de archivos y software que agrupa una gran cantidad de vocabularios, fuentes y estándares de salud y biomedicina. UMLS integra y distribuye terminología elemental, estándares de codificaciones y clasificación, así como todo tipo de recursos para la creación de sistemas y servicios biomédicos, incluyendo los registros médicos electrónicos, con el fin de permitir interoperabilidad entre sistemas informáticos. Ha sido desarrollado por la Biblioteca Nacional de Medicina de los Estados Unidos (NLM², por sus siglas en inglés), que a su vez es parte de los centros y oficinas que conforman los Institutos Nacionales de Salud de los Estados Unidos (NIH³, por sus siglas en inglés).

Específicamente los términos médicos relativos a síntomas que se usan en esta tesis son del vocabulario SNOMED CT⁴, por sus siglas en inglés *Systematized Nomenclature of Medicine – Clinical Terms* (Donnelly, 2006). SNOMED CT proporciona una terminología general básica relativa a las historias clínicas electrónicas (EHR, por sus siglas en inglés). Sus conceptos tienen una organización jerárquica, un significado único y definiciones formales basadas en la lógica. Cada entrega de esta fuente, cuanta con un extenso conjunto de directrices, mapas cruzados con clasificaciones y esquemas de codificación existentes; todo esto como añadido a su núcleo de terminologías (conceptos, descripciones y relaciones). Solo la edición estadounidense de SNOMED CT US⁵ tiene más de 300.000 conceptos únicos, más de un millón de descripciones en SNOMED CT, que incluyen también sinónimos y aproximadamente 903.000 relaciones semánticas entre conceptos SNOMED CT. Su última versión disponible para descargar durante el tiempo de escritura de esta tesis es la de septiembre de 2019.

Así como SNOMED CT está incluido en el metatesauro de UMLS (Information et al., 2009) existen otros vocabularios y bases de datos como el *Medical Subject Headings* (MeSH, por sus siglas en inglés) (Lipscomb, 2000), ICD-10-CM (World Health Organization, 1993), *Logical Observation Identifier Names and Codes* (LOINC, por sus siglas en inglés) (Friedlin & McDonald, 2010), *Centers for Medicare & Medicaid Services* (CPT, por sus siglas en inglés) (CMS, 2019) o RxNorm⁶ bajo la misma licencia⁷. Actualmente UMLS cuenta con más de 100 fuentes diferentes y algunas en varios idiomas, la lista de todos ellos se encuentra disponible en línea⁸.

Sin duda UMLS es un repositorio imprescindible para diversos proyectos y sistemas, ya que es utilizado como un estándar. Para acceder a toda la gama de

¹ <https://www.nlm.nih.gov/research/umls/index.html>

² <https://www.nlm.nih.gov/>

³ <https://www.nih.gov/>

⁴ <https://www.nlm.nih.gov/healthit/snomedct/index.html>

⁵ https://www.nlm.nih.gov/healthit/snomedct/us_edition.html

⁶ <https://www.nlm.nih.gov/research/umls/rxnorm/index.html>

⁷ <https://uts.nlm.nih.gov/license.html>

⁸ <https://www.nlm.nih.gov/research/umls/sourcereleasedocs/index.html>

2.2 Repositorios de conocimiento biomédico

productos y servicios⁹, *The UMLS Terminology Services* (UTS), UMLS pone a disposición de todos, un registro, que otorga licencias de uso de forma gratuita¹⁰; una vez registrado desde su aplicación Web también existe la opción de realizar consultas de conceptos. La última versión descargable de UMLS durante la escritura de esta tesis es la 2019AB y se encuentra disponible en línea¹¹.

Otro repositorio importante es la *Online Mendelian Inheritance in Man*, OMIM¹² por sus siglas en inglés, fue iniciada por el Dr Victor A. McKusick durante la década de 1960 y actualmente es mantenida y editada por el *McKusick-Nathans Institute of Genetic Medicine* y el *Johns Hopkins University School of Medicine*, bajo la dirección del Dr. Ada Hamosh; y disponible en línea por el *National Center for Biotechnology Information* (NCBI) de la NLM de Estados Unidos (Hamosh et al., 2005; Amberger, Bocchini & Hamosh, 2011; Amberger et al., 2015). OMIM es un compendio curado (Buneman, 2009) de genes humanos, fenotipos y trastornos genéticos; cada uno de los más de 15.000 registros de genes en OMIM tiene asociado textos de resúmenes de artículos de PubMed referenciados, a bases de datos de secuencias de proteínas y de mutaciones genéticas. Una de las nomenclaturas utilizadas en OMIM es el *HUGO (Human Genome Organisation) Gene Nomenclature Committee* HGNC¹³ (Povey et al., 2001). Los datos de OMIM se encuentran disponibles a través de la descarga directa de *dumps* desde su página Web oficial¹⁴; también ahí se encuentra disponible una API¹⁵ con la que también se consume su información, tanto la descarga de *dumps* como el acceso a la API requieren que el usuario interesado se registre; lo que no requiere registro es el acceso al motor de búsqueda incorporado dentro de su página Web.

OrphaNet es un repositorio especializado importante, ya que se trata de una iniciativa europea, apoyada por la Comisión Europea y el Ministerio de Sanidad Francés. Tiene como objetivo, mejorar el diagnóstico y tratamiento de las enfermedades raras, y por ello, no solo provee una base de datos dedicada exclusivamente a la información sobre las enfermedades raras y medicamentos huérfanos¹⁶, sino que ofrece una serie de servicios tanto para los pacientes y familiares; extendiéndose y centrándose en gran medida a los investigadores y a los profesionales de la salud, ya que pretende mantenerlos siempre informados (Weinreich et al., 2008; Rath et al., 2012; OrphaNet, 2012). OrphaNet contiene información de 6,172 enfermedades raras únicas, de las cuales, el 71.9% son genéticas (Wakap et al., 2019). En su sitio Web oficial se encuentra disponible en varios idiomas y en él se muestra toda la gama de servicios que ofrece, entre ellos, está un buscador de enfermedades como medio para acceder a sus datos, no tiene alguna forma de

⁹ https://eresources.nlm.nih.gov.nlm_eresources/

¹⁰ <https://uts.nlm.nih.gov/home.html>

¹¹ <https://www.nlm.nih.gov/research/umls/licensedcontent/umlsknowledgesources.html>

¹² <https://omim.org/>

¹³ <https://www.genenames.org/>

¹⁴ <https://omim.org/downloads/>

¹⁵ <https://omim.org/api>

¹⁶ https://www.orpha.net/consor/cgi-bin/Education_AboutOrphanDrugs.php?lng=ES

2.2 Repositorios de conocimiento biomédico

descargar todos sus datos, ni una API (Maiella et al., 2013). Entre otras bases de datos dedicadas a proveer información sobre enfermedades raras, descriptores, pruebas y mutaciones genéticas, se encuentran: OMIM, *Gene Reviews*¹⁷, *Genetics Home Reference*¹⁸, *National Organization for Rare Disorders* (NORD, por sus siglas en inglés)¹⁹ y *Genetic Alliance*²⁰ y DiseaseCard, que se describirá en este capítulo.

Un repositorio de clasificación de suma importancia en el ámbito biomédico es *The Medical Subject Headings*, MeSH²¹ por sus siglas en inglés. Este tesauro, es un vocabulario controlado y jerárquico desarrollado por la NLM, utilizado para buscar, indexar y catalogar información y documentos biomédicos relacionados con la salud. MeSH proporciona búsquedas de datos por tema que se encuentran en la base de datos MEDLINE/PubMed²² de citas de revistas, en el catálogo NLM y en otras bases de datos NLM. Este recurso está continuamente actualizado, incorporando términos de la literatura e investigación científica por especialistas de la salud del NLM, incluso existe la opción de recomendar nuevos términos para ser añadidos. Sus registros se encuentran divididos en tres tipos: descriptores (*main headings*) “*Descriptors*”, calificadores (*subheadings*) “*Qualifiers*” y registros médicos suplementarios “*Supplementary Chemical Records*” (SCR, por sus siglas en inglés). Los descriptores se encuentran organizados en cuatro clasificaciones en una estructura de árbol numerada, que está disponible en línea²³, que va de los temas más generales a los más específicos; los calificadores, se usan con los descriptores para agrupar las citas de los diversos temas; y finalmente, los SCR que indexan sustancias químicas, drogas y términos semejantes, de ellos hay cuatro tipos diferentes. Los datos MeSH se encuentran disponibles en diversos formatos para su descarga desde su página Web oficial, siendo la versión del 2019 la última; también sus datos están disponible a través de la E-utilities API (Sayers, 2018) y por medio de RDF²⁴.

Dentro del ámbito de las ontologías se encuentra la bien conocida *Disease Ontology* (DO) (Schriml et al., 2012, 2019; Kibbe et al., 2015) a cargo de la doctora Lynn M. Schriml de la *University of Maryland, School of Medicine in the Department of Epidemiology and Public Health and at the Institute of Genome Science* (IGS), en Baltimore, Maryland. La DO es una ontología estandarizada de enfermedades humanas hereditarias, del desarrollo y adquiridas, que tiene como fin ofrecer a la comunidad biomédica descripciones consistentes, reutilizables y sostenibles de los conceptos de enfermedades humanas, de características fenotípicas y

¹⁷ <http://www.ncbi.nlm.nih.gov/sites/GeneTests/review?db=GeneTests>

¹⁸ <http://ghr.nlm.nih.gov/>

¹⁹ www.rarediseases.org

²⁰ www.geneticalliance.org

²¹ <https://www.ncbi.nlm.nih.gov/mesh>

²² <https://www.nlm.nih.gov/bsd/pmresources.html>

²³ <https://meshb.nlm.nih.gov/search>

²⁴ <https://id.nlm.nih.gov/mesh/query>

2.2 Repositorios de conocimiento biomédico

de los conceptos relacionados con las enfermedades en otros vocabularios biomédicos. La DO hace un extenso mapeo cruzado de sus conceptos con los de otras fuentes como MeSH, UMLS, ICD-10 e ICD-9 (World Health Organization, 1993), el tesaurus de NCI²⁵, SNOMED y OMIM. Esta ontología de código abierto, en su versión 3 (Schriml et al., 2012), ya tenía una amplia base de conocimiento con 8,043 enfermedades humanas. Todos los datos de la DO se encuentran disponibles a través de un motor de búsqueda incorporado en su página Web oficial; así como también tiene disponible para descargar el *dump* con todos sus datos; y adicionalmente ofrece descargar consultas RDF. La última versión disponible es la de 2019.

En el contexto de las ontologías, se presenta la *Human Phenotype Ontology* (HPO, por sus siglas en inglés)²⁶ que fue desarrollado por la Iniciática Monarch²⁷. La HPO proporciona un vocabulario estandarizado de las manifestaciones fenotípicas que se presentan en las enfermedades humanas, es decir, cada termino en este vocabulario es relativo y describe una anomalía fenotípica. La HPO se ha desarrollado utilizando literatura médica, es decir, que obtiene sus datos a partir de fuentes como, OrphaNet (Weinreich et al., 2008), DECIPHER²⁸ y OMIM (Hamosh et al., 2005). Como resultado, la HPO se forma por tres componentes: el vocabulario del fenotipo, las anotaciones del fenotipo de la enfermedad y los algoritmos que operan sobre ellos. Esta ontología cuenta actualmente con más de 13.000 conceptos médicos y con más de 156.000 anotaciones sobre enfermedades hereditarias. En su sitio Web se encuentra un motor de búsqueda para acceder a sus datos y también se encuentra la opción de ser descargados, no cuenta con una API (Robinson et al., 2008; Groza et al., 2015; Köhler et al., 2017).

La *Gene Ontology* (GO, por sus siglas en inglés), es un recurso fundado en 1998, que proporciona conocimiento estructurado sobre los genes y los productos genéticos, legible tanto para humanos como para ordenadores, abarcando desde los humanos hasta las bacterias. Actualmente su base de datos comprende miles de organismos diferentes y proporciona de forma flexible y dinámica descriptores de secuencias homólogas de genes y proteínas a lo largo del espectro filogenético (Ashburner et al., 2000; The Gene Ontology et al., 2019). Desde su página Web oficial se permite la descarga de su ontología y de sus anotadores; así como la descarga de una herramienta de anotaciones.

Hasta ahora se han presentado repositorios que permiten organizar, clasificar, anotar e identificar información médica, por medio de diferentes estructuras de datos, como lo son las ontologías, nomenclaturas, códigos, tesauros, diccionarios,

²⁵ <https://ncithesaurus.nci.nih.gov/ncitbrowser/>

²⁶ <https://hpo.jax.org/app/>

²⁷ <https://monarchinitiative.org/>

²⁸ <https://decipher.sanger.ac.uk>

2.2 Repositorios de conocimiento biomédico

entre otros. Ahora, presentaremos repositorios que contienen todo tipo de información acerca de las enfermedades humanas como, manifestaciones clínicas, genes, los fármacos que las tratan, entre otras características.

Una de las fuentes de información más conocidas, se encuentra la *Diseases Database* (DD²⁹, por sus siglas en inglés), que es una base de datos en línea editada por el Dr. Malcolm H. Duncan que también es director del *Medical Object Oriented Software Enterprises Ltd*³⁰, entidad por la que es financiada esta base de datos; y registrada en 2001 en la *Health On the Net Foundation*, HON³¹ por sus siglas en inglés; y también se encuentra incluido en UMLS. El objetivo con esta base de datos ha sido desarrollar un índice de referencia cruzada de enfermedades humanas, medicamentos de marca no preferidos, síntomas, signos y hallazgos de investigación anormales; centrando su contenido en la medicina interna, enfermedades hereditarias, bioquímica clínica y farmacología³². DD ofrece adicionalmente a sus datos, una indexación semántica a cada una de las diferentes fuentes de las que puede llegar a conseguir información, alguna de las fuentes son: bases de datos medicas como, MedHunt o MalaCards (Rappaport et al., 2014); fuentes incorporadas a la US NLM como, PubMed (pubmeddev, 2019), PubChem³³, MedlinePlus³⁴; diccionarios, como *OneLook*³⁵, *On-line Medical Dictionary*; libros de texto en línea, como *GP Notebook*³⁶, *MSD (Merck) Manual*³⁷, Radiopedia³⁸, DailyMed³⁹; base de datos de fármacos y medicaciones como, DrugBank⁴⁰ o EMC⁴¹; y páginas o motores de búsqueda variados como, Bing, BioMedSearch⁴², Google, *NIH Clinical Trials*⁴³, SNOMED, WikiDoc y Wikipedia. El índice que se encuentra en su página Web oficial es similar a un libro de texto médico y allí también proveen un motor de búsqueda limitado que sirve como único modo de acceso a sus datos; también incorpora otro buscador con el cual se permite realizar búsqueda por términos SNOMED CT. No posee API, ni la opción de descargar sus datos, lo cual hace difícil el uso de su conocimiento.

Otra fuente de información es DISEASES⁴⁴, es una base de datos en línea que integra evidencias sobre las asociaciones genéticas de las enfermedades, que ha sido desarrollada en el *Novo Nordisk Foundation Center for Protein Research*⁴⁵; sus creadores han desarrollado un método de reconocimiento de entidades (NER,

²⁹ <http://www.diseasesdatabase.com>

³⁰ <http://www.mrtabnet.demon.co.uk/moose1.htm>

³¹ <https://www.hon.ch/home.html>

³² <https://www.nlm.nih.gov/research/umls/sourcereleasedocs/current/DDB/index.html>

³³ <https://pubchem.ncbi.nlm.nih.gov/>

³⁴ <https://medlineplus.gov>

³⁵ <https://www.onelook.com>

³⁶ <https://www.gpnotebook.co.uk>

³⁷ <https://www.msdmanuals.com>

³⁸ <https://radiopaedia.org>

³⁹ <https://dailymed.nlm.nih.gov/dailymed>

⁴⁰ <https://www.drugbank.ca>

⁴¹ <https://www.medicines.org.uk>

⁴² <http://www.biomedsearch.com>

⁴³ <https://www.clinicaltrials.gov>

⁴⁴ <https://diseases.jensenlab.org/Search>

⁴⁵ <https://www.cpr.ku.dk/>

2.2 Repositorios de conocimiento biomédico

por sus siglas en inglés) para etiquetar primero nombres de enfermedades, genes humanos y nombres de proteínas normalizados a identificadores de los recursos: Disease Ontology⁴⁶, STRING (Mering et al., 2003). Y así con las entidades detectadas, se procesan textos biomédicos de MedlinePlus (recuperados usando minería de textos) para encontrar las asociaciones gen-enfermedad. Sus datos y su software de NER se encuentran disponibles gratuitamente a través de su interfaz Web (Pletscher-Frankild et al., 2015).

Un recurso biomédico formado por una herramienta Web que integra colaborativamente información genética y médica, es DiseaseCard⁴⁷, cuyo objetivo es permitir a los profesionales de la salud reducir el esfuerzo y tiempo de búsqueda a través de bases de datos genéticas y bioinformáticas diseminadas por distintos grupos de investigación; y así acceder y relacionar información sobre enfermedades genéticas raras. El sistema desarrollado presenta los datos de las enfermedades en forma de una tarjeta informativa, que contiene información de diferentes fuentes médicas, por ejemplo, para la información general de enfermedades, consultan MalaCards, OrphaNet (Weinreich et al., 2008), OMIM (Hamosh et al., 2005), etc; para información sobre fármacos consulta PharmaGKB⁴⁸, entre otras; también consulta ontologías como GO (Ashburner et al., 2000), ICD-10⁴⁹ o MeSH⁵⁰; y fuentes literarias como PubMed. Desde su página Web se pueden hacer consultas y visualizar los resultados conectados por medio de un árbol de grafos donde se aprecian las relaciones entre los distintos tipos de conocimientos de las distintas bases de datos; también cuenta con una API⁵¹ con la que se accede a todos los datos de enfermedades recogidos en su base de conocimiento por medio de servicios de interfaces LinkedData⁵², un endpoint SPARQL y un conjunto de servicios REST. A pesar de la potencia de la herramienta solo se enfoca a un subconjunto de enfermedades (Oliveira et al., 2004; Dias et al., 2005; Lopes & Oliveira, 2013).

En la Web también se encuentra MalaCards: The Human Disease Database que fue creada en el Weizmann Institute of Sciences en Israel, y pretende ser la base de datos de enfermedades humanas más completa y homogénea disponible; diseñada para permitir a los investigadores y profesionales genómicos y geotécnicos navegar de forma eficiente por el mundo de los genes humanos, las variantes genéticas, las proteínas, las células y las rutas relacionadas con varias enfermedades (Espe, 2018). Sin duda es una robusta base de datos de enfermedades humanas ya que recopila información de enfermedades de al menos 68 fuentes de datos distintas; se encuentra modelada sobre la arquitectura y enfoque de la base de datos GeneCards (Rebhan et al., 1997) de genes humanos (Rappaport et al., 2013,

⁴⁶ <http://www.disease-ontology.org>

⁴⁷ <http://bioinformatics.ua.pt/diseasecard/>

⁴⁸ <https://www.pharmgkb.org>

⁴⁹ <https://www.cdc.gov/nchs/icd/icd10.htm>

⁵⁰ <https://www.ncbi.nlm.nih.gov/mesh>

⁵¹ <http://bioinformatics.ua.pt/diseasecard/about#api>

⁵² <http://linkeddata.org>

2.2 Repositorios de conocimiento biomédico

2014, 2017). MalaCards cuenta con una técnica llamada “*disease name unification*”, que le permite tener una lista homogénea de más de 19.000 enfermedades humanas. Cada enfermedad es representada por medio de tarjetas Web llamadas “MalaCard”, que incluyen información de al menos 18 secciones diferentes como: resúmenes de texto, signos y síntomas, contexto anatómico, fármacos, pruebas genéticas, genes relacionados, enfermedades relacionadas, publicaciones relevantes, entre otras. En su página Web tiene la lista entera de enfermedades humanas categorizada, así como un poderoso motor de búsqueda; se permite descargar a los usuarios registrados cada MalaCards de forma independiente, y no cuenta con una API, ni con la opción de descargar todos sus datos. También dispone de una serie de herramientas de análisis, pero no gratuitamente.

La dispersión de la información relativa a los genes y creciente dificultad de encontrar conocimiento biomédico comprensible sobre determinados genes y sus productos permitió, la concepción de GeneCards: *The Human Gene Database* (Rebhan et al., 1997), tiene como fin ser una comprensible compendio de genes humanos, facilitando la navegar e interrelacionar eficazmente el amplio universo de genes, enfermedades, variantes, proteínas, células y rutas biológicas humana (Safran et al., 2002). De esta base de datos se desprende la fuente de información relativa a enfermedades humanas, MalaCards (Rappaport et al., 2013), ambas fueron creadas en el Weizmann Institute of Sciences en Israel. GeneCards integra información extraída de los que sus autores llaman “base de datos de nivel inferior” que tratan de genes humanos, el número de las bases de datos utilizadas es de al menos 180, que incluye fuentes como, HUGO (Povey et al., 2001), DataMed (Chen et al., 2018), Disease Ontology (Schröml et al., 2012), OMIM (Hamosh et al., 2005), entre otras; estas contienen información sobre los genes, las funciones de las proteínas que codifican y las enfermedades involucradas. En su sitio Web, GeneCards gestiona el conocimiento de los genes a través de elementos llamados “GeneCard”, que proporciona y refunde información relativa a, los nombres de los genes, a sinónimos, a genes homólogos en el ratón, ubicaciones cromosómicas de los genes y de sus homólogos, una lista de las enfermedades en las que el gen está implicado de acuerdo con la evidencia genética, entre otros campos de interés. Así mismo en línea cuenta con un potente motor de búsqueda bien documentado y una serie de herramientas de pago de análisis; no cuenta con una API o cualquier otro mecanismo que de acceso a su conocimiento.

En la línea de las enfermedades y sus genes esta DisGeNet (Bauer-Mehren et al., 2011; Piñero et al., 2017) que fue creada con el fin de almacenar información referente a la base genética de enfermedades humanas creando una base de datos integrada de asociaciones gen-enfermedad, utilizando la información de diversas bases de datos como UniProt (Wu et al., 2006), PsyGeNET (Gutiérrez-Sacristán et al., 2015), ClinVar (Landrum et al., 2014), Orphanet (Weinreich et al., 2008), entre otros y por recursos textuales aptos para la minería de textos como dbSNP (Sherry et al., 2001), EXAC (Karczewski et al., 2017), 1000 Genomes Project

2.2 Repositorios de conocimiento biomédico

(Siva, 2008), entre otros, también hace uso de vocabularios médicos como UMLS, OMIM, MeSH, entre otros. Actualmente en su versión v5.0 contiene 561,119 asociaciones gene-disease, entre 17.074 genes y 20.370 enfermedades y características, y 135.588 asociaciones variant-disease, entre 83.002 variantes y 9.169 enfermedades y fenotipos⁵³. Cuenta con una interfaz Web en la que se permite hacer búsquedas y visualizaciones en forma de grafos, un punto SPARQL, una API y también cuenta con la opción de descargar sus datos en bruto.

Algunas bases de datos suelen estar muy enfocadas y limitadas en ciertos tipos de enfermedades, como lo es PedAM (por sus siglas en inglés, Pediatric Disease Annotations & Medicines) fue construido en el Center of Bioinformatics in East China Normal University (Jia et al., 2018). PedAM, y es una base de datos de libre acceso que permite anotaciones de enfermedades pediátricas y medicina, para estandarizar y clasificar las enfermedades pediátricas. Integra conocimiento de fuentes biomédicas, como los datos clínicos de los Registros Médicos Electrónicos, de ontologías ya existentes, como la DO (Schröml et al., 2012), o la HPO (Robinson et al., 2008) para extraer manifestaciones de enfermedades, que son utilizados como patrones sintácticos para ser reconocidos más de 774.514 artículos de texto completo y 8.848.796 resúmenes MEDLINE⁵⁴. El objetivo de esta fuente es facilitar la medicina de precisión en pediatría para así evitar diagnósticos erróneos y tratamientos ineficaces. Actualmente PedAM cuenta con 8.528 términos médicos referentes a enfermedades pediátricas, de las cuales 4.542 son conceptos únicos de enfermedades y sinónimos 3,986; también cuenta con 8 campos de anotación por cada enfermedad, que incluyen, genes, síntomas, sinónimos, referencias cruzadas, fenotipos humanos y sus correspondientes fenotipos en ratones. En su sitio Web se encuentra la opción de visualizar las conexiones de enfermedad basadas en manifestaciones o genes, mediante el uso de un motor de búsqueda.

Otra base de datos que tiene sus límites muy especificados, ya que se solo se enfoca en las enfermedades de cáncer, es BioXpress (Wan et al., 2015), que es una base de datos compuesta de expresiones genéticas y de asociación al cáncer, donde los niveles de expresión se mapean a los genes usando datos de RNA-seq (secuenciación de ARN), que son recuperados de las bases de datos de The Cancer Genome Atlas (TCGA⁵⁵), International Cancer Genome Consortium (ICGC⁵⁶), Expression Atlas⁵⁷ y de la literatura biomédica, como PubMed (NCBI Resource Coordinators, 2016), UniProtKB/Swiss-Prot (Wu et al., 2006). Para facilitar un análisis pan-cancer⁵⁸, todos los tipos de cáncer han sido mapeados a términos de la DO (Schröml et al., 2012). BioXpress contiene datos de expresión de 64 tipos de cáncer, 6.361 pacientes y 17.469 genes con 9.513 de los genes que muestran

⁵³ <http://www.disgenet.org/web/DisGeNET/menu/dbinfo#sources>

⁵⁴ <https://www.nlm.nih.gov/bsd/medline.html>

⁵⁵ <http://cancergenome.nih.gov/>

⁵⁶ <https://icgc.org/>

⁵⁷ <https://www.ebi.ac.uk/gxa/home>

⁵⁸ <https://www.nature.com/collections/ffebghgeba/>

2.3 Enfoques de extracción y generación de conocimiento biomédico

expresión diferencial entre muestras tumorales y normales. En su página Web se encuentra accesibles libremente sus datos para su descarga, también cuenta con un sencillo motor de búsqueda, que permite consulta por tipo de cáncer y por transcripciones; también cuenta con una API.

La *Human Gene Mutation Database* (HGMD), es una base de datos cuya información proviene totalmente de la curación manual de la literatura científica. HGMD es de acceso abierto dedicada al almacenamiento y conservación de interacciones proteicas, genéticas y químicas, de las principales especies de organismos y de los seres humanos (Stenson et al., 2014). Actualmente tiene información de al menos 1.598.688 interacciones biológicas anotadas manualmente de 55.809 publicaciones y recientemente recoge información sobre interacciones químicas-proteína de la fuente DrugBank (Wishart et al., 2006). Todos los datos de BioGRID están disponibles gratuitamente y sin restricciones desde su página Web, y están disponibles para su descargar directa en formatos estándar que se incorporan fácilmente a través de sus servicios Web.

Finalmente, a modo de complemento a esta subsección, el trabajo de Zou et al. (Zou et al., 2015) cuenta con una extensa lista de bases de datos, recursos biomédicos y bioinformáticos. Esto permite ver la gran cantidad de información que día con día se crea y almacena en Internet.

2.3 Enfoques de extracción y generación de conocimiento biomédico

Con la intención de organizar y complementar el estado del arte del correspondiente documento de tesis, en esta sección se incluyen investigaciones en relación con herramientas, proyectos e iniciativas basadas en la Web, que estén centrados en extraer y generar conocimiento relativos a enfermedades y sus manifestaciones fenotípicas usando fuentes de información biomédicas.

Algunos de los enfoques de extracción son novedosos y sencillos, como el de (Rodríguez González et al., 2015), en el cual sus autores han desarrollado un proceso en el cual hacen uso de *Web scraping*, combinado con técnicas de procesamiento de lenguaje natural para así extraer conceptos relativos a criterios de diagnóstico, y que estén contenidos en los artículos Web referentes a enfermedades infecciosas en MedlinePlus. El proceso de *Web scraping* se encarga de recuperar aquellos textos de las secciones relativas a las manifestaciones fenotípicas de las enfermedades y a las de pruebas y exámenes médicos. Estos textos después son tratados con la herramienta MetaMap (Aronson, 2006) para obtener de ellos, según una configuración especial, términos médicos referentes a, signos y síntomas, enfermedades y síndromes, hallazgos médicos, pruebas y exámenes de laboratorios, entre otros términos médicos. Al final de su proceso de generación de

2.3 Enfoques de extracción y generación de conocimiento biomédico

conocimiento, los autores implementan un proceso de validación de términos médicos que se alimenta, extrayendo conceptos médicos de diversas fuentes de confianza, fuentes ontológicas, fuentes colaborativas, fuentes de UMLS y de MedLine, para que finalmente poder separar aquellos términos médicos que no son correctos.

En la misma línea del trabajo mencionado en el párrafo anterior y desarrollado por los mismos autores (Rodríguez González et al., 2015) expande sus experimentos ya que en esta ocasión hace uso de dos herramientas de NLP: MetaMap (Aronson, 2006) y cTAKES (Savova et al., 2010); obteniendo, por medio de una evaluación, resultados igual de buenos, proponiendo por lo tanto, la integración de sus resultados.

Ahora bien, en los trabajos de (Hassan, Coulet & Toussaint, 2014; Hassan et al., 2015), los autores se centran en la identificación de las relaciones enfermedad-síntoma, ya que consideran que las actuales bases de datos que las catalogan se encuentran incompletas con respecto a la literatura científica disponible. SPARE (*standing for Syntactic PAttern for Relationship Extraction*, por sus siglas en inglés), es la herramienta que desarrollaron en su investigación, la cual que les permite extraer automáticamente relaciones enfermedad – síntoma; por medio de tres pasos: i) el aprendizaje y obtención de los patrones de los gráficos de dependencia, ii) la selección de los mejores patrones según su calidad y especificidad, iii) aplicación de los patrones en los textos para extraer las relaciones enfermedad-síntoma. Utilizaron su herramienta sobre los resúmenes de los artículos del corpus de PubMed que son relativos a 457 enfermedades raras; y la validaron, obteniendo buenos resultados (medida de 55,65%: para la especificidad $\geq 0,5$ y la calidad $\geq 0,5$) que compararon con fuentes como OrphaData (Weinreich et al., 2008) y OMIM (Hamosh et al., 2005), y que han mostrado la viabilidad de extraer automáticamente las relaciones enfermedad-síntoma, incluyendo las relaciones verdaderas que no estaban referenciadas en las bases de datos de fenotipo y que pueden implicar descripciones complejas de los síntomas.

Otro trabajo relacionado con las enfermedades raras es el de (Cousyn et al., 2018), cuya finalidad es la de abordar la difusión del conocimiento sobre las enfermedades raras, ya que mencionan que la investigación de las casi siete mil enfermedades raras existentes es insuficiente. Los autores se centran en la extracción automática de los síntomas de las enfermedades raras de la literatura mediante un enfoque que utiliza un algoritmo de NER (*Named Entity Recognition*, por sus siglas en inglés) que está basado en la estadística numérica Frecuencia de Término – Frecuencia Inversa de Documentos (TF-IDF). Para ello, primero hacen uso de la lista e información básica de enfermedades raras de Orphanet (Weinreich et al., 2008); segundo, utilizan la herramienta *Entrez Programming Utilities API* (Sayers, 2010) para obtener de PubMed todas aquellas publicaciones relacionadas con las enfermedades; tercero, mediante un proceso propio de NLP, que usa información de la HPO (Robinson et al., 2008), obtiene los síntomas. Su diccionario

2.3 Enfoques de extracción y generación de conocimiento biomédico

NER obtuvo una precisión de 1,83%, un *recall* de 59,66% y un *F-score* de 3,55%; estos resultados que describen los autores se deben a que casi el 40% de los síntomas faltan en las publicaciones de las enfermedades, es decir, que PubMed no cuenta con todo el contenido necesario de enfermedades raras.

Dentro de esta semántica de trabajos que enfocan sus esfuerzos en la observación y análisis de las relaciones de enfermedad-manifestaciones (D-M), se encuentra el trabajo de (Xu, Li & Wang, 2013), el cual presenta un método automático para extraer los pares de manifestaciones de enfermedades de la amplia literatura biomédica encontrada en MEDLINE (más de 22 millones de citas) y así reducir el esfuerzo humano requerido. La base de datos resultante se encuentra disponible en línea; su desarrollo se ha realizado, primero utilizando 119.085.682 frases de MEDLINE; segundo, se extrajeron pares D-M de la fuente UMLS como conocimiento previo, para descubrir patrones sintácticos; tercero, obtención de más pares sintácticos de MEDLINE usando los patrones aprendidos; finalmente, los autores han analizado las correlaciones entre las manifestaciones de la enfermedad, los genes y los fármacos. Finalmente, extrajeron 121.359 pares únicos D-M con una alta precisión de 0,924. Entre los pares extraídos, 120.419 (99,2%) no han sido capturados en las fuentes de conocimiento estructurado existentes. Los autores aseguran haber demostrado que las manifestaciones de la enfermedad se correlacionan positivamente tanto con los genes asociados a la enfermedad como con los tratamientos farmacológicos. Este trabajo ha creado una base de datos de conocimiento de relaciones de fenotipo D-M a gran escala y con precisión. Esta base de conocimiento resultante se encuentra disponible en línea con formato de texto plano, para ser descargado directamente; pero lamentablemente no tiene ningún tipo de API o interfaz gráfica que permita interactuar mejor con los datos.

También dentro del ámbito de esta tesis se encuentra un trabajo que presenta su estudio acerca de la explotación de las relaciones enfermedad-síntoma a partir del minado masivo de la literatura biomédica, y también publica la construcción de un gráfico de conocimiento biomédico a partir de este tipo de relaciones (Xia et al., 2018), sus autores han enfocado este trabajo en las enfermedades graves. Para la construcción de la base de conocimiento hicieron uso de 15.970.134 registros de citas de MEDLINE/PubMed (títulos y resúmenes), en los cuales se permitieron descubrir asociaciones entre entidades médicas; después se analizaron y caracterizaron las ocurrencias de 8,514 conceptos de enfermedades provenientes de la Ontología de Enfermedades Humanas (Schriml et al., 2012) y 842 conceptos de síntomas de la Ontología de Síntomas (Schriml, 2009) y su relación. Finalmente, la base de conocimiento se centró en enfermedades complejas o graves, o enfermedades comunes con afecciones más graves, que permitió crear un sistema de diagnóstico de enfoque de inferencia bayesiana con supuestos ingenuos de independencia más preciso de las enfermedades graves. Sus datos no se encuentran disponibles públicamente.

2.3 Enfoques de extracción y generación de conocimiento biomédico

En el trabajo de (Sudeshna, Bhanumathi & Hamlin, 2017) se describen los esfuerzos de sus autores para mejorar el conocimiento, el tiempo de identificación y de predicción de las enfermedades del corazón; ya que ellos han apreciado que los médicos tienen que observar el estado de salud y también los hábitos alimenticios de los pacientes, y así también el paciente tiene que realizar distintas pruebas, convergiendo todo en un proceso muy largo. Según esta investigación, este tipo de enfermedades son la causa principal de muerte en todo el mundo y por ello han centrado en su estudio. En su publicación desarrollan una técnica que predice la enfermedad y la probabilidad de la enfermedad, así como la probabilidad de la enfermedad y la probabilidad de sufrir ataques; todo ello en base al análisis del conjunto de datos del paciente. El análisis de la medicina se realiza mediante pruebas, utilizando la técnica de minería de datos como, las técnicas de clasificación *Support Vector Machine* (SVM), *Sequential Minimal Optimization* (SMO) o de agrupamiento (*clustering*, en inglés) que ayudan a segmentar los datos en grupos centralizados; el agrupamiento de los datos se realiza en pequeños grupos basados en las similitudes. El proceso se logra mediante el análisis del estado de salud del corazón del paciente, la formulación de preguntas de las enfermedades del corazón, y la recopilación de evidencias de análisis sobre las enfermedades del corazón. Este sistema desarrollado también reconoce los síntomas, y en base a los síntomas la enfermedad se identifica y se identifica también el fármaco para esa enfermedad y sobre la base del conjunto de datos del paciente se identifica el mejor fármaco que se establecerá sus condiciones de salud, para finalmente, sugerir al paciente el mejor fármaco. No se encuentran publicados sus datos, ni ninguna interfaz de usuario.

Finalmente, una indiscutible fuente de conocimiento, son los registros electrónicos de pacientes (EHR, por sus siglas en inglés). Estos tipos de conocimiento son ricos en textos, en donde están escritos casi todos los procesos de toma de decisiones del personal médico de los hospitales. Por lo tanto, minarlos supone adquirir conocimiento relativo a la toma de decisiones y el diagnóstico de las enfermedades. El trabajo de investigación de (Tsumoto et al., 2017) se centra en la minería de textos para los resúmenes de alta, que incluyen la explicación compacta del ingreso del paciente; un registro de sus quejas; hallazgos físicos; resultados de laboratorio y estudios radiográficos durante la hospitalización; una lista de los cambios en sus medicamentos al ser dada de alta; y recomendaciones para el seguimiento de la atención. El proceso de minería de textos desarrollados por los autores consiste en cuatro procesos diferentes: en primer lugar, aplicar análisis morfológicos a un conjunto de resúmenes, para generar una matriz de términos. En segundo lugar, la aplicación del análisis correspondiente a las etiquetas de clasificación y a la matriz de términos, para generar dos coordenadas dimensionales. Mediante la medición de las distancias entre las categorías y los puntos asignados, generan la clasificación de las palabras clave. Y a continuación, seleccionan las palabras clave como atributos según el rango, y se generan ejemplos de formación para los clasificadores. Por último, aplican métodos de aprendizaje a

2.4 Redes de enfermedades

los ejemplos de formación. Para la tarea de minería de texto los autores utilizaron, árboles de decisión, *Support Vector Machine* (SVM), *Binned Nearest Neighbors* (BNN) y aprendizaje profundo (*Deep Learning* en inglés). La validación de su sistema resultó en que el bosque aleatorio obtuvo el mejor rendimiento y el segundo mejor fue el de aprendizaje profundo con una pequeña diferencia, pero los métodos de árbol de decisión con muchas palabras clave obtuvieron un rendimiento sólo un poco peor que los métodos de red neuronal o de aprendizaje profundo. Sus datos no se encuentran disponibles.

2.4 Redes de enfermedades

El enfoque de redes de enfermedades es una parte indispensable para este trabajo de tesis ya que será uno de los medios que se usaran para llevar a cabo el análisis propuesto. A continuación, se describen cada uno de los trabajos relacionados a esta tesis en donde se hacen uso del enfoque de redes de enfermedades para llevar a cabo sus investigaciones.

The human disease network (HDN, por sus siglas en inglés), es considerada la primera red compleja de enfermedades que ha sido publicada por (Goh et al., 2007). Actualmente este trabajo es una referencia y base para los estudios posteriormente desarrollados en este campo. La HDN está basada en un grafo que se encuentra dividido en dos y denominado *Diseasome*, es decir, este grafo bipartito relaciona un conjunto de nodos de enfermedades disjunto de otro conjunto de nodos de genes. El primer conjunto de nodos es referente a todas las enfermedades genéticas conocidas y el segundo, por todos los genes asociados a cada enfermedad. La conexión enfermedad – gen se establece si la mutación de dicho gen está implicada en la enfermedad.

El conjunto de enfermedades, genes implicados en enfermedad y sus asociaciones fueron obtenidos de la fuente OMIM (Hamosh et al., 2005).

A partir de la *Diseasome* se han creado dos redes biológicamente relevantes: la HDN y la *Disease Gene Network* (DGN, por sus siglas en inglés). En la HDN, los nodos son las enfermedades, y se encuentran conectadas entre sí sólo en el caso de que compartan al menos un gen en el cual las mutaciones estén asociadas a ambas enfermedades. En la DGN, los nodos son los genes asociados a las enfermedades, y éstos se encuentran relacionados entre sí sólo si se encuentran asociados con la misma enfermedad.

En la HDN se muestran múltiples conexiones no aleatorias, donde se observan agrupaciones de datos (*clústeres*) formados entre las distintas enfermedades individuales y sus clases, según la fisiopatología. En el estudio revelan que estos *clústeres* indican a su vez orígenes genéticos comunes entre ellas. En el mismo sentido,

2.4 Redes de enfermedades

en la DGN se observa que, mientras que la mayoría de los genes se encuentran implicados en un bajo número de enfermedades, otros son causantes de más de diez enfermedades, dando de nuevo lugar a *clústeres*.

Debido al alto número de genes implicados en las enfermedades, en esta red de enfermedades se muestra un *clúster* voluminoso que es relativo a la enfermedad del cáncer y las enfermedades asociadas. Se encuentra altamente interconectado gracias a diversos oncogenes y/o genes supresores de tumores comunes; es posible que el cáncer sea de las enfermedades más conectadas ya que tienen diversos subtipos. En otro caso, enfermedades complejas como diabetes o la obesidad cumplen el rol de conectores, uniendo enfermedades de diferentes clases (Wang, Gulbahce & Yu, 2011).

Los autores llegaron a determinar a través del análisis en conjunto de estas redes de enfermedades que, los genes que están involucrados en una misma enfermedad se agrupan en módulos de enfermedad, debido a que: **i)** tienden a mostrar elevados niveles de co-expresión, **ii)** tienden a ser expresados juntos en tejidos específicos, **iii)** tienden a compartir la clasificación en términos de la GO (Ashburner et al., 2000), **iv)** sus productos muestran una fuerte tendencia a interaccionar entre ellos mediante *Protein-Protein Interaction* (PPIs), y **v)** exhiben una expresión sincronizada como grupo.

Finalmente, el modelo propuesto y desarrollado en (Goh et al., 2007), sugiere que una enfermedad representa la perturbación o ruptura de un módulo funcional específico, ya que las redes celulares son modulares y están constituidas por grupos de proteínas altamente conectadas responsables de las funciones celulares específicas. Por lo tanto, esta correlación entre enfermedad y módulo funcional puede ayudar a identificar que genes están involucrados en la misma función celular o módulo de la red (Prieto Santamaría, 2018).

Y es en este contexto, con el desarrollo de este tipo de redes nace la *Human symptoms–disease network* (HSDN, por sus siglas en inglés), que fue presentada en (Zhou et al., 2014b), en donde se creó una HDN pero usando los síntomas como medida principal de similitud para conectar enfermedades. Siguiendo el principio de que, si dos enfermedades comparten síntomas, es posible pensar que compartan alguna característica más, en un sentido diferente al fenotípico.

En la HSDN para obtener información de relaciones síntoma–enfermedad hizo uso de técnicas de minería de datos sobre la base de datos bibliográfica PubMed (Trueba-Gómez & Estrada-Lorenzo, 2010; Canese & Weis, 2013). Su proceso ha consistido en utilizar una API (Sayers, 2018) también desarrollada por la NLM, que recoge los identificadores de los artículos PubMed cuyas palabras clave incluyan cualquiera de los términos de enfermedad o síntoma humano definidos por

2.4 Redes de enfermedades

MeSH (Lipscomb, 2000), el vocabulario controlado utilizado para indexar los artículos PubMed. La asociación entre los síntomas y enfermedades se realiza en base a sus co-ocurrencias en los campos de metadatos MeSH en PubMed.

Los autores de la HSDN también integraron bases de datos de genotipo-fenotipo y de PPIs, determinando que la compartición de síntomas indica tanto genes como PPIs compartidos entre enfermedades.

Adicionalmente, se llevó a cabo un análisis sobre las similitudes entre las enfermedades en base a sus síntomas mediante una medida frecuentemente utilizada en minería de textos y en la literatura biomédica: la similitud coseno, que es la similitud entre dos vectores de síntomas A y B de dos enfermedades distintas. La fórmula para calcularla es:

$$sim_{cosine}(A, B) = \frac{|A \cap B|}{|A|^{\frac{1}{2}} + |B|^{\frac{1}{2}}}$$

En donde un valor de similitud coseno de 0 apunta que no hay síntomas compartidos y un valor de 1 indica síntomas idénticos entre las dos enfermedades. El grosor de las aristas varia con respecto al índice de similitud.

En (Zhou et al., 2014a) se presentó una red de fenotipos clínicos. La red tiene como objetivo ayudar a comprender el mecanismo subyacente de la medicina personalizada de la medicina tradicional china (MTC) y proponer una nueva perspectiva para el perfeccionamiento de las habilidades clínicas individualizadas de la MTC. La MTC investiga el diagnóstico clínico y las regularidades de tratamiento en un esquema típico de medicina personalizada, indicando que los pacientes individualizados con las mismas enfermedades obtendrían un diagnóstico distinto y un tratamiento óptimo de diferentes médicos de la MTC. Este principio ha sido reconocido e inculcado por los practicantes clínicos de la MTC durante miles de años. Y a pesar de ello, los mecanismos subyacentes de la medicina personalizada de la MTC no se han investigado plenamente y siguen siendo desconocidos.

Los autores de (Zhou et al., 2014a) citan, que así como las enfermedades los fenotipos clínicos y los síntomas son entidades complicadas de estudiar y generalmente sucede simultáneamente que tienen en común los mecanismos moleculares subyacentes. En la MTC los fenotipos clínicos son los únicos elementos usados para el diagnóstico y el tratamiento, la sintomatología y la sinología regular que mantienen contribuye a como los factores principales al principio de diagnóstico y tratamiento personalizados de la MTC. Por lo tanto, en lugar de centrarse en captar las manifestaciones específicas de la enfermedad como hacen los médicos biomédicos modernos, los profesionales de la MTC aplican habilidades de observación sistemática para detectar los síntomas y signos de los pacientes, y generan descripciones elaboradas de las manifestaciones sintomáticas de los pacientes.

2.4 Redes de enfermedades

Esta investigación se basa en 246 registros ambulatorios bien diseñados sobre insomnio, y al evaluar los sesgos personales de la observación de las manifestaciones y las preferencias de las prescripciones de hierbas, observamos similitudes significativas entre las prescripciones de cada hierba y las similitudes de los síntomas entre cada encuentro. Estas similitudes al igual que en la HSDN, han sido calculadas utilizando la similitud coseno.

Usando las características de los síntomas del conjunto de datos sobre el insomnio de dos estudios relacionados previos, se calculó la co-ocurrencia de los síntomas en los pacientes y se construyó una red con los síntomas como nodos y las co-ocurrencias de los síntomas como aristas. La red de síntomas obtenida está formada por 442 nodos y 5.671 enlaces o relaciones. En ella se observa que la longitud media del camino más corto de la red de síntomas es de 2,16 y su diámetro es de 3, por lo que tiene una estructura de mundo pequeño en la que la mayoría de los síntomas están estrechamente relacionados entre sí. Los nodos con los 5 más altos grados son la disforia (*dysphoria*, 222), el mareo (*dizziness*, 205), la lengua roja (*tongue*, 197), el pelaje amarillo fino (*thin yellow fur*, 163) y la irritabilidad (*irritable*, 157); que también forman parte de la sintomatología frecuente en el insomnio. Además, el coeficiente medio de agrupación de esta red es de 0,789. Esto es mucho mayor que una red aleatoria, lo que indica que la mayoría de los síntomas se agrupan en grupos densos. Esto ofrece una información perspicaz para el diagnóstico personalizado de la MTC: los síntomas de los pacientes detectados por un médico pueden concebir síntomas relacionados ocultos, que serían enfatizados por otro médico para un diagnóstico personalizado.

DisGeNET es una plataforma de conocimiento que contiene un catálogo público completo de genes y variantes asociados a enfermedades humanas (Bauer-Mehren et al., 2011; Piñero et al., 2017) creada en 2010. Aunque presentada previamente en la sección donde se analizaban los recursos de información biomédica, merece ser comentada nuevamente en esta sección dadas sus características. DisGeNET contiene casi la totalidad de las enfermedades humanas, incluyendo enfermedades Mendelianas, complejas, ambientales y raras, así como características asociadas a ellas. DisGeNET, para extraer información biomédica referente a relaciones genotipo-fenotipo hace uso de múltiples bases o conjunto de datos como, UniProt (Wu et al., 2006), PsyGeNET (Gutiérrez-Sacristán et al., 2015), ClinVar (Landrum et al., 2014) u Orphanet (Weinreich et al., 2008); de recursos textuales, aptos para la minería de textos como, dbSNP (Sherry et al., 2001), EXAC (Karczewski et al., 2017) o 1000 Genomes Project (Siva, 2008), en este sentido DisGeNET hace uso de diversas técnicas de NLP para explotar los textos de contenido biomédico; también de vocabularios médicos como UMLS (Bodenreider, 2004), OMIM (Hamosh et al., 2005) o MeSH (Lipscomb, 2000).

DisGeNET ofrece variedad de anotaciones y diferentes métricas para reforzar, priorizar y facilitar la recuperación y la interpretación de esta información. En su

2.4 Redes de enfermedades

compendio refunde datos de repositorios curados por expertos, de catálogos GWAS⁵⁹, de modelos animales y de literatura científica. Los datos de DisGeNET están homogéneamente anotados con vocabularios controlados y ontologías impulsadas por la comunidad científica. DisGeNET cuenta con alrededor de 20.370 enfermedades y desórdenes. La distribución de éstas se encuentra disponible en el siguiente enlace⁶⁰ (Prieto Santamaría, 2018).

Actualmente, a principios de 2020 se encuentra en la versión seis que contiene 628.685 asociaciones gen–enfermedad, entre 17.549 genes y 24.166 enfermedades y características y 210.498 asociaciones variante – enfermedad, entre 117.337 variantes y 10.358 enfermedades, fenotipos y tratamientos⁶¹. Cuenta con una interfaz Web en la que se permite realizar búsquedas y visualizaciones en forma de grafos. Desde la página Web se puede acceder al DisGeNET–RDF Linked Dataset que permite consumir sus datos por medio de consultas SPARQL y también permite integrar sus datos a otros conjuntos de datos RDF externos. Por último, también se puede acceder a sus datos por medio de la descarga de ficheros o por medio del uso de un API.

Otra red de enfermedades relevante es DisNor: *a disease network open resource*, desarrollada con el objetivo de explotar el gran volumen de datos sobre la identificación de genes asociados a enfermedades para ensamblar vías de enfermedad inferidas. DisNor pretende ser un nuevo esfuerzo de curación de datos médicos destinado a conectar los genes de enfermedades mal caracterizadas a la red de señalización global en *The SIGnaling Network Open Resource* (SIGNOR, por sus siglas en inglés) (Perfetto et al., 2016) (Lo Surdo et al., 2018).

La creación de esta red se ha realizado mediante la combinación de los datos de asociación genético–enfermedad (GDA) que se encuentran anotados en el recurso DisGeNET (Piñero et al., 2015) con un nuevo esfuerzo de curación destinado a poblar la base de datos SIGNOR con interacciones causales relacionadas con los genes de la enfermedad con la mayor cobertura posible. Para cada enfermedad curada en DisGeNET, DisNor vincula los genes de la enfermedad mediante relaciones causales anotadas manualmente y ofrece una visualización intuitiva de las “vías de acceso” inferidas a diferentes niveles de complejidad.

DisNor al ser un recurso Web de libre acceso, todos sus usuarios tienen la oportunidad de explorar más de 3.700 vías de enfermedades, uniendo cerca de 2.600 genes de enfermedades. cada que se consulta una enfermedad en su visualizador se proyecta una red en la que se indican las conexiones con sus productos genéticos. Cuenta con un sencillo motor de búsqueda y un potente método que permite generar una red por enfermedad, incluidas enfermedades raras y cánceres.

⁵⁹ <https://www.ebi.ac.uk/gwas/>

⁶⁰ <https://www.disgenet.org/dbinfo>

⁶¹ <http://www.disgenet.org/web/DisGeNET/menu/dbinfo#sources>

2.4 Redes de enfermedades

Las redes de genes se han explotado mucho gracias a que se ha demostrado que son útiles en diversos aspectos de la investigación de enfermedades, con estrategias basadas en redes desarrolladas para generar hipótesis sobre asociaciones genético-enfermedad-fármacos. En este contexto se ha mostrado relevante la capacidad de predicción y organización de los genes más relevantes para una enfermedad determinada. Y con ese sentido es como nació HumanNet v2 que integra diversos tipos de datos ómicos mediante un marco estadístico bayesiano con el que se creó un método que tiene la capacidad de recuperar genes de enfermedades (Hwang et al., 2019). HumanNet v2 es una base de datos de redes de genes humanos actualizada que comprende una jerarquía de redes de genes humanos que aseguran sus desarrolladores que permite una incorporación más flexible de la información de la red.

La red de HumanNet v2 se basa en ~300.000 artículos de texto completos de PubMed Central, mientras que ~750.000 resúmenes de Medline. También se ha hecho uso de otros tipos de bases de datos como KEGG (Kanehisa & Goto, 2000), BioCarta (Rouillard et al., 2016), Reactome (Fabregat et al., 2018), entre otras. El resultado es que el número de PPI no redundantes en la HumanNet v2 es de 158.499 (que conectan 15.352 genes). HumanNet v2 incluye redes basadas en asociaciones de contextos genómicos (GN y PG). Utilizando 1.748 genomas procariotas (1.626 bacterianos y 122 arqueas) y 996 metagenomas (754 humanos y 242 oceánicos) (27,28) para analizar las asociaciones de contexto genómico para HumanNet v2.

La HumanNet v2 desenvuelve un buen papel durante la clasificación de conjuntos de genes vinculados a enfermedades con un mínimo de sesgos dependientes de la literatura. En este sentido se observó que la incorporación de las interacciones proteína-proteína de los organismos modelo no mejora notablemente las predicciones de los genes de las enfermedades, lo que permitió a los autores sugerir que muchas de las asociaciones de genes de enfermedades se capturan directamente en los conjuntos de datos obtenidos de los humanos.

Finalmente, la HumanNet v2 cuenta con una interfaz de usuario interactiva y potente para el análisis de la red de enfermedades y con la opción de descargar sus datos en un fichero.

Como se comentó previamente, el reposicionamiento de fármacos es un tema de mucho interés para investigadores y empresarios ya que pretende entre muchas más cosas, reducir los costes y el tiempo de desarrollo y producción de los medicamentos tratando de buscar nuevos usos a fármacos existentes. Este concepto es ampliamente explotado en los enfoques de redes de enfermedades, ya que permiten relacionar enfermedades en base a sus diferentes características, y en base a estas poder crear posibles estrategias de reposicionamiento. En esta línea se encuentra este trabajo relacionado publicado en (Himmelstein et al., 2017), donde hablan de

2.4 Redes de enfermedades

la capacidad de predecir computacionalmente si un compuesto trata una enfermedad mejoraría la economía y la tasa de éxito de la aprobación de la droga. Los autores presentan en esta publicación el proyecto Rephetio para modelar sistemáticamente la eficacia de los medicamentos en base a 755 tratamientos existentes. Para ello han desarrollado una red llamada Hetionet, que es una red integradora que codifica el conocimiento de millones de estudios biomédicos. Hetionet consiste en 47.031 nodos de 11 tipos y 2.250.197 relaciones de 24 tipos. En ella se integraron datos de 29 recursos públicos para conectar compuestos, enfermedades, genes, anatomías, vías, procesos biológicos, funciones moleculares, componentes celulares, clases farmacológicas, efectos secundarios y síntomas.

Creada la red, se identificaron patrones de red que destacan los tratamientos de los que no lo son. Después, se calculó la probabilidad de tratamiento para 209.168 pares de enfermedades compuestas. Las predicciones desarrolladas se validaron en dos conjuntos de tratamiento externo y proporcionaron conocimientos farmacológicos sobre la epilepsia, sugiriendo que ayudarán a priorizar los candidatos a la readaptación de los medicamentos.

Un tipo de red de enfermedad interesante, es aquella que está formada por múltiples capas. En el artículo titulado *The multiplex network of human diseases* publicado en (Halu et al., 2019) se afronta el reto de desenredar la compleja interacción entre fenotipo y genotipo, ya que sus autores creen que es crucial para la caracterización y sub tipificación efectiva de las enfermedades. En su publicación construyen una red múltiple de 779 enfermedades humanas, que consiste en una capa basada en el genotipo y una capa basada en el fenotipo.

La red bipartita de genes de enfermedades se ha construido a partir del conocido conjunto de datos de OMIM (Hamosh et al., 2005). La versión de OMIM utilizada en el estudio contiene información sobre 3.369 genes y 4.239 enfermedades, para formar un total de 5.308 aristas. La red bipartita de interacciones genes – enfermedades se proyectó sobre el componente de enfermedades para construir la siguiente capa de la red, la de enfermedades–enfermedades que está basada en el genotipo. De la misma forma, la red bipartita de enfermedades y síntomas se construye a partir del conocido conjunto de datos de la HPO (Robinson et al., 2008). Esta ontología proporciona un conjunto estructurado, exhaustivo y bien definido de 10.088 términos que describen las anormalidades del fenotipo humano y 13.326 relaciones de subclase entre las clases de HPO. El resultado ha sido que se han encontrado 6.662 enfermedades con relaciones a 11.052 síntomas, para un total de 100.281 aristas. La red bipartita de interacciones síntoma – enfermedad también se proyectó sobre el componente de la enfermedad para construir la capa de la red enfermedad–enfermedad basada en el fenotipo.

Así se mostró que, las enfermedades con componentes genéticos comunes tienden a compartir síntomas, y se descubrió cómo la información del fenotipo ayuda

2.4 Redes de enfermedades

a aumentar la información del genotipo. Además, se realizó una clasificación de las enfermedades que considera sus fundamentos moleculares junto con sus manifestaciones clínicas que se ha basado en la similitud tanto fenotípica como genética y creado nuevas conexiones que mostraron que las comunidades de enfermedades múltiples propuestas son capaces de captar la base molecular de las conexiones entre las enfermedades, incluso si no se han asociado entre sí simplemente por sus manifestaciones clínicas. También se detectaron grupos cohesivos de enfermedades que tienen una gran similitud intragrupal tanto a nivel molecular como fenotípico. Inspeccionando las comunidades de enfermedades, los autores demostraron las vías subyacentes que conectan las enfermedades mecánicamente. Así mismo, se observa que los trastornos mono génicos agrupados con enfermedades complejas para las que aumentan el factor de riesgo. Finalmente, en este estudio se propone asociaciones de enfermedades potencialmente nuevas que surgen como una característica única del flujo de información dentro y a través de las dos capas.

La baja tasa de prevalencia de las enfermedades huérfanas (OD) requiere esfuerzos especiales combinados para mejorar el diagnóstico, la prevención y el descubrimiento de nuevas estrategias terapéuticas. Esta motivación llevó a los autores de (Zhang et al., 2011) a identificar e investigar las relaciones basadas en genes compartidos o características funcionales compartidas, hemos realizado un análisis global basado en la bioinformática de todas las enfermedades huérfanas con genes mutantes causantes de enfermedades conocidas. Comenzando con una red bipartita de genes mutantes conocidos causantes de la sobredosis y de la sobredosis y utilizando el interactoma de la proteína humana, primero construimos y analizamos topológicamente tres redes: la red de enfermedades huérfanas, la red de genes mutantes causantes de enfermedades huérfanas y el interactoma de genes mutantes causantes de enfermedades huérfanas. Nuestros resultados demuestran que, en contraste con los genes mutantes comunes causantes de enfermedades que son predominantemente no esenciales, la mayoría de los genes mutantes huérfanos causantes de enfermedades son esenciales. Para confirmar este hallazgo, encontramos que los genes mutantes causantes de sobredosis son topológicamente importantes en el interactoma de la proteína y se expresan de forma ubicua. Además, el análisis de enriquecimiento funcional de los genes en los que las mutaciones causan sobredosis muestra que la mayoría de ellos provocan una muerte prematura o son letales en los modelos ortólogos de genes de ratón *knockout* o ratón KO⁶², un ratón modificado por ingeniería genética para que uno o más de sus genes estén inactivados mediante una técnica llamada bloqueo de genes. Para hacer frente a las limitaciones de las redes de enfermedades tradicionales basadas en genes, también construimos y analizamos redes de DO sobre la base de características enriquecidas compartidas (procesos biológicos, componentes celulares, vías, fenotipos y citas bibliográficas). Analizando estas redes de DO vinculadas funcionalmente, identificamos varias relaciones adicionales de DO-OD que son

⁶² <https://www.genoway.com/services/customized-mouse/knockout-models/overview.htm>

2.4 Redes de enfermedades

tanto fenotípicamente similares como fenotípicamente diversas. Sorprendentemente, observamos que el cableado de las redes de DO basadas en el gen y otras basadas en características son en gran medida diferentes; esto sugiere que la relación entre las DO no puede ser captada completamente por la red basada en el gen por sí sola.

(Menche et al., 2015) es otro trabajo relacionado que también enfoca sus esfuerzos en la problemática de que el mapa del interactoma se encuentra incompleto. Si este estuviera completo supondría un tremendo impacto en la capacidad de entender los fundamentos moleculares de las enfermedades humanas; actualmente es imposible evaluar con precisión cuán lejos está un determinado módulo de enfermedad de ser completado. De esta forma, en (Menche et al., 2015) se describe este problema y propone una solución con un enfoque de red que ha permitido demostrar que, a pesar del carácter incompleto, el interactoma disponible tiene suficiente cobertura para aplicar un enfoque sistemático de las enfermedades humanas basado en redes. Esta investigación además ofrece evidencias cuantitativas para la identificación para algunos módulos de enfermedades, al tiempo que determinan que, en el caso de otras enfermedades, la condición de identificación no se cumple todavía con el nivel actual de incompletitud del interactoma. El último aporte que se presenta en esta investigación es, que han llegado a determinar que la posición relativa basada en el interactoma de dos módulos de enfermedades es un fuerte predictor de su similitud biológica y fenotípica.

Para llevar a cabo la investigación en (Menche et al., 2015) se hizo uso de diferentes fuentes de información, primero, compilaron 141.296 interacciones físicas entre 13.460 proteínas documentadas experimentalmente en las células humanas, incluyendo interacciones proteína-proteína y reguladoras, interacciones en la vía metabólica e interacciones cinasa-sustrato. También, obtuvieron 299 enfermedades definidas por el vocabulario biomédico MeSH (Lipscomb, 2000) con el sesgo de que tengan al menos 20 genes asociados en la base de datos OMIM (Hamosh et al., 2005), y además obtuvieron información del *genom--wide association studies* (GWAS, por sus siglas en inglés) (Ramos et al., 2014), que incluyen 2436 proteínas asociadas a enfermedades.

Después de filtrar las enfermedades con al menos 20 genes asociados y los genes para los que se tienen información de interacción, se obtuvieron 299 enfermedades y 3.173 genes asociados.

Para llevar a cabo el análisis de la similitud entre genes y enfermedades, los autores utilizaron: **a)** anotaciones de la GO (Ashburner et al., 2000), **b)** datos de expresión génica específicos de los tejidos, **c)** asociaciones de síntomas y enfermedades, **d)** datos de comorbilidades, y **e)** anotaciones de rutas de la base de datos MSigDB (Liberzon et al., 2011).

2.5 Enfoques de similitud entre enfermedades

Por su parte (Chen et al., 2015) describe una realidad latente: las redes genéticas están fácilmente disponibles, a diferencia de las redes de fenotipo de la enfermedad que siguen siendo en gran parte incompletas. Es por ello que en (Chen et al., 2015) se presentó la construcción de una Red de Manifestación de Enfermedades (DMN, por sus siglas en inglés) a gran escala con 50.543 relaciones semánticas de manifestación de enfermedades altamente precisas porque son obtenidas del lenguaje UMLS. Esta red de fenotipos propuesta contiene 2.305 nodos y 373.527 relaciones ponderadas para así representar las similitudes fenotípicas de las enfermedades.

A través de un análisis comparativo entre la DMN y la HDN y una red de fenotipos de enfermedades de descubrimiento de genes, llamada mimMiner (van Driel et al., 2006). El primer análisis comparativo elaborado ha sido contra mimMiner, en él se ha encontrado que las similitudes de manifestación en el DMN tienen correlaciones con medidas cuantificadas de las asociaciones genéticas de la distensión. En la segunda comparativa se han comparado los bordes y las estructuras comunitarias de la DMN con la HDN, en ella se muestra que el número de bordes comunes entre la DMN y la HDN es significativamente mayor que la distribución aleatoria. Se mostró que el mimMiner también contiene 520 bordes comunes con la HDN basada en OMIM y 14 con la HDN basada en GWAS⁶³. Sin embargo, DMN y mimMiner comparten diferentes conexiones de enfermedades con las HDN: 76 de las 278 (27%) superposiciones de bordes entre DMN y HDN basada en OMIM no aparecen en el mimMiner, y 5 de las 6 superposiciones de bordes entre DMN y HDN basada en GWAS no aparecen en el mimMiner.

2.5 Enfoques de similitud entre enfermedades

Actualmente existen dos enfoques bien diferenciados para el cálculo de similitudes de enfermedades: aquellos que lo realizan según los rasgos genéticos y aquellos que los tratan según los rasgos fenotípicos. En esta tesis se enfocan todos los esfuerzos a los cálculos de similitud de enfermedades según sus características fenotípicas, información que se encuentra almacena en el ya mencionado conjunto de enfermedades y de sus manifestaciones fenotípicas.

En la actualidad, la medición cuantitativa de la similitud entre enfermedades es una cuestión necesaria para ayudar a revelar la psicofisiología común y a mejorar los sistemas de toma de decisiones clínicas, comprender mejor las enfermedades humanas y clasificarlas con mayor precisión (Deng et al., 2019).

En el estudio de (Valle et al., 2018) ha evaluado a la Wikipedia como fuente de información fenotípica de enfermedades comparándola con la base de datos PubMed (Trueba-Gómez & Estrada-Lorenzo, 2010). Los autores de este artículo

⁶³ <https://www.ebi.ac.uk/gwas/docs/about>

2.5 Enfoques de similitud entre enfermedades

utilizaron las tres medidas de similitud basadas en la construcción de vectores de características fenotípicas por cada enfermedad: el índice *Coseno*, el índice *Jaccard* y el índice *Dice*.

Ahora se describirán algunos de los trabajos relacionados en los cuales se han implementado alguno de los coeficientes de similitud anteriormente descritos y algunos más novedosos que pretenden mejorar estos tipos de algoritmos.

En diferentes trabajos como, (Zhou et al., 2014b; Yang et al., 2019) se ha calculado la similitud *Coseno* para comparar los vectores de síntomas entre dos enfermedades. Además del índice coseno, en otras investigaciones también han utilizado el índice de *Jaccard* para calcular la similitud fenotípica (Pesquita et al., 2008; Hoehndorf, Schofield & Gkoutos, 2011, 2015; Peng et al., 2016).

En el trabajo de (Hoehndorf, Schofield & Gkoutos, 2015) se ha desarrollado un enfoque de minería de texto semántico para identificar los fenotipos (signos y síntomas) asociados con más de 6.000 enfermedades. En este estudio se han utilizado datos de la HPO (Robinson et al., 2008) y de la *Mammalian Phenotype Ontology* (MP, por sus siglas en inglés) (Eppig et al., 2015) para los experimentos. El análisis realizado demostró que se pueden identificar correctamente los genes conocidos asociados a las enfermedades en ratones y humanos con gran precisión. Utilizando una medida de similitud fenotípica, los autores crearon una red de enfermedades humanas en la que se agrupan las enfermedades que tienen signos y síntomas similares. La red fue usada para identificar enfermedades estrechamente relacionadas sobre la base de fundamentos etiológicos, anatómicos y fisiológicos comunes.

En (Hoehndorf, Schofield & Gkoutos, 2015) se llevó a cabo el cálculo de la similitud semántica de enfermedades utilizando el sistema PhenomeNET (Hoehndorf, Schofield & Gkoutos, 2011). PhenomeNET usa una ontología de fenotipo multiespecie que integra las clases de ontologías de fenotipo específicas de la especie en una única estructura en la que las clases se relacionan basándose en sus definiciones formales.

Por otro lado para comparar conjuntos de fenotipos, se utilizó la medida *simGIC* (Pesquita et al., 2008) basada en conjuntos; esta medida se basa en el índice de similitud *Jaccard* ponderado por el contenido de información de una clase dentro del corpus que consiste en modelos de ratones y enfermedades:

$$simGIC(P, R) = \frac{\sum_{x \in Cl(R) \cap Cl(P)} IC(x)}{\sum_{x \in Cl(R) \cup Cl(P)} IC(y)}$$

Donde $Cl(X)$ es el conjunto más pequeño que contiene X y que se cierra contra la relación de superclases. $IC(x)$ es el contenido de información de una clase dentro del corpus de modelos y enfermedades de ratones.

2.5 Enfoques de similitud entre enfermedades

Por otra parte (Peng et al., 2016) propone una nueva medida de similitud llamada PhenoSim. Esta medida presentada se distingue de otras ya que toma en cuenta las propiedades únicas de la HPO (Robinson et al., 2008). Los autores de PhenoSim sugieren que en comparación con otras medidas de similitud: i) PhenoSim es el primer enfoque de similitud semántica que está especialmente optimizado para HPO; ii) que se ha desarrollado un nuevo Contenido de Información (CI) con restricciones de trayectoria para calcular la similitud entre dos términos de HPO; iii) PhenoSim construye una red de fenotipos y explota un método basado en PageRank para modelar los ruidos en el conjunto de datos del fenotipo del paciente.

Todos los datos utilizados en (Peng et al., 2016) han sido obtenidos de la HPO. PhenoSim está constituido por cuatro pasos: 1) construcción de la red de fenotipos, 2) reducción del ruido de la red de fenotipos, 3) medición de la similitud de fenotipo, y 4) calcular la similitud del conjunto de fenotipos. En el primer paso, construyeron una red de fenotipos basada en la calificación del índice Jaccard por pares entre los conjuntos de genes de dos fenotipos. En el segundo paso, para la reducción del ruido usaron el algoritmo del PageRank. En el tercer paso, fue donde se definió una nueva medición de similitud semántica del término HPO:

$$sim(t_1, t_2) = \begin{cases} \min(IC(t_1), IC(t_2)) \\ 0 \end{cases}$$

donde $IC(t)$ es el contenido de información del término fenotipo t , definido como $IC(t) = \ln \frac{U}{U_t}$, donde U y U_t son el número de anotaciones asociadas al término raíz y t , respectivamente (también se incluyen las anotaciones asociadas a todos sus descendientes). Si t_1 y t_2 son alcanzables, la similitud es el mínimo de su contenido informativo. Si t_1 y t_2 son inalcanzables, la similitud es 0.

DisSim: an online system for exploring significant similar diseases and exhibiting potential therapeutic drugs (Cheng et al., 2016).

Ahora en el trabajo de (Traverso-Ribón & Vidal, 2018) se ha desarrollado una medida de similitud semántica basada en el aprendizaje máquina y las características de la entidad: atributos, vecinos, jerarquía de clases o grados de asociación. GARUM (*GrAph entity Regression sUpported similarity Measure*) una medida de similitud semántica para los grafos de conocimiento. Esta medida se basa en las similitudes de las características de las entidades y calcula los valores de similitud considerando varias características de la entidad de forma simulada.

GARUM fue evaluada sobre entidades de tres diferentes grafos de conocimiento: El primer gráfico de conocimiento describe artículos de noticias anotados con entidades de DBpedia (Auer et al., 2007); y los otros dos gráficos restantes describen proteínas anotadas en la OG (Ashburner et al., 2000). GARUM se compara con medidas de similitud de última generación con el objetivo de determinar

2.6 Conclusiones

si los valores de similitud de GARUM están más correlacionados con los patrones de oro.

GARUM considera cuatro medidas de similitud individuales: Similitud jerárquica, similitud de vecindario, información compartida y atributos. Por lo tanto, GARUM se define en términos de una función α :

$$GARUM(e_1, e_2) = \alpha(Sim_{hier}, Sim_{neigh}, Sim_{shared}, Sim_{attr})$$

Dependiendo del tipo de regresión, α puede ser una combinación lineal o no lineal de los predictores. En los dos casos e independientemente del algoritmo de regresión implementado, α se calcula minimizando una función de pérdida, que es el error cuadrático medio (MSE) y se define de la siguiente manera:

$$MSE = \frac{1}{n} \sum_{i=1}^n (\hat{Y}_i - Y_i)^2$$

Y es un vector de n valores observados, es decir, valores de patrón oro, y $\hat{Y}$ es un vector de n predicciones, en otras palabras, $\hat{Y}$ corresponde a los resultados de la función calculada por α . Por lo tanto, el algoritmo de regresión implementado en GARUM aprende, a partir de un conjunto de datos de capacitación, a combinar las medidas de similitud individuales mediante una función α , de manera que se minimiza la MSE entre los resultados producidos por α y el patrón oro correspondiente.

Finalmente, los resultados han sugerido que: i) el hecho de considerar todas las características de la entidad permite calcular valores de similitud más precisos; ii) GARUM es capaz de superar los enfoques de vanguardia obteniendo valores de correlación más altos; y iii) los enfoques de aprendizaje automático son capaces de predecir las funciones de agregación que superan las funciones manuales definidas por los humanos.

2.6 Conclusiones

El análisis de los trabajos relacionados muestra la importancia de las fuentes de información biomédicas de calidad. Iniciativas como: *The Human Disease Network* (Goh et al., 2007), *Human Symptoms-Disease Network* (Zhou et al., 2014b), MalaCards (Espe, 2018) o DisGeNet (Piñero et al., 2017) son buenos ejemplos que se han valido de datos de fuentes de calidad como PubMed, MeSH, UMLS, entre otras, para desarrollar sus investigaciones y así ayudar a mejorar el conocimiento que se tiene sobre las enfermedades.

2.6 Conclusiones

Los estudios relacionados también permiten centrar esfuerzos en conseguir que los conjuntos de datos médicos consideren tener una lista de enfermedades consistente, completa y lo más homogénea posible. La lista de enfermedades es la columna vertebral del conjunto de datos, por la cual se pueden establecer relaciones enfermedad–enfermedad y tener interconexiones de conocimiento relevante.

Varias contribuciones anteriormente descritas se han aprovechado para recuperar conocimientos de diferentes fuentes estructuradas y no estructuradas, a fin de integrarlos en una única base de datos. Reunir conocimiento de diferentes fuentes conlleva muchos retos pero contribuyen con conocimiento diverso que complementa, mejora y da mucho más valor. A su vez, se ha demostrado que la utilización de esos recursos tiene importantes repercusiones en diferentes áreas, entre ellas la de lograr una mejor comprensión de las enfermedades y sus relaciones.

Además, los recientes trabajos de investigación sobre la representación de las enfermedades y la interrelación (similitud) mediante el uso de redes han proporcionado resultados prometedores, permitiendo a los científicos obtener una mejor comprensión de las enfermedades desde una perspectiva diferente.

También, es importante reconocer que aunque la información de una fuente no se encuentre estructurada y sea totalmente textual no supone un impedimento para descartar dichos tipos de fuentes. Esto es todo lo contrario, ya que actualmente por medio de técnicas modernas como el procesamiento de lenguaje natural (NLP, por sus siglas en inglés) es posible obtener de los textos, información relevante estructurada para que sistemas informáticos puedan valerse de ellos y generar nuevo conocimiento útil.

Los trabajos relacionados también destacan el uso de múltiples fuentes textuales públicas para recopilar información fenotípica e incluir a Wikipedia como principal fuente médica pública. Esto brinda al presente documento de tesis la oportunidad de analizar una fuente controversial, que se encuentra en constante crecimiento y que está en constante actualización. La obtención de conocimiento de fuentes públicas y de acceso libre como lo es Wikipedia representa un importante paso en la investigación que muchas veces se ve limitada porque para acceder a conocimiento relevante es necesario asimilar costes.

Finalmente, la revisión de la literatura ha puesto de manifiesto que las propuestas que se plantean en esta tesis para mejorar el conocimiento sobre las enfermedades están bien enfocadas, por ejemplo: la creación de un conjunto de datos estructurados con información médica proveniente de fuentes de datos públicas textuales, el uso del NLP para obtener información diagnóstica de los textos de las fuentes, analizar y validar la información generada y por último, desarrollar un medio para compartir la información generada y apoyar a otros investigadores con un nuevo conjunto de datos enfermedad–manifestaciones fenotípicas.

2.6 Conclusiones

Capítulo 3 Metodología

En este capítulo se describirá explícitamente la metodología general llevada a cabo en la presente tesis, con especial énfasis en los procesos de extracción y generación del conocimiento de la plataforma DISNET, al representar ésta el núcleo de la presente tesis.

En primer lugar, se hará referencia a las fuentes de información públicas utilizadas para extraer conocimiento tales como Mayo Clinic, PubMed y, especialmente, Wikipedia.

Se analizarán los elementos tenidos en cuenta para catalogar dichas fuentes de información como relevantes describiendo, a continuación, cada una de estas fuentes.

En segundo lugar, se hará referencia a la lista de enfermedades extraídas de cada una de las fuentes de información del punto anterior.

En tercer lugar, se centrará la atención en la extracción y generación de conocimiento, haciendo referencia a los componentes de la arquitectura propuesta las diferentes formas de extracción de ese conocimiento, con especial énfasis en cómo se detectan los términos médicos y como es el proceso de almacenamiento posterior en la base de datos.

Como elemento característico de este trabajo y en este punto, se hará referencia a la constante actualización de la información extraída, enfatizando que no se trata de un proceso estático, sino que se actualiza automática y periódicamente cada 15 días con el fin de capturar los cambios producidos en las fuentes (concretamente en Wikipedia, como se explicará más adelante) a lo largo del tiempo.

Además, se hace referencia a la creación de la API, que permitirá que toda esta información extraída y estructurada pueda ser accesible a todo el público.

Finalmente, a modo de complementación de la información y de cara a realizar un análisis evolutivo del contenido de los artículos Wikipedia, se describen los diferentes enfoques que se han desarrollado con el objetivo de minar la información histórica de Wikipedia.

3.1 Fuentes de información

Los libros, manuales y artículos han dejado de ser los únicos prescriptores de información; Actualmente, debido a la interactividad digital, las ideas, información y el conocimiento, en general, circulan libremente.

En el ámbito de la medicina ocurre exactamente lo mismo: si tradicionalmente era necesario remitirse a libros o manuales de medicina para obtener y extraer el conocimiento, esto ha cambiado drásticamente con la creación de la Web; la información se expande de forma imparable y cualquier usuario con acceso a internet puede acceder a ella.

Aunque los libros y textos médicos contienen una parte vital de la información disponible en éste ámbito, los cuales están escritos por especialistas en la materia, siguiendo siendo usados por una gran parte de los profesionales, esta forma de acceder a la información no deja de estar exenta de inconvenientes, siendo el principal su actualización o su acceso por medios digitales, algo que no siempre es posible.

Actualizar el contenido de un libro supone reescribir el libro: eliminar el contenido cambiante; editar, imprimir y comercializarlo nuevamente. Esto impide que las actualizaciones se encuentren a disposición de los usuarios en cortos periodos de tiempo. A todo esto se le añade la complejidad de estructurar la información contenida para realizar tareas tan importantes como la minería de datos.

Cosa distinta es la información contenida en la Web donde la información y el conocimiento se crea, modifica y estructura, pero sobre todo y lo más importante, es accesible de forma gratuita (en la mayoría de los casos) y de forma casi inmediata para cualquier usuario.

Las fuentes de información médica que se encuentran en la Web pueden ser de varios tipos: **especializada o no especializada**, dependiendo de si la información que contiene se refiere a una única materia o si contiene información de diversa índole (así, por ejemplo, PubMed o Medline Plus solo recogen información médica); también puede ser **textual o gráfica**, dependiendo de si la información se expresa únicamente mediante palabras o si utiliza además imágenes, gráficos, videos, etc.; de **acceso libre o restringido**, dependiendo de si es accesible gratuitamente para cualquier usuario o si, por el contrario, solo puede accederse a ella si se cumplen ciertos requisitos como registro y/o pago.

Pero la clasificación que más interesa para el desarrollo de esta tesis es la que distingue entre información **estructurada y no estructurada**, siendo la primera aquella que se encuentra ordenada en filas, columnas, listas, etc. de forma orde-

3.1 Fuentes de información

nada e identificada, siendo perfectamente apta para ser procesada mediante herramientas automáticas; y siendo de la segunda categoría aquella que no tiene un orden interno, que se basa en un conjunto masivo de objetos que necesitan ser identificados y almacenados de forma ordenada para ser aptos para su procesamiento automático.

Se pueden establecer muchas clasificaciones mas de la tipologías de fuentes de información pública (dependiendo del formato utilizado –Word, pdf, etc–, el lenguaje utilizado –inglés, castellano, francés, italiano, etc.–), pero las referenciadas son las mas importantes.

El 80% de la información contenida en la Web se genera de forma no estructurada, siendo esta forma no estructurada principalmente el lenguaje escrito. Por lo tanto, una de las técnicas para la obtención de conocimiento de estos datos no estructurados es lo que se conoce como procesamiento de lenguaje natural, técnica abordada en esta tesis en la sección 3.2 “Enfoque principal de extracción y generación de conocimiento”.

Del estudio de los trabajos relacionados se ha observado que mientras algunos han construido sus bases de datos con la extracción de información de una única fuente (por ejemplo: *abstracts*/resúmenes de artículos científicos), los trabajos de la HDN (Goh et al., 2007) y la HSDN (Zhou et al., 2014b) han usado principalmente OMIN y PubMed para construir su contenido. Otros trabajos como DisGeNet o MalaCards se han servido de una gran cantidad de fuentes para construir sus bases de datos; MalaCards utiliza hasta 70 fuentes para obtener conocimiento, y su conocimiento no abarca únicamente enfermedades, sino que incluye hasta 14 categorías de conocimiento dentro de cada enfermedad (fármacos, signos y síntomas, genes, publicaciones, etc.).

Es evidente que a mayor cantidad de fuentes médicas analizadas mayor cantidad de conocimiento médico obtendremos, por ello, se suscita la idea de crear un enfoque que permita obtener información fenotípica (dado el objetivo/orientación de la presente tesis) de tantas fuentes médicas como sea posible, teniendo en cuenta que cada una de estas fuentes está creada por usuarios muy diversos, con distintos grados de formación y diferentes estilos. Por lo tanto, el objetivo es intentar que cada fuente de información se complemente y contraste con las demás, creando un conocimiento más extenso y, sobre todo, más fiable.

Los criterios utilizados para seleccionar las fuentes de información que nutrirán el sistema propuesto en esta tesis deben cumplir las siguientes características: Que sean de libre acceso; que la calidad y fiabilidad de las mismas sea reconocida y, sobre todo, que disponga de una gran cantidad de datos, estructurados o no.

Así, la primera versión del sistema incluye las siguientes fuentes, que serán objeto de desarrollo en los próximos apartados: **i)** Wikipedia, **ii)** PubMed y **iii)**

3.1.1 Wikipedia

Mayo Clinic; aunque se debe destacar que el sistema creado, derivado de esta tesis, permitiría la incorporación de nuevas fuentes en el futuro.

3.1.1 Wikipedia

La primera y principal fuente de información de la que nos hemos servido es Wikipedia, pues se trata de una fuente de información en línea, abierta y colaborativa.

Wikipedia ha sido desarrollada por la Fundación Wikimedia (WMF)⁶⁴ y contiene más de 43 millones de artículos publicados en 298 idiomas⁶⁵. Es uno de los cinco sitios Web más visitados de mundo⁶⁶, siendo su edición en inglés la más grande y activa con más de cinco millones de artículos⁶⁷.

En marzo de 2017 en el campo de la medicina, su edición en inglés contaba con alrededor de 30,000 artículos médicos (Shafee et al., 2017), habiéndose incrementado un año después a 36,850 artículos (Al Tamime, Giordano & Hall, 2018) y considerándose como una de las fuentes médicas más utilizadas por los profesionales de la salud (Azer, 2015; Shafee et al., 2017) así como por otros grupos no especializados en el ámbito médico (Friedlin & McDonald, 2010; Aibar, 2017).

Wikipedia tiene una colaboración profundamente enriquecedora en el ámbito médico (Cohen, 2013; Azzam et al., 2017), siendo ejemplo de ello la iniciativa Cochrane/Wikipedia, cuyo objetivo es aumentar la fiabilidad de los artículos con contenido médico (Matheson & Matheson, 2017). Cochrane reúne a un grupo de investigadores de ciencias de la salud de más de 11.500 voluntarios en más de 90 países que aplican un riguroso y sistemático proceso de revisión de las intervenciones realizadas sobre Wikipedia relativas a la salud⁶⁸.

Wikipedia fue calificada en 2014 como “*the single leading source of medical information for patients and health care professionals*” por el *Institute of Medical Science* (IMS) (Heilman & West, 2015). Esta calificación permite reflexionar sobre el número creciente de profesionales médicos que toma cada vez más conciencia sobre la importancia de colaborar y producir contenidos de calidad en la que ya es la mayor enciclopedia del mundo.

Esta enciclopedia, por su naturaleza pública y colaborativa, fomenta la inclusión y la edición de los contenidos de sus artículos mediante actualizaciones constantes. Por esto, el contenido de Wikipedia se encuentra en consonancia con la

⁶⁴ <https://wikimediafoundation.org>

⁶⁵ https://en.wikipedia.org/wiki/List_of_Wikipedias

⁶⁶ <https://www.alexa.com/topsites>

⁶⁷ https://en.wikipedia.org/wiki/Wikipedia:Size_of_Wikipedia

⁶⁸ https://es.wikipedia.org/wiki/Colaboraci%C3%B3n_Cochrane

3.1.1.1 Anatomía de un artículo de enfermedad

realidad social, característica destacable que ha despertado el interés entre la comunidad científica, debido a la diversidad de conocimientos que la integran (Valle et al., 2018).

Antes de comenzar con la exposición de Wikipedia se debe hacer referencia brevemente a Wikidata⁶⁹, base de datos estructurados, libre, colaborativa, multilingüe y secundaria, que sirve de soporte a todos los proyectos Wikimedia, y creada con el fin de reducir los esfuerzos, coordinar y automatizar los trabajos relacionados dentro de los proyectos de Wikimedia.

En relación con Wikipedia, Wikidata es, sin mas, un repositorio que almacena de forma estructurada los datos mas esenciales de cada artículo de Wikipedia, concretamente la información contenida en cada “*infobox*”, proporcionando un lugar central para almacenarlos.

Aunque se pudiera pensar que hubiera sido mas sencillo utilizar la información contenida en Wikidata y no la contenida en Wikipedia, por su sencillez y estructura, es necesario aclarar que Wikidata no contiene toda la información contenida en los artículos de Wikipedia, por ello se ha descartado esa opción.

Así, aunque Wikidata contiene una lista de los elementos fenotípicos de cada enfermedad, no los incluye todos, y deja fuera los elementos de diagnóstico, esenciales en el desarrollo del presente trabajo de tesis.

A continuación, se describen los elementos principales de los artículos de Wikipedia que son sujeto de extracción y análisis en la presente tesis.

3.1.1.1 Anatomía de un artículo de enfermedad

Como ya se ha mencionado, la versión en inglés de Wikipedia contiene un importante número de artículos médicos, interesando a este trabajo aquellos artículos de contenido médico relativos a enfermedades o categorizados como tales.

Wikipedia permite la creación y edición de sus artículos a cualquier persona sugiriendo una serie de pautas estructurales que deberían ser tenidas en cuenta. Estas indicaciones se encuentran publicadas online⁷⁰, y contempla las secciones sugeridas para los artículos de contenido médico; estas secciones son: “*Signs and symptoms*”, “*Symptoms and causes*”, “*Symptoms and diagnosis*”, “*Signs*”, “*Sign*”, “*Symptoms*”, “*Symptom*”, “*Cause*”, “*Causes*”, “*Diagnosis*”, “*Diagnostic*”, “*Diagnostic approach*”, “*Causes of injury*”, “*Presentation*”, “*Clinical presentation*”, “*Clinical significance*”, “*Symptoms of ...*”, “*Causes of ...*”.

⁶⁹ https://www.wikidata.org/wiki/Wikidata:Main_Page

⁷⁰ https://en.wikipedia.org/wiki/Wikipedia:Manual_of_Style/Medicine-related_articles

3.1.1.1 Anatomía de un artículo de enfermedad

El contenido de dichas secciones podemos esquematizarla de la siguiente forma:

- > “*Signs and symptoms*”; donde se contienen los elementos fenóticos de cada enfermedad –ver sección 2.1.2–.
- > “*Cause*”, donde se contienen los factores que producen los signos y síntomas de una enfermedad. Estas causas o factores pueden ser endógenos (cuando vienen determinadas por la genética del sujeto) o exógenos (cuando dependen de factores externos a él –medioambientales, alimenticios, etc.–).
- > “*Diagnosis*”, donde se contienen los diferentes procesos realizados por los especialistas de la salud para identificar o descartar una enfermedad.
- > “*Clinical significance*”, es donde se contiene la información relativa a los efectos de los tratamientos aplicados a cada enfermedad, valorando si dichos tratamientos son notables en la vida diaria o no.

Es esencial señalar este aspecto, el cual otorga a la Wikipedia, que es una fuente principalmente no estructurada (textual), la posible categorización o estructuración de parte de su contenido en base a las secciones mencionadas. Esto es de vital importancia como se verá más adelante para la extracción de conocimiento de los artículos médicos.

Además, los artículos incorporan unas casillas informativas *infobox* que, generalmente, se ubican tanto al principio como al final del documento, donde se encuentran los códigos de enfermedades que nos remiten a otras fuentes externas y que complementan el contenido de Wikipedia, siendo de gran utilidad para permitir el mapeo de la información publicada en un determinado artículo en Wikipedia con otras fuentes de datos o vocabularios/taxonomías a través de estos códigos (en la Figura 3 se muestra un ejemplo).

Estas pautas permiten que este tipo de documentos (artículos sobre enfermedades) tenga un contenido con un cierto grado de homogeneidad. Sin embargo, debe aclararse que debido a la propia naturaleza de Wikipedia como fuente de información pública accesible a cualquier persona, es habitual encontrar artículos que no cumplen dichas pautas, generando problemas en su extracción y procesamiento.

En aquellos artículos en los que siguen las pautas estructurales de Wikipedia es considerablemente fácil identificar los términos relativos, por ejemplo, a la información fenotípica de una enfermedad, información que sirve de guía para establecer un diagnóstico.

En la Figura 2 se muestra el índice de contenido de un artículo Wikipedia, en el que se aprecian las secciones relevantes como, “*Signs and symptoms*”, “*Causes*” y “*Diagnosis*”.

3.1.1.1 Anatomía de un artículo de enfermedad

a.

Ebstein's anomaly

Pathological specimen and ultrasound image of a heart with Ebstein's anomaly: Abbreviations: RA: right atrium; ARV: atrialized right ventricle; FRV: functional right ventricle; AL: anterior leaflet; SL: septal leaflet; LA: left atrium; LV: left ventricle; asterisk: grade II tethering of the tricuspid septal leaflet

Classification and external resources	
Specialty	cardiology
ICD-10	Q22.5 ↗
ICD-9-CM	746.2 ↗
OMIM	224700 ↗
DiseasesDB	4039 ↗
MedlinePlus	007321 ↗
eMedicine	med/627 ↗
MeSH	D004437 ↗

[\[edit on Wikidata\]](#)

b.

External links

Classification	ICD-10: J10 ↗ , J11 ↗ · ICD-9-CM: 487 ↗ · OMIM: 614680 ↗ · MeSH: D007251 ↗ · DiseasesDB: 6791 ↗
External resources	MedlinePlus: 000080 ↗ · eMedicine: med/1170 ↗ ped/3006 ↗ · Patient UK: Influenza ↗

Figura 3. **a)** “Infobox” ubicado al inicio del documento, capturado a finales de febrero de 2018 del artículo Wikipedia de la “*Ebstein’s anomaly*”; **b)** infobox al final del documento, capturado a mediados de julio de 2019 del artículo Wikipedia de la “*Influenza*”.

Contents [\[hide\]](#)

- 1 Signs and symptoms
- 2 Causes
- 3 Diagnosis
- 4 Management
- 5 Epidemiology
- 6 History
- 7 References
- 8 Further reading
- 9 External Resources
- 10 External link

Figura 2. Índice de contenido del documento, capturado a mediados de julio de 2019 del artículo Wikipedia referente a la “*Gorham's disease*”.

Finalmente, es importante mencionar que a lo largo del desarrollo de esta tesis la organización de la información de los artículos de enfermedades en la versión

3.1.1.1 Anatomía de un artículo de enfermedad

inglesa de Wikipedia ha cambiado. Esta renovación es relativa al contenido de las casillas *infobox* que se encuentra al principio del documento, ya que han dejado de guardar información relativa a códigos de enfermedades para describir de forma reducida algunas características de la condición presentada en el artículo⁷¹. El contenido de este elemento reestructurado abarca: sinónimos, especialidad sanitaria, síntomas, inicio de la enfermedad, duración, causas, factores de riesgos, diagnóstico, diagnóstico diferencial, prevención, tratamiento, medicación, pronóstico, frecuencia y muertes. En la Figura 4 se muestra un *infobox* con contenido reestructurado.

Lung cancer	
Other names	Lung carcinoma

A chest X-ray showing a tumor in the lung (marked by arrow)	
Specialty	Oncology
Symptoms	Coughing (including coughing up blood), weight loss, shortness of breath, chest pains ^[1]
Usual onset	~70 years ^[2]
Types	Small-cell lung carcinoma (SCLC), non-small-cell lung carcinoma (NSCLC) ^[3]
Risk factors	Tobacco smoking · genetic factors · radon gas · asbestos · air pollution ^{[4][5]}
Diagnostic method	Medical imaging, tissue biopsy ^{[6][7]}
Treatment	Surgery, chemotherapy, radiotherapy ^[7]
Prognosis	Five-year survival rate 19.4% (US) ^[2] 41.4% (Japan) ^[8]
Frequency	3.3 million affected as of 2015 ^[9]
Deaths	1.7 million (2015) ^[10]

Figura 4. Contenido reestructurado de la casilla informativa ubicada al principio del documento Wikipedia referente al “*Lung cancer*”. Captura realizada a principios de marzo de 2020.

Debe destacarse que los procesos desarrollados en esta tesis tienen en cuenta ambas posibilidades de cara a la extracción de la información.

⁷¹ https://en.wikipedia.org/wiki/Template:Infobox_medical_condition

3.1.1.2 Enfoque de extracción

El gran reto a afrontar es que la información que se encuentra en Wikipedia, y que forma parte de la base de este trabajo, no se encuentra estructurada (más allá de las secciones ya descritas), suponiendo un incremento de la complejidad a la hora de extraer sus datos y obtener conocimiento.

Existen iniciativas que han desarrollado herramientas capaces de acceder a los datos almacenados en esta fuente, habiendo analizado en este trabajo tres de dichas herramientas (por ser las principales) para la extracción de información:

- a) La API de Wikipedia (AW)⁷², desarrollada en base a la plataforma⁷³ MediaWiki (Barrett, 2008; Kumar, 2010) y soportado por la WMF⁷⁴.

La AW permite hacer consultas de un determinado artículo en un tiempo pasado, de forma que se puede obtener la versión completa (wikitext) de un mismo artículo de diferentes épocas y comparar así su evolución.

Además puede recuperarse, no solo una redacción anterior, sino que también permite consultar diferentes ediciones (en inglés, castellano, alemán, etc.).

A través de esta API también es posible modificar el contenido de artículos, también ofrece búsquedas por un solo artículo o una lista de artículos, usando diversos parámetros de búsqueda como por el título, el *id* principal del documento o el *id* de revisión de este. Es una aplicación nativa, que presenta la desventaja de contar con una documentación muy escasa y dispersa, limitando por lo tanto su utilización.

- b) El proyecto DBpedia⁷⁵ (Auer et al., 2007; Lehmann et al., 2015) surge sobre la base de la formalización del conocimiento de los artículos de Wikipedia. DBpedia es un esfuerzo para extraer datos estructurados de la información creada en varios proyectos de Wikimedia, Wikipedia entre ellos. DBpedia los comparte en línea por medio de tecnologías de Web semántica y Linked Data.

Actualmente incorpora dos tipos de versiones: la versión regular, que se actualiza semestralmente, y la versión “Live”, que tiene la cualidad de obtener información sincronizada en tiempo real. La Figura 5 muestra una consulta SPARQL que permite obtener información de DBpedia.

⁷² https://www.mediawiki.org/wiki/API:Main_page

⁷³ <https://www.mediawiki.org/wiki/MediaWiki>

⁷⁴ <https://en.wikipedia.org/wiki/MediaWiki>

⁷⁵ <http://dbpedia.org>

3.1.1.2 Enfoque de extracción

Otra forma de obtener información contenida en Wikipedia es a través de la técnica del *Web Scraping* que mediante diferentes programas de software permite navegar y extraer información de sitios web.

Esta técnica del *Web Scraping* puede realizarse utilizando programas desarrollados por terceros o desarrollando un programa propio de *Web Scraping*.

```
1 PREFIX dcterms: <http://purl.org/dc/terms/>
2 PREFIX category: <http://dbpedia.org/resource/Category:>
3 PREFIX foaf: <http://xmlns.com/foaf/0.1/>
4 PREFIX owl: <http://www.w3.org/2002/07/owl#>
5 PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
6
7 SELECT DISTINCT ?d ?dn ?wikiPage ?icd9 ?icd10 ?disDB ?meshID ?mlp ?omim ?emed ?frbase
8 WHERE {
9     ?d a <http://dbpedia.org/ontology/Disease>.
10    ?d foaf:isPrimaryTopicOf ?wikiPage.
11    OPTIONAL { ?d <http://dbpedia.org/ontology/icd9> ?icd9. }
12    OPTIONAL { ?d <http://dbpedia.org/ontology/icd10> ?icd10. }
13    OPTIONAL { ?d <http://dbpedia.org/ontology/diseasesDb> ?disDB. }
14    OPTIONAL { ?d <http://dbpedia.org/ontology/meshId> ?meshID. }
15    OPTIONAL { ?d <http://dbpedia.org/property/medlineplus> ?mlp. }
16    OPTIONAL { ?d <http://dbpedia.org/ontology/omim> ?omim. }
17    OPTIONAL { ?d <http://dbpedia.org/property/emedicineTopic> ?emed. }
18    OPTIONAL {
19        ?d owl:sameAs ?frbase.
20        FILTER(regex(str(?frbase), "freebase", "i")).
21    }
22    OPTIONAL {
23        ?d rdfs:label ?dn.
24        FILTER (lang(?dn) = 'en')
25    }
26 }
```

Figura 5. Consulta SPARQL para obtener la lista de enfermedades de Wikipedia mediante SPARQL.

El sistema utilizado en esta tesis para la extracción de información de Wikipedia se basa en dos métodos:

El primero, es a través de DBpedia para obtener el conjunto de artículos contenidos en Wikipedia catalogados como enfermedades humanas, es decir, artículos que son del tipo de categoría “*Disease*”⁷⁶. Este tipo de categoría en Wikipedia tiene la siguiente definición: “Es una condición anormal particular que afecta negativamente a la estructura o función de todo o parte de un organismo y que no

⁷⁶ <https://en.wikipedia.org/wiki/Disease>

3.1.1.2 Enfoque de extracción

se debe a ninguna lesión externa inmediata.”. Además el tipo “Disease” tiene re direcciones de los tipos “*Maladies*” y “*Ailment*”, que son tipos semánticamente iguales.

En la consulta SPARQL de la Figura 5 el tipo “Disease” de Wikipedia esta representada mediante la clase ontológica DBpedia⁷⁷: “disease” <<http://dbpedia.org/ontology/Disease>>).

El segundo, es a través de la creación de un software propio de *Web scrapping* que nos permite acceder a los diversos artículos de Wikipedia y extraer el contenido médico relevante para esta tesis.

Para acceder a la información estructurada contenida en DBpedia se hace uso de consultas SPARQL⁷⁸ que son ejecutadas en el “*Virtuoso environment SPARQL Query Editor of DBpedia*”⁷⁹. La consulta desarrollada obtiene todos los artículos de Wikipedia en inglés referidos a enfermedades humanas así como los códigos de enfermedad asociados de fuentes como: ICD-9, ICD-10, diseasesDb, MeSH, MedlinePlus, OMIM, eMedicine y Freebase, ver Figura 5. El código de la consulta esta disponible en línea en el siguiente enlace⁸⁰.

Debe destacarse que en base a la información que se recopila de DBpedia, donde se obtienen las URLs de los artículos, la aplicación de la técnica de *Web Scraping* es la mejor opción para la extracción de información de Wikipedia porque, como ya se expuso con anterioridad en la sección 3.1.1.1 “*Anatomía de un artículo de enfermedad*”, los propios creadores de Wikipedia han establecido una serie de pautas estructurales para la creación de este tipo de artículos y dichas pautas se siguen con mucha asiduidad, de forma que, a pesar de que el conocimiento no se detiene y la edición de los artículos tampoco, los cambios producidos suelen ser de contenido y no de estructura. Por lo tanto, la extracción de la información de los artículos se centra en extraer la información de aquellas secciones que se consideran relevantes para el objetivo de esta tesis.

El ejemplo claro de esto es que en esta tesis las secciones relevantes, por ser donde se encuentran contenidos los elementos fenotípicos, son: síntomas y signos, causas, diagnósticos, entre otras; al contener la gran mayoría de estos artículos estas secciones bien diferenciadas, se ha aplicado el *Web Scraping* en estas secciones, pues dicha estructura apenas se modificaba, no teniendo que acudir a otras secciones del mismo artículo para extraer información relevante.

Por último se hace referencia a la lista de los elementos recuperados, es decir, aquellos elementos de importancia para la tesis que son: i) los textos en bruto como párrafos, listas y tablas (en la Figura 6 se aprecia un ejemplo de ellos); ii)

⁷⁷ <http://dbpedia.org/ontology/>

⁷⁸ <https://www.w3.org/TR/rdf-sparql-query/>

⁷⁹ <https://dbpedia.org/sparql>

⁸⁰ https://github.com/disnet-project/paperdisnet/blob/master/get_diseases_query.sparql

3.1.1.2 Enfoque de extracción

los enlaces contenidos en cada uno de los textos recuperados; **iii)** los códigos de identificación de enfermedades de diversos vocabularios médicos externos a Wikipedia, encontrados en la casilla informativa al pie del documento (ver Figura 3) y; **iv)** la información de la enfermedad en forma de resumen, encontrada en la casilla informativa al inicio del documento (ver Figura 4).

Symptoms [edit]

Besides [peripheral neuropathy](#) (presenting as [paresthesia](#) or [itching](#), burning or [pain](#)) and discoloration, swelling ([edema](#)) and [desquamation](#) may occur.

Since mercury blocks the degradation pathway of [catecholamines](#), [epinephrine](#) excess causes profuse sweating ([diaphora](#)), [tachycardia](#), [salivation](#) and [elevated blood pressure](#). Mercury is suggested to inactivate [S-adenosyl-methionine](#), which is necessary for catecholamine [catabolism](#) by [catechol-o-methyl transferase](#).

Affected children may show red [cheeks](#) and [nose](#), red ([erythematous](#)) [lips](#), loss of [hair](#), [teeth](#), and [nails](#), transient rashes, [hypotonia](#) and [photophobia](#). Other symptoms may include [kidney](#) dysfunction (e.g. [Fanconi syndrome](#)) or neuropsychiatric symptoms (emotional lability, [memory](#) impairment, [insomnia](#)).

Signs and symptoms [edit]

Thus, the clinical picture is variable.

There is some evidence that signs and symptoms which may suggest lung cancer include:^[1]

- Respiratory symptoms: [coughing](#), [coughing up blood](#), [wheezing](#), or [shortness of breath](#)
- Systemic symptoms: [weight loss](#), [weakness](#), [fever](#), or [clubbing of the fingernails](#)
- Symptoms due to the cancer mass pressing on adjacent structures: [chest pain](#), [bone pain](#), [superior vena cava obstruction](#), or [difficulty swallowing](#)

If the cancer grows in the [airways](#), it may obstruct airflow, causing breathing difficulties. The obstruction can also lead to accumulation of secretions behind the blockage, and increase the risk of [pneumonia](#).^[1]

Depending on the type of tumor, [paraneoplastic phenomena](#) — symptoms not due to the local presence of cancer — may initially attract attention to the disease.^[22] In lung cancer, these phenomena may include [hypercalcemia](#), [syndrome of inappropriate antidiuretic hormone](#) (SIADH, abnormally concentrated urine and diluted blood), ectopic ACTH production, or [Lambert-Eaton myasthenic syndrome](#) (muscle weakness due to [autoantibodies](#)). Tumors in the top of the lung, known as [Pancoast tumors](#), may invade the local part of the [sympathetic nervous system](#), resulting in [Horner's syndrome](#) (drooping of the eyelid and a small pupil on that side), as well as damage to the [brachial plexus](#).^[1]

Many of the symptoms of lung cancer (poor appetite, weight loss, fever, fatigue) are not specific.^[6] In many people, the cancer has already spread beyond the original site by the time they have symptoms and seek medical attention.^[23] Symptoms that suggest the presence of metastatic disease include weight loss, bone pain, and neurological symptoms (headaches, [fainting](#), [convulsions](#), or limb weakness).^[1] Common sites of spread include the brain, bone, [adrenal glands](#), opposite lung, [liver](#), [pericardium](#), and [kidneys](#).^[23] About 10% of people with lung cancer do not have symptoms at diagnosis; these cancers are [incidentally found](#) on routine [chest radiography](#).^[17]

Causes [edit]

Cancer develops after [genetic damage to DNA](#) and [epigenetic](#) changes. Those changes affect the [cell's normal functions](#), including [cell proliferation](#), programmed cell death ([apoptosis](#)), and [DNA repair](#). As more damage accumulates, the risk for cancer increases.^[24]

Smoking [edit]

Tobacco smoking is by far the main contributor to lung cancer.^[4] Cigarette smoke contains at least 73 known [carcinogens](#),^[25] including [benzo\[a\]pyrene](#),^[26] [NNK](#), [1,3-butadiene](#), and a radioactive isotope of polonium — [polonium-210](#).^[25] Across the developed world, 90% of lung cancer deaths in men and 70% of those in women during the year 2000 were attributed to smoking.^[27] Smoking accounts for about 85% of lung cancer cases.^[7] A 2014 review found that [vaping](#) may be a risk factor for lung cancer but less than that of cigarettes.^[28]

Passive smoking — the inhalation of smoke from another's smoking — is a cause of lung cancer in nonsmokers. A passive smoker can be defined as someone either living or working with a smoker. Studies from the US,^[29]^[30]^[31] Europe,^[32] and the UK^[33] have consistently shown a significantly-increased risk among those exposed to passive smoking.^[34] Those who live with someone who smokes have a 20–30% increase in risk while those who work in an environment with secondhand smoke have a 16–19% increase in risk.^[35] Investigations of [sidestream smoke](#) suggest that it is more dangerous than direct smoke.^[36] Passive smoking results in roughly 3,400 lung cancer-related deaths each year in the US.^[31]

Marijuana smoke contains many of the same [carcinogens](#) as those found in tobacco smoke,^[37] however, the effect of smoking [cannabis](#) on lung cancer risk is not clear.^[38]^[39] A 2013 review did not find an increased risk from light to moderate use.^[40] A 2014 review found that smoking cannabis doubled the risk of lung cancer, though cannabis is in many countries commonly mixed with tobacco.^[41]

Relationship between cigarette consumption per person (blue) and male lung cancer rates (dark yellow) in the US over the century.

Cumulative risk of death from lung cancer.

Year	Men (Cumulative risk)	Women (Cumulative risk)
1900	~10	~5
1950	~15	~10
2000	~20	~15
2050	~25	~20

Figura 6. Textos médicos en secciones relevantes. Capturada la imagen de fondo a mediados de julio de 2019 del artículo Wikipedia de la “*Acrodermatitis gangrenosa*” y la imagen de frente capturada a principios de marzo de 2020 del artículo Wikipedia del “*Lung cancer*”.

De todos estos elementos recuperados debemos destacar la importancia de los códigos de enfermedad de diferentes vocabularios, que nos permite conectar la información que Wikipedia contiene de una enfermedad con la información que otra fuente de información posee con respecto a la misma enfermedad.

Finalmente, para una mejor comprensión en la Figura 7 se muestra el flujo de trabajo del proceso de extracción de información médica de Wikipedia .

3.1.2 PubMed

Figura 7. Flujo de trabajo del proceso de extracción de conocimiento de Wikipedia.

3.1.2 PubMed

PubMed⁸¹ es una base de datos de acceso libre y especializada en las ciencias de la salud conformada únicamente por artículos científicos médicos. Actualmente PubMed contiene más de 30 millones de artículos extraídos de MEDLINE (otra base de datos médica textual) (Miller, Lacroix & Backus, 2000), revistas de ciencias de la vida y libros en línea, incluyendo trabajos publicados de más de 5.300 revistas médicas, consolidándose como un referente para los profesionales de la salud en lo que las fuentes de información médica se refieren.

No obstante, no todos los artículos de PubMed son accesibles gratuitamente de forma completa, sino que en muchas ocasiones el contenido gratuito al que se puede acceder es solo un extracto del mismo (*abstract*).

Dentro de PubMed se encuentra PubMed Central⁸², que es el repositorio en el que se contienen los artículos completos de los que se nutre PubMed, de forma que muchos de los artículos de PubMed contienen enlaces al contenido completo del artículo a través de PubMed central.

⁸¹ <https://www.ncbi.nlm.nih.gov/pubmed>

⁸² <https://www.ncbi.nlm.nih.gov/pmc/>

3.1.2.1 Enfoque de extracción

Una característica de los artículos de PubMed es que estos se encuentran anotados con descriptores del vocabulario MeSH, los cuales son de gran utilidad para realizar búsquedas (Lowe & Barnett, 1994; Lipscomb, 2000; Trueba-Gómez & Estrada-Lorenzo, 2010, p.; Valle et al., 2018) de artículos que pertenecen a una o varias de las categorías que describe MeSH.

En el caso de PubMed, y a diferencia de lo que ocurre con Wikipedia, solo se han podido tener en consideración al objeto de esta tesis, los resúmenes de los artículos por dos motivos: en primer lugar, porque como ya se comentó previamente no siempre es posible acceder al texto completo del artículo; y en segundo lugar, porque los artículos de PubMed no tienen una estructura estándar, por lo que considerar el artículo en su totalidad supondría extraer y almacenar información que podría generar gran cantidad de ruido dentro del contexto de la tesis.

No obstante, no son ajenas a este estudio doctoral las limitaciones que conlleva la extracción de conocimiento únicamente de los resúmenes de los artículos (Westergaard et al., 2018), debiendo considerarse, para estudios y trabajos futuros, la extracción del contenido del artículo completo cuando esto sea posible, y desechando del mismo la información irrelevante.

Es importante enfatizar que en PubMed la información sobre una sola enfermedad se distribuye entre múltiples documentos –a diferencia de Wikipedia, en la que existe una relación biyectiva entre los artículos y las enfermedades–; Esto supone una importante consideración en el diseño del conjunto de datos multi-fuente propuesto en esta tesis y que es descrito en la sección 3.2.2 Conjunto de datos.

3.1.2.1 Enfoque de extracción

Del estudio de los trabajos relacionados se ha identificado el método más común para la extracción de información de PubMed, y por lo tanto ha sido este método el que se ha aplicado para desarrollo la presente tesis. El método es una combinación del vocabulario MeSH y la base de datos de literatura PubMed (Li & Agarwal, 2009; Zhou et al., 2014b; Hoehndorf, Schofield & Gkoutos, 2015; Cousyn et al., 2018).

En primer lugar se debe decir que el método utilizado se basa en el uso vocabulario MeSH. Este vocabulario está compuesto por muchas categorías, tales como: Anatomía (*A*), Organismos (*B*), Químicos y Drogas (*D*), etc. En el contexto de la presente tesis, es de interés la categoría (*C*), que hace referencia a las enfermedades, y más concretamente las subcategorías *C01* a *C20*, que hacen referencia a enfermedades humanas, siendo las restantes a enfermedades animales, signos y

3.1.2.1 Enfoque de extracción

síntomas o heridas y lesiones, entre otras, tal y como se muestra en el árbol de clasificación que se encuentra disponible en línea en el siguiente enlace⁸³.

Dentro del proceso de extracción, el primer paso es obtener aquellos términos categorizados como *Diseases* en la categoría *C* del vocabulario MeSH para que, una vez extraídos dichos términos se pueda mapear cada enfermedad con sus equivalentes en el vocabulario *Disease Ontology* (DO) (Schriml et al., 2019) para obtener los códigos de enfermedades del vocabulario ICD-10 (World Health Organization, 1993), OMIM (Hamosh et al., 2005), MeSH (Lipscomb, 2000), SNOMED CT (Donnelly, 2006) y del *Unified Medical Language System* (UMLS) (Bodenreider, 2004).

La *Disease Ontology* como proveedor de identificadores de enfermedad es relevante porque es una ontología referenciada que integra y conecta cerca de 46.000 términos médicos a través de un amplio mapeo de referencias cruzadas en diferentes vocabularios: MeSH, ICD, NCI's thesaurus, SNOMED CT y OMIM (Schriml et al., 2019).

Es importante enfatizar que el uso de múltiples vocabularios tiene por objeto obtener la mayor cantidad de códigos entre vocabularios (para mapear las enfermedades a través diferentes fuentes de información).

Como segundo paso, es necesario obtener el conjunto de artículos PubMed cuyos términos estén asociados a cada uno de los elementos de la lista de enfermedades MeSH previamente descripta. A través de la API de PubMed Entrez (Sayers, 2010, 2018) es posible llevar a cabo esta tarea, dado que con ella se tiene acceso a todas las bases de datos de Entrez⁸⁴ incluyendo PubMed, PubMed Central (PMC), Gene (Maglott et al., 2005), Nucleotide⁸⁵ y Protein⁸⁶.

Una característica importante a mencionar de la API de Entrez y que también ha sido utilizada en este trabajo, es la posibilidad de clasificar los artículos por su relevancia (Fiorini et al., 2018), logrando así centrar los esfuerzos en aquellos artículos más significativos, sobresalientes o destacados asociados a un término MeSH. Por lo tanto, se ha limitado la extracción a los 100 artículos más relevantes asociados a cada término, con el objetivo de reducir el tiempo de extracción, pero garantizando obtener una lista de artículos lo suficientemente relevantes.

⁸³ <https://meshb.nlm.nih.gov/treeView>

⁸⁴ <https://www.ncbi.nlm.nih.gov/Class/MLACourse/Original8Hour/Entrez/>

⁸⁵ <https://www.ncbi.nlm.nih.gov/nucleotide/>

⁸⁶ <https://www.ncbi.nlm.nih.gov/protein/>

3.1.3 MayoClinic

Por último se debe hacer referencia a la lista de los elementos recuperados sobre los artículos, es decir, aquellos elementos de importancia para la tesis que son, en este caso: **i)** el título, **ii)** el resumen, **iii)** los nombres de los autores, **iv)** el identificador único en PubMed y PubMed Central, **v)** el DOI (*Digital Object Identifier*, por sus siglas en inglés), **vi)** los términos MeSH asociados al artículo y **vii)** las palabras clave (*keywords*).

Figura 8. Flujo de trabajo del enfoque propuesto de extracción de información de PubMed.

El flujo de trabajo para la extracción de información de los documentos de PubMed se encuentra representado en la Figura 8.

3.1.3 MayoClinic

Según el sitio web oficial Mayo Clinic⁸⁷ es una organización sin ánimo de lucro comprometida con la práctica clínica, la educación y la investigación, que proporciona atención experta y personalizada a todos los que la necesitan. Según la lista *U.S. News & World Report Best Hospitals Honor Roll* en este último año (2019 – 2020) Mayo Clinic ha sido considerada, una vez más, como el mejor hospital de los Estados Unidos de América (EUA) y el primero en doce especialidades, siendo

⁸⁷ <https://www.mayoclinic.org>

3.1.3.1 Enfoque de extracción

la única organización de atención médica que figura constantemente entre los principales proveedores de todo EUA^{88,89}.

Más allá de ser un proveedor de servicios de salud también dedica esfuerzos a la investigación y publicación de conocimiento científico a través de su equipo de médicos e investigadores.

Mayo Clinic tiene una amplia gama de productos y servicios entre las cuales se encuentra en su página Web oficial la sección referente a sus trabajos publicados en diferentes revistas científicas⁹⁰ y una más referente a la revista propia del hospital llamada *Mayo Clinic Proceedings* (MCP)⁹¹, siendo ésta una de las publicaciones científicas más leídas y citadas por los médicos, con una difusión mensual de, aproximadamente, 127.000 ejemplares, conteniendo gran cantidad de información médica relevante sobre las enfermedades y sus manifestaciones fenotípicas.

Adicionalmente, entre sus fuentes de datos disponibles, Mayo Clinic proporciona en su página web oficial un catálogo de descripción de enfermedades o afecciones, la cual a principios de 2019 contaba con 1.180 artículos sobre enfermedades, que están disponibles para su consulta en el siguiente enlace⁹².

3.1.3.1 Enfoque de extracción

Existen un enfoque principal del que ya se ha hecho mención con anterioridad para extraer información relativa a los elementos fenotípicos y de diagnóstico de las enfermedades de Mayo Clinic. El enfoque es a través de los artículos de enfermedades que se encuentran disponibles en su página Web oficial⁹², que se muestra en la Figura 9.

Esta lista de artículos web o catálogo de enfermedades, es considerada una fuente viable para recuperar información ya que los artículos online presentan una estructura genérica que permite identificar de forma sencilla el contenido fenotípico relativo al diagnóstico de las enfermedades.

Los artículos contenidos en Mayo Clinic están configurados en diversas secciones. Sin embargo, tras un análisis pormenorizado de la información contenida en su web, aquellas en las que se ha detectado una mayor concentración de contenido fenotípico son las secciones “*Symptoms*”, “*Causes*” y “*Diagnosis*”. Estas secciones contienen información relevante para esta tesis como: los síntomas que presenta una enfermedad, las causas, los diagnósticos, los tratamientos, los tipos de médicos

⁸⁸ <https://www.mayoclinic.org/es-es/about-mayo-clinic/quality/rankings>

⁸⁹ <https://health.usnews.com/best-hospitals>

⁹⁰ <https://www.mayo.edu/research/publications>

⁹¹ <https://www.mayoclinic.org/medical-professionals/publications>

⁹² <https://www.mayoclinic.org/diseases-conditions>

3.1.3.1 Enfoque de extracción

que la tratan, sus departamentos o centros e información relativa a los servicios relacionados que ofrece Mayo Clinic.

Figura 9. Lista principal de enfermedades en Mayo Clinic. Ubicada en su página web.

Al igual que con Wikipedia, se ha desarrollado un *pipeline/workflow* usando *Web scraping* sobre cada una de éstas secciones relevantes de cada uno de los artículos relativos a enfermedades que contiene.

Por último, al igual que hemos hecho con el resto de fuentes de información, se hará referencia a la lista de los elementos recuperados. Es importante mencionar que los artículos de enfermedades de Mayo Clinic no cuentan con códigos de enfermedades de bases de datos médicas externas, por lo que resulta muy complejo mapear enfermedades y complementar su conocimiento usando otras fuentes. También se debe mencionar que la lista de enfermedades que contiene es considerablemente menos extensa que la de Wikipedia. Sin embargo, presenta la ventaja de ser un sitio Web oficial y completamente revisado por especialistas de la salud, siendo por lo tanto la información obtenida de una mayor fiabilidad.

Por todo ello, de Mayo Clinic solo se pueden recuperan textos médicos de las secciones relevantes: “*Symptoms*”, “*Causes*” y “*Diagnosis*”.

Con respecto al proceso de extracción de información, este conlleva tres pasos que se describen a continuación:

- a) Se obtiene la lista de enfermedades contenidas en Mayo Clinic. Dicha lista de enfermedades⁹² se encuentra ordenada alfabéticamente de la A a la Z, incluyendo una categoría mas, la #, que contiene información sobre el “*22q11.2 deletion syndrome*”, una enfermedad a la que la propia sección redirige a la lera D, “*DiGeorge syndrome*”.

3.1.3.1 Enfoque de extracción

Cada una de estas letras (de la *A* a la *Z*), contiene la lista de aquellas enfermedades cuya inicial es la letra en la que aparecen, de forma que la lista *A* contiene todas aquellas enfermedades humanas cuyos nombres inician con esta letra. Ver Figura 9.

- b) Se han extraído todos los enlaces que contienen cada una de las enfermedades incluidas en cada una de las categorías. Ver Figura 10.
- c) Finalmente se extrae el contenido diagnóstico de cada una de ellas. Ver Figura 11.

En la Figura 12 se muestra el flujo de trabajo del proceso de extracción de información textual de Mayo Clinic.

Figura 10. Lista de la categoría *A* relativa a todas las enfermedades cuyo nombre inicia con la primera letra del abecedario inglés.

Figura 11. Contenido del artículo Mayo Clinic relativo a la enfermedad "*Acoustic neuroma*". Secciones relevantes: "*Symptoms & causes*" y "*Diagnosis & treatment*".

Figura 12. Flujo de trabajo del proceso de extracción de conocimiento de Mayo Clinic.

3.2 Enfoque principal de extracción y generación de conocimiento

Esta sección describe la arquitectura general del sistema desarrollado, incluyendo la extracción de datos y la subsiguiente extracción de conocimiento, profundizando en cada uno de los componentes que lo integran. En aras a la claridad, dicha arquitectura se describe en la Figura 13.

3.2.1 Flujo de trabajo

El primer paso del proceso de extracción y generación de conocimiento es la recuperación de la información de las fuentes de información ya identificadas y descritas.

3.2.1 Flujo de trabajo

Figura 13. Arquitectura y flujo de trabajo del enfoque de extracción y generación de conocimiento diagnóstico.

Mediante técnicas de *Web scraping* se ejecutará el proceso GDLP (por sus siglas en inglés *Get Disease List Process*), que es el encargado de obtener la lista de enfermedades de cada una de las fuentes de información públicas configuradas (y descritas en apartados previos), proporcionando los enlaces de todos aquellos documentos que se encuentren relacionados con dichas enfermedades en cada fuente.

El procedimiento GDLP se adapta a cada una de las fuentes de información. Para la extracción del contenido de Wikipedia utiliza la consulta SPARQL anteriormente descrita⁸⁰. En el caso de PubMed la extracción de información la realiza a través de una lista de términos de MeSH. Para Mayo Clinic los enlaces para los artículos se recuperan directamente del contenido de la página oficial usando *Web scraping*.

Una vez recogida la lista de URLs, el módulo *Web scraping* (WSP) se encarga de conectar con cada uno de los hipervínculos y de extraer el texto específico que describe las manifestaciones y diagnósticos fenotípicos de cada enfermedad, así como los enlaces (referencias) contenidos en los mismos textos. Además, y siempre

3.2.1 Flujo de trabajo

que sea posible, se extrae la información relacionada con la codificación de las enfermedades, es decir, los códigos utilizados para identificar la enfermedad en diferentes bases de datos o vocabularios. Actualmente se puede recuperar información de más de 7.261 artículos en Wikipedia, de 229.160 resúmenes de artículos en PubMed que corresponden a 2.355 enfermedades y de 1.138 artículos de enfermedades en Mayo Clinic. La información extraída por el proceso WSP se almacena en un conjunto de datos llamado “Raw DS”, que contiene el texto crudo, sin procesar.

El segundo paso dentro del flujo de trabajo es la primera parte del proceso de NLP (por las siglas en inglés: *Natural Language Processing*). Este componente es responsable de: **i)** leer todos los textos de una “*snapshot*”, y **ii)** obtener para cada texto una lista de conceptos/términos clínicos relevantes, descartando aquellos no relacionados/relevantes.

El procesamiento de lenguaje natural (NLP, por sus siglas en inglés) es la disciplina que se enfoca en la comprensión mediante sistemas informáticos del lenguaje humano y así interpretarlo y manipularlo. Dentro del contexto de la medicina este tipo de sistemas permiten procesar lenguaje médico y obtener conceptos referentes a diversos aspectos de esta área del conocimiento: anatomía, químicos y fármacos, comportamientos, genes, moléculas, procedimientos, síntomas, enfermedades, entre otros.

El uso de una herramienta de NLP es indispensable ya que toda la información que se trata son textos de contenido médico en inglés.

Actualmente el proceso NLP utiliza una herramienta denominada MetaMap (Aronson, 2001a) como herramienta de procesamiento de lenguaje médico natural para extraer términos clínicos de interés –véase la sección 3.2.4.1 sobre la configuración de MetaMap–. El proceso de extracción de términos del texto se basa en la definición de los tipos semánticos que MetaMap usa y que se van a recuperar. En este contexto los tipos semánticos (SM, por sus siglas en inglés “*Semantic types*”) son elementos definidos en UMLS para definir categorías de conceptos. MetaMap utiliza los SM para encontrar elementos pertenecientes a una categoría semántica, es decir, en el conjunto de datos propuesto todos los términos médicos tienen asociados uno o más SM de los definidos como entrada a MetaMap. La lista posible de *semantic types* que se pueden pasar a MetaMap para la identificación de términos está disponible en línea en el siguiente enlace⁹³.

El resultado del proceso de NLP se almacena en el “DISNET Medical DS” (DMDS). En él se almacenan, de manera estructurada, los conceptos médicos que han sido obtenidos por la primera parte del proceso de NLP, así como toda la

⁹³ <https://metamap.nlm.nih.gov/SemanticTypesAndGroups.shtml>

3.2.1 Flujo de trabajo

información necesaria para rastrear el origen de dichos conceptos –con el fin de investigar cualquier error que pueda ser detectado posteriormente.

En resumen, la información almacenada de Wikipedia, PubMed y Mayo Clinic de forma estructurada en el DMDS es: **i)** los conceptos médicos con su ubicación, información y tipos semánticos, **ii)** los textos de los que se han extraído y los enlaces que contienen, **iii)** las secciones a las que pertenecen los textos, **iv)** el documento o documentos que describen la enfermedad (enlace web) y **v)** los códigos de identificación de la enfermedad en diferentes vocabularios o bases de datos. La información adicional, como el día de la extracción y la fuente, también son almacenados.

Antes de llegar al último paso del proceso, es importante destacar la naturaleza de la información hasta ahora almacenada. Específicamente, el sistema desarrollado no solo permite extraer signos o síntomas de una enfermedad: también permite extraer términos médicos que en esta tesis son considerados como manifestaciones fenotípicas y elementos de diagnóstico de la enfermedad.

El proceso de extracción de estos términos es realizado actualmente por la herramienta MetaMap. Aunque este proceso garantiza un alto porcentaje de acierto en la extracción, trabajos previos han demostrado que pueden existir errores relativos por ejemplo a modificadores de características fenotípicas: ejemplo: se detecta “*lymph nodes*” como elemento fenotípico, pero también el adjetivo “red” (para formar “*red lymph nodes*” – enrojecimiento de nodos linfáticos) de forma aislada. El proceso MetaMap identifica ambos conceptos por separado (“*lymph nodes*” y “*red*”) como elementos fenotípicos, siendo en todo caso la unión de ambos el elemento buscado. La identificación de este tipo de términos (*red*, de manera aislada) introduce posibles falsos positivos. Por esa razón, el proceso de extracción basado en MetaMap necesita de un proceso adicional que permita validar/verificar que los términos extraídos son correctos. Para este objetivo, en la presente tesis se hace uso de la segunda parte del proceso de NLP, el método/proceso TVP desarrollado previamente en (Rodríguez González et al., 2015, 2018) y que se basa en la comparación de los elementos extraídos con otros vocabularios mediante diferentes estrategias de identificación (comparación de *Concept Unique Identifiers* (CUI) (McInnes, Pedersen & Carlis, 2007), equiparación léxica (exacta o mediante distancia Levenshtein (Levenshtein, 1966)), etc).

Finalmente, el resultado del proceso de TVP también es almacenado en el conjunto de datos, por lo que se sabe para cada concepto médico si ha sido evaluado tanto positivamente como negativamente.

3.2.1 Compartición de la información

Tras la extracción y procesamiento de las enfermedades y sus características, el sistema desarrollado como parte de esta tesis permite proporcionar una forma con la que acceder y recuperar la información. En este contexto, se ha desarrollado una interfaz basada en la tecnología REST⁹⁴ que proporciona a los usuarios registrados en la plataforma DISNET⁹⁵ un TOKEN de acceso que autoriza a consumir la API y consultar todos los datos disponibles en el sistema. Esta interfaz se describe en detalle en el sitio web del sistema (<http://disnet.ctb.upm.es/apis/disnet>). Adicionalmente, en la sección 3.2.5 Proyecto DISNET se describe el proyecto desarrollado como parte de esta tesis doctoral en su totalidad.

Finalmente, es imperativo volver a mencionar que el proceso principal de extracción y generación de conocimiento se ejecuta automáticamente dos veces por mes sobre la fuente de Wikipedia, habiéndose iniciado su actividad el 1 de febrero de 2018. En el Capítulo 3 se describe detalladamente el contenido completo del conjunto de datos.

3.2.2 Conjunto de datos

En esta sección se describen todos los elementos que conforman el conjunto de datos y como interactúan entre sí. No obstante, antes de dicha descripción se describen las particularidades que incorpora el conjunto de datos creado:

- i. Capacidad para incorporar datos provenientes de múltiples fuentes de información: el sistema permite generar nuevas fuentes de datos donde se almacene información como el nombre de la fuente de datos, nombre de enfermedades, enlaces a los documentos de las enfermedades, identificadores de enfermedades, textos relativos al diagnóstico, términos médicos referentes al diagnóstico de enfermedades, etc.
- ii. El modelo de datos creado permite el almacenamiento de dos tipos de estructuras de datos: **1.** Estructuras uno a uno, que son aquellas estructuras en las que, para abarcar la totalidad del conocimiento que se tiene respecto de una enfermedad, es suficiente acudir a un único artículo (es el caso de Wikipedia o Mayo Clinic, que cada artículo posee todo el conocimiento que Wikipedia y Mayo Clinic tiene sobre dicha enfermedad, es decir, no hay dos artículos de Wikipedia o Mayo Clinic que hablen sobre una misma enfermedad) y, **2.** Estructuras uno a muchos, que son aquellas estructuras en las que para abarcar la totalidad de conocimiento que se tiene sobre una enfermedad es necesario acudir a

⁹⁴ https://es.wikipedia.org/wiki/Transferencia_de_Estado_Representacional

⁹⁵ <http://disnet.ctb.upm.es>

3.2.2 Conjunto de datos

dos o más artículos (como ocurre con PubMed, que al estar conformado por un gran número de artículos científicos de una misma enfermedad, existen varios artículos que hablan de dicha enfermedad y cada uno de ellos puede añadir información diferente, de forma que es necesario almacenar y analizar todos los artículos relativos a la misma enfermedad para obtener todo el conocimiento posible).

- iii. Capacidad para almacenar la información de las múltiples fuentes capturada en diferentes lapsos de tiempo, con el fin de complementar con cada captura la información anteriormente almacenada y observar cómo se encontraba el conocimiento médico en una determinada fecha y por lo tanto analizar su evolución temporal.
- iv. Capacidad para registrar todos los elementos que han conformado ese artículo (el documento en sí, la fuente de donde proviene el documento, las diferentes secciones y sus textos, los conceptos médicos hallados en esos textos, etc.) así como capacidad para guardar la configuración de las distintas herramientas de NLP que se han podido utilizar para obtener los conceptos médicos de los textos.

En la Figura 14 se muestra el diagrama entidad – relación de la base de datos creada a tal efecto y en el siguiente enlace⁹⁶ se visualiza el mismo diagrama de forma más detallada.

Como se aprecia, dicha figura muestra el conjunto de datos, que gira en torno a la entidad documento, siendo ésta entidad un “documento compuesto” porque permite integrar las dos tipologías de estructuras a las que antes se hizo referencia: estructuras uno a uno (Wikipedia y Mayo Clinic); y estructuras uno a muchos (PubMed) –ver (i).

En la gráfica se muestra la forma en la que este documento se relaciona con el resto de entidades: **entidad *disease*** (que es la que hace referencia a la enfermedad a la que se refiere ese documento); la **entidad *synonym*** (que nos muestra los sinónimos de los nombres de las enfermedades); la **entidad *source*** (que hace referencia a la fuente de la que proviene ese documento); la **entidad *paper*** (que nos muestra todos los artículos relacionados con la enfermedad a la que se refiere dicho documento –ésta entidad es de exclusiva aplicación a PubMed y fuentes similares–); la **entidad *term*** (que nos dice cuáles son aquellos conceptos MeSH asociados a los artículos de PubMed); la **entidad *code*** (que muestra los códigos de enfermedad encontrados en dicho documento); la **entidad *resource*** (que nos indica las bases de datos externas a la que pertenecen los códigos contenidos en el documento); la **entidad *url*** (que nos muestra el enlace al documento así como el enlace a los códigos contenidos en los documentos); la **entidad *section*** (que

⁹⁶ https://github.com/disnet-project/main_configuration_directory/blob/master/disnet_e-r_model_v2020.png

3.2.2 Conjunto de datos

nos dice cuáles son las diferentes secciones en la que se divide ese documento); la **entidad *text*** (que nos muestra los textos de las secciones relevantes de ese documento); la **entidad *symptom***, (íntimamente relacionada con la entidad *text* y que nos muestra los conceptos médicos obtenidos de los textos); la **entidad *semantic type*** (que nos proporciona los tipos semánticos de cada concepto médico); la **entidad *retrieval method*** (que es la que nos muestra cual es el método por el cual se han recuperado los códigos y sinónimos de las enfermedades) y, por último, la **entidad *configuration*** (que nos muestra las configuraciones de las distintas herramientas que se han usado para crear una captura de conocimiento –*snapshot*–).

Además entre estas entidades existen también las entidades de conexión que son las que relacionan unas entidades con otras: *has_disease*, *has_symptom*, *has_text*, *has_semantic_type*, etc. En línea⁹⁷ se encuentra disponible los datos de la enfermedad “*Otic polyp*” en cada una de las entidades del conjunto de datos descrito a modo de ejemplo.

⁹⁷ https://github.com/disnet-project/thesis/blob/master/oticpolyp_data_dmds.xlsx

3.2.2 Conjunto de datos

Figura 14. Modelo entidad – relación del conjunto de datos.

3.2.3 Componentes técnicos

En esta sección se describen todos los componentes técnicos del sistema que llevan a cabo la actividad del proceso principal de extracción y generación de conocimiento –ver sección 3.2.

En la Figura 15 se presenta un diagrama que permite visualizar como interactúan los diferentes componentes.

En primer lugar, se hace referencia a las tecnologías usadas para desarrollar este sistema:

- i. El sistema se ha ejecutado en un servidor con un sistema basado en Ubuntu⁹⁸. El sistema operativo es el Pop!_OS en su versión 18.04 LTS que ha sido desarrollado por system76⁹⁹.
- ii. El lenguaje de programación del *back-end* es Java. Se ha hecho de Spring boot¹⁰⁰ que permite implementar una arquitectura de micro servicios¹⁰¹. La arquitectura de micro servicios está compuesta por pequeñas aplicaciones independientes que se comunican a través de APIs. Esta arquitectura también hace que los servicios sean más fáciles de escalar y más rápidos de desarrollar. Además, la lógica del negocio se encuentra repartida entre cada uno de los servicios que la componen.
- iii. Para completar la arquitectura de micro servicios se usa la plataforma de software Docker¹⁰² que permite crear, probar e implementar aplicaciones rápidamente. Docker empaqueta el software en unidades estandarizadas llamadas contenedores¹⁰³ que incluyen todo lo necesario para que el software (en este caso los servicios Web desarrollados con Java) se ejecuten, incluyendo las bibliotecas, herramientas de sistema y códigos necesarios. Docker proporciona un estándar para la ejecución del código. Docker se instala en el servidor y a través de comandos sencillos se puede crear, iniciar o detener contenedores. Los contenedores son mucho más ligeros que las máquinas virtuales. De hecho no se indican qué recursos se van a necesitar, sino que es el mismo Docker, en función de las necesidades de cada momento, el encargado de asignar los recursos que sean necesarios para que los contenedores funcionen adecuadamente.

⁹⁸ <https://ubuntu.com/>

⁹⁹ <https://system76.com/pop>

¹⁰⁰ <https://spring.io/projects/spring-boot>

¹⁰¹ <https://aws.amazon.com/es/microservices/>

¹⁰² <https://aws.amazon.com/es/docker/>

¹⁰³ <https://www.docker.com/resources/what-container>

3.2.3 Componentes técnicos

Figura 15. Estructura del sistema Web encargada de ejecutar el proceso de extracción y generación de conocimiento médico.

- iv. El enfoque de multi contenedor requiere el uso de una herramienta que permita orquestar el *cluster* de aplicaciones. Docker provee esta herramienta a través de un complemento llamado Docker Compose¹⁰⁴, que define y ejecuta aplicaciones de varios contenedores Docker. Con Docker Compose se utiliza un archivo YAML¹⁰⁵ para configurar y administrar los servicios de la aplicación en forma de contenedores Docker. Luego,

¹⁰⁴ <https://docs.docker.com/compose/>

¹⁰⁵ <https://es.wikipedia.org/wiki/YAML>

3.2.3 Componentes técnicos

con un solo comando, se puede crear e iniciar todos los servicios de la configuración.

- v. Las bases de datos han sido creadas por medio del gestor de bases de datos MySQL¹⁰⁶. La edición utilizada es la *MySQL Community Edition*.
- vi. Finalmente, para desarrollar la aplicación Web se ha hecho uso del marco de trabajo responsivo Materialize¹⁰⁷ para el *front-end* (la interfaz gráfica que permite que el usuario pueda visualizar de una forma clara el contenido digital) y se ha hecho uso del marco de trabajo Thymeleaf¹⁰⁸ para el back-end.

Los motivos de dicha elección son, con respecto a Materialize, la facilidad para crear de forma sencilla una interfaz web que sea amigable al usuario y con respecto a Thymeleaf, es un marco de trabajo que permite implementar el patrón MVC¹⁰⁹ de aplicaciones Web.

Las versiones utilizadas de las tecnologías ya descritas son las siguientes: **i)** Java 8¹¹⁰, **ii)** Spring boot 1.5.9.RELEASE¹¹¹, **iii)** Docker 19.03.6, **iv)** Docker Compose 1.25.1, **v)** HTML5, **vi)** Thymeleaf 2.1.5.RELEASE, **vii)** Materialize v0.100.2 y **viii)** MySQL 5.7.

Una vez se han descrito las tecnologías de desarrollo, a continuación, se describirán cada uno de los servicios que se muestran en la Figura 15.

Como se puede observar en la imagen, existen 14 contenedores, cada uno de ellos con una función muy concreta:

- Los contenedores que se muestran en color blanco son los encargados de desplegar las tres bases de datos necesarias en contenedores diferentes.
 - > El primero de ellos (DB Registro de Lista de Enfermedades), gestiona los datos recuperados de DBpedia.
 - > El segundo de ellos (Conjunto de datos médicos), almacena toda la información referente al conjunto de datos médicos, es decir, las fuentes de donde proviene dicha información, las enfermedades halladas en cada una de esas fuentes, los documentos que

¹⁰⁶ <https://www.mysql.com/>

¹⁰⁷ <https://materializecss.com/>

¹⁰⁸ <https://www.thymeleaf.org/>

¹⁰⁹ <https://si.ua.es/es/documentacion/asp-net-mvc-3/1-dia/modelo-vista-controlador-mvc.html>

¹¹⁰ <https://www.java.com/es/download/faq/java8.xml>

¹¹¹ <https://spring.io/projects/spring-boot#overview>

3.2.3 Componentes técnicos

describen dichas enfermedades, los textos, los códigos y los conceptos médicos obtenidos mediante el proceso de NLP y demás información relevante.

- > El tercero de ellos (Base de datos de usuario), almacena toda la información referente a los usuarios registrados, el TOKEN de acceso otorgado y las consultas realizadas con el mismo que se almacenan a modo de histórico.
- Los contenedores que se muestran en color verde son los servicios Web con los que el usuario interactúa directamente.

- > El primero de ellos DWA (*DISNET Web Application*), es el relativo a la página Web a través de la cual los usuarios, mediante su registro, obtienen un TOKEN de acceso autorizado.

Con este TOKEN los usuarios tienen la posibilidad de visualizar todas las consultas realizadas por ellos mismos.

- > El segundo de ellos MAPI (*Main API*), es el servicio principal que se pone a disposición de los usuarios (la API).

Es aquí donde se encuentran todas las consultas a las que el usuario registrado tiene acceso mediante la conexión existente entre MAPI y el conjunto de datos médicos, es decir, es esta la forma en la que los usuarios tienen acceso a la información generada mediante el proceso de extracción y generación de conocimiento.

Todos y cada uno de los servicios que componen esta API se encuentran documentados en línea en el siguiente enlace¹¹².

- Los contenedores que se muestran en color gris son los encargados de extraer y generar la información contenida en el conjunto de datos.
 - > Mediante el servicio DAP (*DISNET Autorization Process*) el sistema autoriza las consultas solicitadas por los usuarios a través de sus TOKEN de acceso. En el momento en el que un usuario hace una consulta, DAP verifica que el TOKEN se encuentre autorizado, continuando el procesamiento de la consulta en caso afirmativo, y denegándolo en caso contrario.
 - > Por otro lado, nos encontramos los procesos de extracción de textos, mediante los cuales se recuperan los textos y demás elemen-

¹¹² <http://disnet.ctb.upm.es/apis/disnet>

3.2.3 Componentes técnicos

tos relevantes relativos a las manifestaciones fenotípicas y la diagnosis de las enfermedades contenidas en las fuentes de información seleccionadas.

Una particularidad de todos los servicios de esta área es que tienen una respuesta genérica con el fin de estandarizar la comunicación y simplificar la inserción de los datos provenientes de distintas fuentes.

Los servicios que lo integran son:

- DLEP (*Disease List Extraction Process*), que obtiene la lista de artículos Wikipedia mediante DBpedia (Lehmann et al., 2015).
 - WTE (*Wikipedia Text Extraction*), que consume la lista obtenida por DLEP y se conecta con cada elemento de la lista para recuperar los elementos relevantes.
 - PMTE (*PubMed Text Extraction*), que es el servicio que obtiene textos relevantes de PubMed.
 - MCTE (*Mayo Clinic Text Extraction*), que es el servicio que obtiene textos relevantes de Mayo Clinic.
 - El último servicio de esta área es simbólico, indicando que cuando se desee agregar una nueva fuente se deberá crear un nuevo servicio e integrarlo al sistema.
- > En la parte inferior centro se encuentran dos servicios relacionados: MM (*MetaMap Service*) y MMC (*MetaMap Client*).
- Con respecto a MM es el encargado de guardar la instalación local de MetaMap, mientras que MMC es el encargado de realizar las peticiones a MM a través de listas de textos con el fin de obtener los conceptos médicos relevantes encontrados en ellos.
- > En la parte inferior derecha se encuentra el servicio TVP que es el que filtra una lista de conceptos médicos e indica cuales de ellos son considerados como elementos de diagnóstico o manifestaciones fenotípicas y cuáles no.
- > Finalmente, en el centro del diagrama se encuentra el servicio que orquesta el funcionamiento de todos los demás mediante tareas programadas automáticamente, el servicio DSVP (*Data Storage and Validation Process*).

3.2.4 MetaMap

Es aquí donde se almacenan las instrucciones para minar las fuentes seleccionadas dos veces al mes, solicitando a los servicios contiguos que se activen y recuperen los textos para, una vez recuperados, toda la información sea almacenada de forma estructurada en el conjunto de datos médico.

Este servicio también es el encargado de enviar esta información estructurada (los textos) a MMC (*MetaMap Client*) para que sea procesada y éste la reenvíe nuevamente ya actualizada con la nueva información (los conceptos obtenidos de los textos).

Por último, la lista de conceptos de los textos recibidos de MMC son enviados a TVP para que puedan ser filtrados y nuevamente actualizado el conjunto de datos médico.

3.2.4 MetaMap

MetaMap¹¹³ es una herramienta altamente configurable que descubre en textos biomédicos en inglés conceptos médicos del tesoro UMLS¹¹⁴ mediante el uso de técnicas de NLP y lingüística computacional (Aronson, 2001b). Fue desarrollado por la Biblioteca Nacional de Medicina de Estados Unidos de América (NLM), siendo una de las aplicaciones más ampliamente aceptadas de NLP en biomedicina; según su página oficial MetaMap es utilizado para el análisis de textos en inglés.

MetaMap cuenta con dos versiones descargables para ser ejecutadas localmente, la primera es la versión regular que tiene dos opciones de descarga: i) descargar de los archivos comprimidos para su instalación o ii) descargar una imagen Docker lista para ser ejecutada, siendo la segunda una versión *lite* (MetaMapLite), que es una versión menos pesada de la versión regular. Además, la página Web oficial de MetaMap pone a disposición a todo aquel usuario registrado la opción de usarlo de manera interactiva accediendo también a todas sus funcionalidades mediante una API. Adicionalmente en la página Web se encuentra toda su documentación, información acerca de las distintas herramientas que provee y noticias sobre sus lanzamientos y actualizaciones. La última versión de MetaMap fue publicada en diciembre de 2018 (2018AA).

MetaMap analiza textos e identifica palabras o frases, las relaciona y las asigna a los conceptos UMLS ((*Unified Medical Language System*) es un estándar en cuanto de codificación ampliamente utilizado por todo tipo de aplicaciones biomédicas). Los conceptos UMLS detectados tienen asignado un CUI. El *Concept*

¹¹³ <https://metamap.nlm.nih.gov/>

¹¹⁴ <https://www.nlm.nih.gov/research/umls/index.html>

3.2.4.1 Configuración

Unique Identifier (CUI)¹¹⁵ es el identificador principal de los conceptos del metatesauro (Aronson, 2001b; Park et al., 2015). Estos conceptos tienen establecidos a que tipo de concepto semánticos pertenecen. UMLS proporciona una lista de estos tipos semánticos “*semantic types*” (ST) que se encuentran disponible online¹¹⁶. Algunos ejemplos de tipos semánticos relevantes para esta tesis y para los procesos de extracción e identificación de conceptos es por ejemplo el tipo *sosy* que hace referencia a signos o síntomas, o *dsyn* que hace referencia a desordenes y enfermedades. Además, los *semantic types* se encuentran agrupados por categorías superiores¹¹⁷. La lista de estas categorías o grupos también está disponible en línea en el siguiente enlace¹¹⁶. Uno de estos grupos es, por ejemplo: el grupo *disorders* que abarca doce *semantic types* que son relativos a enfermedades o síndromes¹¹⁶. Para ver los tipos semánticos usados en esta tesis como parte del proceso de extracción/identificación de términos relevantes se recomienda ver la sección 3.2.4 MetaMap.

Finalmente, como indicativo de la eficiencia de la herramienta cabe destacar que MetaMap ha sido evaluada en el pasado en multitud de contextos de extracción de información clínica/médica con buenos resultados (Al-Safadi, Alomran & Almutairi, 2013; Osborne, Gyawali & Solorio, 2014; Rodríguez González et al., 2018; Reátegui & Ratté, 2018) y que ha sido y sigue siendo utilizada por una gran cantidad de investigaciones científicas relevantes en la generación del conocimiento médico (van Driel et al., 2006; Gooch & Roudsari, 2011; Leaman, Islamaj Dogan & Lu, 2013; Rappaport et al., 2014; Cui et al., 2014; Schlegel, Crouner & Elkin, 2015; Mouriño-García, Pérez-Rodríguez & Anido-Rifón, 2017; Guo et al., 2018).

3.2.4.1 Configuración

En esta sección se describirá la configuración completa usada durante el proceso de extracción y generación de conocimiento del sistema desarrollado en esta tesis.

La versión¹¹⁸ de MetaMap utilizada y con la que se han extraído todos los términos médicos de textos en esta tesis es la 2016v2¹¹⁹; Esta versión se ejecuta, por defecto, con las siguientes opciones de datos: **i)** fuente de conocimiento “*Knowledge Source (-Z)*”: “2016AA”, **ii)** versión de datos “*Data Version (-V)*”: “USABase” y, **iii)** modelo de datos “*Data Model (-A)*”: “*Strinct Model*”. Además, MetaMap inicia el proceso de *Word Sense Disambiguation Server*, que sirve para determinar cuál es la mejor opción en los casos en los cuales MetaMap mapea dos o

¹¹⁵ https://www.nlm.nih.gov/research/umls/new_users/glossary.html#c

¹¹⁶ <https://metamap.nlm.nih.gov/SemanticTypesAndGroups.shtml>

¹¹⁷ <https://semanticnetwork.nlm.nih.gov/>

¹¹⁸ <https://metamap.nlm.nih.gov/DescriptionOfDataVersions.shtml>

¹¹⁹ <https://metamap.nlm.nih.gov/Sources.shtml>

3.2.4.1 Configuración

más conceptos para una sola entidad reconocida en el texto (Aronson, 2001b). La configuración utilizada por defecto es explicada a continuación:

En la documentación de la versión 2016 de Metamap, la *MetaMap2016 Usage Notes*¹²⁰ (MMUN), se encuentran todas las opciones disponibles de esta herramienta.

Ahora bien, la configuración propuesta en esta tesis para MetaMap incluye:

- En primer lugar, la lista de tipos semánticos –ver sección 3.2- que, según en el MMUN se establece con la instrucción “-J”.

Con esta lista se le indica a MetaMap que busque en el texto todos aquellos conceptos UMLS que sean de los *semantic types* indicados previamente. La lista completa de los *semantic types* usados en esta tesis se encuentra descrita en detalle en la Tabla 1.

- En segundo lugar, la selección de un vocabulario fuente específico que se encuentre dentro de la base de datos de UMLS (la lista completa de estos vocabularios se encuentra en el siguiente enlace¹²¹); es la instrucción -R la que, según el manual MMUN, establece el vocabulario fuente que se quiere utilizar.

El vocabulario fuente seleccionado es SNOMEDCT_US¹²² que integra la versión internacional de SNOMED CT, incluyendo la estadounidense.

SNOMED CT proporciona una terminología general básica para el Registro Electrónico de Salud (EHR, por sus siglas en inglés) y los conceptos tienen significados únicos y definiciones formales basados en una lógica organizadas en jerarquías.

SNOMED CT incluye más de 300.000 conceptos únicos y hay más de 1.000.000 de descripciones, incluyendo sinónimos que pueden ser utilizados para referirse a un concepto; Además, hay aproximadamente 903.000 enlaces o relaciones semánticas entre los conceptos de SNOMED CT.

- El tercer y último elemento de configuración a tener en consideración es el uso del *Word Sense Disambiguation Server* que se activa según en el MMUN estableciendo la instrucción “-y”.

Es importante destacar que en el proceso desarrollado como parte de la plataforma DISNET, la configuración de MetaMap (o de cualquier otra herramienta

¹²⁰ https://metamap.nlm.nih.gov/Docs/MM_2016_Usage.pdf

¹²¹ <https://www.nlm.nih.gov/research/umls/sourcereleasedocs/index.html>

¹²² https://www.nlm.nih.gov/research/umls/sourcereleasedocs/current/SNOMEDCT_US/index.html

3.2.4.1 Configuración

de NLP que se pueda utilizar como sustitutiva de esta) puede cambiarse su configuración en base de las necesidades específicas en el proceso de extracción del conocimiento.

Son dieciséis los tipos semánticos que se usan como parte de la configuración actual descrita en esta tesis y se pueden observar en la siguiente tabla. Estos tipos semánticos están relacionados, todos ellos, con el diagnóstico y las manifestaciones fenotípicas de las enfermedades.

El amplio número de tipos semánticos definido es necesario para obtener el mayor número de manifestaciones fenotípicas y elementos de diagnósticos posibles. Si no se contara con todos ellos, los elementos a descubrir estarían muy limitados.

Tabla 1. Lista de *semantic types* relevantes para esta tesis.

SEMANTIC TYPE	ABBREVIATION	CODE	GROUP	GROUP CODE
Acquired Abnormality	acab	T020	Disorders	DISO
Anatomical Abnormality	anab	T190	Disorders	DISO
Cell or Molecular Dysfunction	comd	T049	Disorders	DISO
Congenital Abnormality	cgab	T019	Disorders	DISO
Disease or Syndrome	dsyn	T047	Disorders	DISO
Experimental Model of Disease	emod	T050	Disorders	DISO
Finding	findg	T033	Disorders	DISO
Mental or Behavioral Dysfunction	mobd	T048	Disorders	DISO
Neoplastic Process	neop	T191	Disorders	DISO
Pathologic Function	patf	T046	Disorders	DISO
Sign or Symptom	sosy	T184	Disorders	DISO
Laboratory or Test Result	lbtr	T034	Phenomena	PHEN
Clinical Attribute	clna	T201	Physiology	PHYS
Mental Process	menp	T041	Physiology	PHYS
Diagnostic Procedure	diap	T060	Procedures	PROC
Laboratory Procedure	lbpr	T059	Procedures	PROC

Finalmente, se describe un ejemplo de la forma en que MetaMap, con la configuración descrita anteriormente, encuentra conceptos médicos en los diferentes textos de enfermedad.

Se ha escogido un fragmento del texto contenido en la sección *Signs and symptoms* del artículo Wikipedia relativo al *Cerebral salt-wasting syndrome*¹²³, extraído el 1 de febrero de 2018; el texto establece lo siguiente:

“Advanced symptoms include muscle cramps, lightheadedness, dizziness or vertigo, feelings of anxiety or panic (not mentally induced), increased heart

¹²³ https://en.wikipedia.org/w/index.php?title=Cerebral_salt-wasting_syndrome&oldid=768604979

3.2.5 Proyecto DISNET

rate or slowed heart rate, low blood pressure and orthostatic hypotension sometimes resulting in fainting”.

Después de procesar el texto con MetaMap se ha obtenido el siguiente resultado:

*Advanced **symptoms** [sosy Symptom: C1457887] include muscle cramps [sosy Muscle cramp: C0026821], lightheadedness [sosy Dizziness: C0012833], dizziness [sosy Dizziness: C0012833] or vértigo [sosy Vertigo: C0042571], feelings [menp Emotions: C0013987] of anxiety [mobd Anxiety: C0003467] or panic [fndg Panic: C0030318] (not mentally induced), increased heart rate [patf Tachycardia: C0039231] or slowed heart rate [patf Bradycardia: C0428977], low blood pressure [fndg Hypotension: C0020649] and orthostatic hypotension [fndg Orthostatic hypotension: C0020651] sometimes resulting in fainting [sosy Syncope: C0039070].*

En negrita se indican las palabras o conjuntos de palabras (n-gramas) que MetaMap obtiene del texto así como sus tipos semánticos asociados. Como se observa, se han obtenido cuatro conceptos médicos cuyo tipo semántico es *fndg* (*findings*) relativo a las manifestaciones fenotípicas de esa enfermedad, de forma que de haberse excluido el tipo semántico, no hubieran sido detectados.

Por ello es tan importante incluir el máximo número de tipos semánticos (dentro de la lógica de la información que se pretende detectar), pues es preferible excedernos en la cantidad de información y depurar de ella la que no es relevante al objeto de esta tesis, que limitar el conocimiento relativo a las distintas enfermedades.

3.2.5 Proyecto DISNET

Todo lo descrito hasta ahora como parte de la metodología tiene como objetivo la creación del sistema DISNET (*Drug repositioning and **DIS**ease understanding through complex **NET**works creation and analysis*). DISNET es un sistema que hace referencia a un proyecto que pretende dar un paso más en la mejora del conocimiento humano sobre las enfermedades, no sólo para buscar y analizar las distintas conexiones entre todas ellas, sino también y, sobre todo, con el objetivo principal de encontrar una solución real a aquellas enfermedades que no tienen cura mediante estrategias de reposicionamiento de fármacos. El proyecto DISNET, del cual se deriva esta tesis, ha sido financiado en la convocatoria de Proyectos I+D de Retos orientados a la sociedad del año 2018 (RTI2018-094576-A-I00).

3.3 Enfoque de extracción de información histórica de los artículos sobre enfermedades en Wikipedia

Este proyecto ha creado un marco de trabajo para generar y analizar una gran red de enfermedades en la que se concebirán tres capas importantes: la primera capa referente a los datos de enfermedades y síntomas (capa fenotípica); la segunda capa referente a los datos biológicos de esas enfermedades (genes, proteínas, rutas metabólicas, etc. Asociados); y la tercera capa referente a los datos de los fármacos previstos para esas enfermedades; Para establecer, posteriormente, relaciones entre las enfermedades y los elementos de las diferentes capas.

El resultado será una red multicapa compleja, denominada “Red de enfermedades multicapa” o MLDN, que permitirá el análisis de las relaciones entre enfermedades: que elementos comparten a nivel fenotípico, a nivel biológico y a nivel farmacológico.

3.3 Enfoque de extracción de información histórica de los artículos sobre enfermedades en Wikipedia

En esta sección se describe un enfoque que permite extraer la información histórica referente a las manifestaciones y al diagnóstico fenotípico contenida en los artículos Wikipedia catalogados como enfermedades humanas. Estos artículos Wikipedia son los mismos que se han obtenido mediante el proceso de extracción y generación de conocimiento (KEGP) descrito en la sección 3.2.

3.3.1 Justificación

La necesidad de un nuevo enfoque de extracción de información en Wikipedia se debe, en primer lugar, a que Wikipedia es la fuente central de estudio en esta tesis, y en segundo lugar, porque uno de los propósitos de esta tesis es, precisamente, analizar cómo ha evolucionado el contenido médico de los artículos catalogados como enfermedades en la versión inglesa de Wikipedia y verificar que su información es competente para ser utilizada en enfoques de análisis relativos a enfermedades al permitir caracterizar patologías

Éste análisis se centra específicamente en tres elementos: **i)** la evolución del número de caracteres, **ii)** de las referencias y **iii)** de los términos médicos.

Sin embargo, los términos médicos no son las únicas fuentes de información, ya que también pueden ser de interés otras secciones además de las que se refieren a las manifestaciones y al diagnóstico fenotípico.

Para ilustrar esta limitación hay que observar la evolución del artículo Wikipedia sobre la enfermedad *Otic polyp*¹²⁴. En la Figura 16 se muestran dos capturas

¹²⁴ https://en.wikipedia.org/wiki/Otic_polyp

3.3.1 Justificación

de pantalla que corresponden respectivamente a la captura de conocimiento (“*snapshot*”, de aquí en adelante) realizado por el KEGP el 1 de febrero de 2019 y a la del 15 de febrero de 2019.

Figura 16. Comparación de dos revisiones del artículo de enfermedad Wikipedia sobre *Otic polyp*. a) Revisión del artículo que corresponde con la captura de conocimiento (*snapshot*) de Wikipedia del primero de febrero de 2019 contenida en el conjunto de datos propio; b) Revisión del artículo que corresponde con la captura de conocimiento (*snapshot*) de Wikipedia del quince de febrero de 2019 contenida en el conjunto de datos propio.

Se observa que el texto es el mismo en ambos casos, sin embargo, en el conjunto de datos se detecta un importante aumento en el número de caracteres en los textos de las secciones relevantes de la segunda captura con respecto a la primera –véase Figura 17. La razón reside en cómo se organiza la información: mientras que en la primera captura de conocimiento sólo se reconoce como relevante una sección “*Signs and symptoms*” y se descarta la información de “*Imaging findings*”, “*Pathology findings*” y “*Differential diagnoses*”; ahora se debe observar que estas tres últimas secciones se han agrupado en una sola sección en la segunda captura

3.3.2 Proceso para recuperar la información histórica

de conocimiento, es decir, se ha refundido todo su contenido en la sección “*Diagnosis*”, de forma que la información contenida en estas tres secciones refundidas, en la nueva “*snapshot*”, ha sido procesada por KEGP. Por lo tanto, una reestructuración del contenido, en la que sólo se mueven textos, es interpretado erróneamente en el análisis para esta tesis como un aumento de la cantidad de información disponible.

DIS000004	Otic polyp	2019-01-15	2019-01-01	143	0
DIS000004	Otic polyp	2019-02-01	2019-01-15	143	0
DIS000004	Otic polyp	2019-02-15	2019-02-01	1227	1084
DIS000004	Otic polyp	2019-03-01	2019-02-15	1227	0

Figura 17. Identificación gráfica de la limitación del contenido del conjunto de datos de cara al análisis de la evolución propuesto en esta sección. Se muestra un incremento anormal del número de caracteres de los textos de las secciones relevantes en el artículo Wikipedia de la enfermedad *Otic polyp*, entre las capturas de conocimiento (“*snapshot*”) del 1 de febrero de 2019 y del 15 de febrero de 2019.

Y no es que sea un error sino que los textos en los que se basa en un principio esta tesis para realizar el análisis de la evolución del contenido médico en Wikipedia no son suficientes.

Los primeros textos solo ofrecen una visión reducida del artículo y es necesario tener una visión global y total de todos ellos. Y es que con la medida de la totalidad de cada artículo es posible observar cuando han ocurrido verdaderos cambios en el contenido, es decir, cuando se ha eliminado o agregado conocimiento.

Para alcanzar éste objetivo particular y tomando en cuenta la limitación descrita con anterioridad, es necesario completar la información existente en el conjunto de datos médicos generado por el KEGP.

Para completar la información se hizo uso del historial de revisiones¹²⁵, es decir, el lugar en el que se encuentran todas las actualizaciones y ediciones que se han realizado en un artículo Wikipedia a lo largo del tiempo.

3.3.2 Proceso para recuperar la información histórica

Este proceso de recuperación aprovecha el conjunto de datos médico para obtener la información inicial (la información con la que se parte). Específicamente, de cada enfermedad se tiene en cuenta: **i)** el nombre de la enfermedad, es decir, el título del artículo de Wikipedia; **ii)** el enlace al artículo de Wikipedia (URL); **iii)** y la lista de *snapshots* con marcas de tiempo de acceso asociadas. Cada una de

¹²⁵ https://en.wikipedia.org/wiki/Help:Page_history

3.3.2 Proceso para recuperar la información histórica

estas capturas de conocimiento corresponden a una revisión específica de Wikipedia, actualizada con una frecuencia quincenal, y que abarca desde el 1 de febrero de 2018 hasta el 15 de julio de 2019 (18 meses, 36 *snapshots* de extracción de información de Wikipedia) en el actual momento de análisis.

Con la información descrita anteriormente es posible acceder al contenido histórico de los artículos de Wikipedia a través del historial de revisiones, que abarca todas las actualizaciones o ediciones que se han realizado sobre un artículo específico, en una fecha determinada y por un usuario determinado. Wikipedia identifica los artículos por su título o id de página¹²⁶ y, de manera similar, las versiones anteriores de cada documento se identifican por un id de revisión único¹²⁷. A modo de ilustración, el siguiente enlace “https://en.wikipedia.org/wiki/Otic_polyp” lleva al artículo sobre la enfermedad “*Otic polyp*”; y el enlace “https://en.wikipedia.org/w/index.php?title=Otic_polyp&action=history” lleva a su historial de revisiones. El enlace “https://en.wikipedia.org/w/index.php?title=Otic_polyp&oldid=877794912” lleva a una revisión específica del artículo, en este caso la revisión es la del 10 de enero de 2019. La Figura 18 muestra cómo se muestran los enlaces descritos anteriormente en Wikipedia. En este ejemplo se destacan dos elementos clave para llevar a cabo el proceso de recuperación de información histórica: **a)** el título del artículo (“*title*”) y **b)** el id de la revisión (“*oldid*”). Aun así, en el conjunto de datos sólo se encuentra el primer elemento clave, pero no la lista completa de los ids de las revisiones, de manera que se ha tenido que desarrollar un proceso adicional.

Se evaluaron dos posibles enfoques para acceder al contenido histórico de los artículos (historial de revisiones): por un lado, un proceso de *Web scraping* similar al desarrollado por para Wikipedia en la sección 3.1.1; y por otro, el uso de la API de Wikipedia en inglés¹²⁸, un servicio Web de la MediaWiki API¹²⁹, de la familia de la Fundación Wikimedia, Inc.¹³⁰.

La API de Wikipedia permite realizar una gran variedad de consultas de una forma sencilla, en comparación con el proceso de *Web scraping* que, para alcanzar los mismos resultados, requerirá un considerable esfuerzo de desarrollo. En este contexto, es necesario dejar claro que la API de Wikipedia solo facilita el acceso al historial de revisiones de los artículos. Porque una vez recuperado el contenido entero del artículo es imprescindible procesarlo, tarea que se logra mediante el *Web scraping*.

Lo que verdaderamente no es viable es usar *Web scraping* para acceder al historial de revisiones. Ya que como se muestra en la Figura 18 navegar solo con *Web scraping* a través de cada elemento del historial de revisiones y así mismo

¹²⁶ https://en.wikipedia.org/wiki/Template:Pageid_to_title

¹²⁷ https://en.wikipedia.org/wiki/Wikipedia:Revision_id

¹²⁸ <https://en.wikipedia.org/w/api.php>

¹²⁹ https://www.mediawiki.org/wiki/API:Main_page/en

¹³⁰ https://es.wikipedia.org/wiki/Fundaci%C3%B3n_Wikimedia

3.3.2 Proceso para recuperar la información histórica

por las páginas que los contienen, supone usar muchos más recursos de cómputo y tiempo.

The figure consists of three screenshots of the Wikipedia article 'Otic polyp'.

a) The first screenshot shows the current version of the article as of September 2019. The title is 'Otic polyp'. The text states: 'An **otic polyp** is a benign proliferation of chronic inflammatory cells associated with **granulation tissue**, in response to a longstanding inflammatory process of the **middle ear**.^{[1][2]}'

b) The second screenshot shows the 'Revision history' of the article. It lists several revisions, including one by 'JCW-CleanerBot' on 23:11, 10 January 2019. The revision summary for this edit is: '(-top: task, replaced: Archives of otolaryngology--head & neck surgery -> Archives of Otolaryngology) (Tag: AWB)'.

c) The third screenshot shows the content of the specific revision from January 2019. It includes a warning box: 'This is an old revision of this page, as edited by JCW-CleanerBot (talk | contribs) at 23:11, 10 January 2019 (-top: task, replaced: Archives of otolaryngology--head & neck surgery -> Archives of Otolaryngology is a permanent link to this revision, which may differ significantly from the current revision.)'. Below the warning, the text reads: 'An **otic polyp** (also called **aural polyp**) is a benign proliferation of chronic inflammatory cells associated with **granulation tissue**, in response to a longstanding inflammatory process of the **middle ear**.'

Figura 18. Seguimiento para ver una revisión del artículo de Wikipedia sobre la enfermedad *Otic polyp*. a) Versión actual de este artículo Wikipedia, a septiembre de 2019; b) historial de revisiones del artículo en septiembre de 2019; y c) contenido de una de sus revisiones específicas (la del 10 de enero de 2019).

Por ello, el enfoque de extracción histórica más efectivo es el uso de la API de Wikipedia, requiriéndose para ello dos consultas diferentes que se describen a continuación:

> La primera consulta (**WQone**) es:

3.3.2 Proceso para recuperar la información histórica

<https://en.wikipedia.org/w/api.php?action=query&prop=revisions|redirects&format=json&rvprop=ids|flags|timestamp|userid|user|size|comment&rvstart=2019-01-15T00:00:00Z&rvdir=older&rvlimit=1&redirects&titles=Otic%20polyp>

Mediante la consulta **WQone** se recupera la información básica que Wikipedia posee con respecto a una enfermedad concreta en un tiempo pasado, de forma que, especificando en dicha consulta una fecha concreta, ésta te devuelve la información contenida en dicha fecha. En caso de que no se tuviera información de la fecha solicitada, devolverá la última información obtenida antes de la misma.

La configuración de la consulta consta de seis elementos básicos: En primer lugar, es necesario especificar que la acción a realizar es una consulta, es decir, buscar información sobre una materia de forma que, a través del parámetro *action*¹³¹ y su valor *query*, se obtiene; En segundo lugar, es necesario especificar el objeto de la consulta, es decir, sobre qué asunto se requiere esa información y, a través del parámetro *prop* y su valor *revisions*, se indica que, en este caso, la consulta se enfoca en información sobre las revisiones. Además *prop* soporta más de un valor añadiendo el carácter “|”: en este caso a la consulta se le ha añadido el valor *redirects* el cual permite encontrar la revisión a pesar de que el nombre del artículo indicado en el parámetro *titles* haya cambiado.

En tercer lugar, se deben proporcionar un conjunto de parámetros exclusivos de consultas con *prop = revisions*¹³²:

- 1) **rvprop**, que especifica la lista de atributos a devolver: {(*ids*, el id de la revisión), (*flags*, los flags de la revisión "minor"), (*timestamp*, indica el timestamp de la revisión), (*userid* , el id de usuario que hizo de la revisión), (*user*, el nombre del usuario que hizo la revisión), (*size*, la longitud (bytes) de la revisión) y (*comment*, el comentario proporcionado por el usuario para la revisión)}.
- 2) **rvstart**, que define la fecha de revisión; nótese que, si no se especifica nada aquí, la consulta devolverá información sobre la revisión definida por el parámetro *rvdir*. La fecha concreta que se introduce en este parámetro es la *snapshot*.
- 3) **rvdir**, que especifica cómo deben ordenarse las revisiones; En este caso se utiliza el valor más antiguo “*older*”, a fin de obtener la revisión más antigua antes de la fecha prevista.

¹³¹ <https://en.wikipedia.org/w/api.php?action=help&modules=main>

¹³² <https://www.mediawiki.org/wiki/API:Revisions>

3.3.2 Proceso para recuperar la información histórica

- 4) **rvlimit**, es el número de revisiones a devolver, en este caso solo interesa una, la correspondiente a la “*snapshot*” indicada en el parámetro **rvstart**.

En cuarto lugar, es necesario especificar el parámetro **titles** cuyo valor es el nombre del artículo Wikipedia, es decir, el nombre de la enfermedad. Además **titles** soporta el ingreso de más de un nombre de artículo mediante el uso del carácter “|” que actúa como separador.

En quinto lugar, se especifica el parámetro **redirects**, que es el que permite visualizar, en el caso que existan, las redirecciones del artículo o sus cambios de nombre (un caso muy común en Wikipedia).

Y por último, el parámetro **format**, en el que se indica el formato de la respuesta, que para este caso siempre será JSON.

Para más información sobre las consultas sobre revisiones Wikipedia y el tipo de consulta **action = query**, se encuentra disponible online la documentación pertinente¹³³, ¹³⁴.

En la Figura 19 se muestra un ejemplo de una respuesta JSON de la consulta **WQone**.

No obstante, a pesar de que con la consulta **WQone** es posible la recuperación textual completa del artículo solicitado, dicha recuperación no tiene una estructura ni un formato sencillo, de forma que la información que nos devuelve de un tiempo pasado se basa en un texto en bloque, no siendo sencilla su explotación.

Por ello, se desarrolló la segunda consulta que viene a solucionar dicho inconveniente:

> La segunda consulta (**WQtwo**) es:

<https://en.wikipedia.org/w/api.php?action=parse&format=json&prop=sections|text&oldid=877794912>

Mediante la consulta **WQtwo** también se puede recuperar el contenido textual completo de una revisión concreta de un artículo de Wikipedia, al igual que con la consulta anterior, con la ventaja de que, con esta segunda consulta, también se obtiene la lista completa de las secciones que lo componen.

La particularidad más significativa para elegir este segundo tipo de consulta es usar el parámetro **action = parse**, en el cual por defecto el contenido textual es proporcionado en formato HTML, haciendo mucho más sencilla el uso de técnicas

¹³³ <https://www.mediawiki.org/wiki/API:Revisions>

¹³⁴ <https://www.mediawiki.org/wiki/API:Query>

3.3.2 Proceso para recuperar la información histórica

de *Web scraping* sobre este tipo de estructura. El contenido textual esta originalmente construido como un wikitext¹³⁵, un texto elaborado mediante un lenguaje de marcado especial para la creación de páginas wikis¹³⁶.

```
1.  {
2.    "continue": {
3.      "rvcontinue": "20160527211659|722399173",
4.      "continue": "||redirects"
5.    },
6.    "query": {
7.      "pages": {
8.        "35510049": {
9.          "pageid": 35510049,
10.         "ns": 0,
11.         "title": "Otic polyp",
12.         "revisions": [
13.           {
14.             "revid": 877794912,
15.             "parentid": 722399173,
16.             "minor": "",
17.             "user": "JCW-CleanerBot",
18.             "userid": 31737083,
19.             "timestamp": "2019-01-10T23:11:39Z",
20.             "size": 5761,
21.             "comment": "/* top */[[User:JCW-CleanerBot#Logic|task]], replaced: Archives of otolaryngology--head & neck surgery → Archives of Otolaryngology–Head & Neck Surgery"
22.           }
23.         ],
24.         "redirects": [
25.           {
26.             "pageid": 35512271,
27.             "ns": 0,
28.             "title": "Otic Polyp"
29.           }
30.         ]
31.       }
32.     }
33.   }
34. }
```

Figura 19. Respuesta JSON de la consulta WQone usando como elementos de búsqueda: *titles = Otic polypy rvstart = 2019 – 01 – 15.*

La configuración de esta consulta consta de cuatro elementos básicos: En primer lugar, es necesario especificar que la acción a realizar con la consulta es el procesamiento de datos, por ello indicamos en la consulta el parámetro *action = parse* para habilitar la opción descrita.

¹³⁵ <https://en.wikipedia.org/wiki/Help:Wikitext>

¹³⁶ <https://en.wikipedia.org/wiki/Wiki>

3.3.2 Proceso para recuperar la información histórica

En segundo lugar, es necesario especificar las propiedades o elementos del artículo que se quieren obtener; Para ello se indican en forma de lista los elementos de interés por medio del parámetro *prop = sections|text*; En este caso, dado que se quiere obtener dos valores, se han introducido dos parámetros: *sections* (secciones) y *text* (el contenido textual del artículo).

En tercer lugar, es necesario indicar el parámetro *format = json* para establecer el formato de respuesta de la consulta.

Finalmente, es necesario indicar el parámetro *oldid*, ligado al id de la revisión del artículo (obtenido con la consulta WQone).

El ejemplo de la respuesta de la consulta WQtwo se muestra en la Figura 20.

```
1. {
2.   "parse": {
3.     "title": "Otic polyp",
4.     "pageid": 35510049,
5.     "revid": 877794912,
6.     "text": "...",
7.     "sections": [
8.       {
9.         "toclevel": 1,
10.        "level": "2",
11.        "line": "Signs and symptoms",
12.        "number": "1",
13.        "index": "1",
14.        "fromtitle": "Otic_polyp",
15.        "byteoffset": 1265,
16.        "anchor": "Signs_and_symptoms"
17.      },
18.      {
19.        "toclevel": 1,
20.        "level": "2",
21.        "line": "Imaging findings",
22.        "number": "2",
23.        "index": "2",
24.        "fromtitle": "Otic_polyp",
25.        "byteoffset": 1466,
26.        "anchor": "Imaging_findings"
27.      },
```

Figura 20. Respuesta JSON de la consulta WQtwo usando como elemento de búsqueda: *oldid = 877794912*.

3.4 Conclusiones

En la Figura 21 se muestra el flujo de trabajo completo del proceso de extracción de información histórica de Wikipedia. El flujo inicia con el módulo encargado de recuperar del conjunto de datos médico, la lista de enfermedades (nombre e identificación de enfermedades) y su respectiva lista de *snapshots*. A continuación, un módulo diferente consume el API de Wikipedia mediante la ejecución de las consultas **WQone** para cada enfermedad y **WQtwo** para cada una de sus *snapshot*. Además, se calcula por cada *snapshot* de enfermedad el número de caracteres y de referencias. Finalmente todos los datos recuperados se guardan en archivos JSON para ser procesados posteriormente

Toda la información histórica recolectada se analiza en profundidad en el Capítulo 4 .

Figura 21. Flujo de trabajo del proceso de extracción de información histórica de artículos Wikipedia.

3.4 Conclusiones

Como conclusión, en lo relativo a lo que a la metodología se refiere, podemos asentar las siguientes afirmaciones:

1. Es perfectamente viable la extracción de conocimiento médico de fuentes de información públicas, libres y textuales tales como Wikipedia, PubMed o Mayo Clinic y, sobre todo, que en dichas fuentes de información se puede hallar una gran cantidad de conocimiento médico relevante.

Además en el caso concreto de Wikipedia y Mayo Clinic, al remitirse a fuentes externas que contrastan la información y la asocian con otras fuen-

3.4 Conclusiones

tes de datos, vocabularios y taxonomías, existe la posibilidad de poder asociar y completar la información de estas fuentes de información con las fuentes externas a las que se remiten.

Así, los artículos contenidos en Wikipedia cuentan con códigos de enfermedades en fuentes externas que permiten relacionar estos códigos de enfermedades en Wikipedia con enfermedades en vocabularios o taxonomías como MeSH, OMIM, ICD-10 entre otras.

No ocurre lo mismo con las enfermedades contenidas en Mayo Clinic, que se encuentran completamente aisladas en términos de mapeos sin posibilidad de ser comparada su información con otras fuentes de información, al menos mediante códigos de vocabularios, resultando por lo tanto en una desventaja en este contexto la información ofrecida por dicha fuente.

2. Mediante enfoques de procesamiento de lenguaje natural (NLP) o de minería de texto es posible transformar la información textual de las fuentes de información analizadas en datos completamente estructurados, permitiendo su posterior análisis de una forma mucho más sencilla y cómoda.
3. Que mediante la integración de diferentes fuentes de información es posible crear un conjunto de datos relativos a enfermedades, signos y síntomas, así como elementos de diagnóstico, sin un gran perjuicio de las complejidades que conlleva el mapeo de dichas enfermedades por provenir de distintas fuentes de información, o el hecho de que las expresión y redacción de los textos son heterogéneos.
4. El hecho de que la información médica recuperada se encuentre siempre actualizada es un gran punto a favor, además de que obtener conocimiento médico de distintas fuentes de información supone un enriquecimiento en lo que a los datos médicos se refiere. Además, esta actualización permite estudiar la evolución del conocimiento y podría evaluarse y valorar si esta evaluación es positiva o negativa.
5. La simplicidad y una estandarización cada vez más rigurosa en la estructura de los artículos Wikipedia han resultado ser elementos clave para la recuperación del contenido médico relevante para esta tesis.
6. A pesar de los juicios a los que se ha visto sometida Wikipedia, en especial por la desconfianza que genera que los artículos que contiene puedan ser editados por cualquier persona, es actualmente una fuente de información que ha mejorado notablemente con el transcurso de tiempo, integrando nuevos y mejores mecanismos que han hecho que el contenido de sus artículos de ciencia, y en especial de salud y enfermedades sea más robusto y válido, tal y como la literatura científica ha demostrado.

3.4 Conclusiones

7. La herramienta MetaMap es sin duda una referencia en cuanto a la obtención de conocimiento médico se refiere; no obstante, supondría una mejora en el sistema de generación de conocimiento médico la implementación de otras herramientas NLP, como Apache cTakes u otras de desarrollo propio.

Capítulo 4 Resultados

En este capítulo se describen los diferentes resultados de los experimentos y análisis realizados para el presente documento de tesis y con los cuales se han permitido elaborar una serie de caracterizaciones en las enfermedades del conjunto de datos DISNET en base a sus elementos de diagnóstico. Se hace especial énfasis en los análisis desarrollados sobre la información recuperada de Wikipedia.

En primer lugar, se describen algunos conceptos básicos que son importantes para la comprensión correcta del presente capítulo: *snapshot*, enfermedad y lista de enfermedades.

En segundo lugar, se describe la información recuperada de las distintas fuentes de información seleccionadas que conforma el conjunto de datos médicos desarrollado para esta tesis.

Se analizan y comparan los elementos recuperados como: los enlaces a los documentos Web, número de enfermedades, número de textos, secciones relevantes de contenido diagnóstico, número de términos médicos obtenidos, enlaces en los textos, entre otros. Esta observación sobre los elementos recuperados permite observar que el hecho de tener un software de NLP posibilita caracterizar las enfermedades porque extrae información que las define.

En tercer lugar, se describe en detalle el proceso de validación del sistema de extracción y generación de conocimiento diagnóstico –ver sección 3.2–. Se contabilizan verdaderos positivos, falsos negativos, falsos positivos y verdaderos negativos, con el fin de calcular la precisión que tiene el sistema de obtener conocimiento diagnóstico.

En cuarto lugar, se describe un análisis descriptivo que permite caracterizar la información de enfermedades humanas recuperada de Wikipedia, PubMed y Mayo Clinic. Además de una comparativa entre ellas.

En quinto lugar, se describe un análisis sobre los identificadores de enfermedades recuperados de las diferentes fuentes de información. Enfocando los esfuerzos en observa la conectividad y relaciones que se pueden llegar a crear con otros vocabularios, taxonomías o bases de datos externas a DISNET.

En sexto lugar, se describen varios análisis sobre la evolución de la información médica contenida en los artículos Wikipedia que permite definir mejor las enfermedades. Los análisis están basados en la evolución incremental de las métricas: el número de caracteres, términos y referencias relevantes en cada página.

Finalmente, se describe el proceso que define las enfermedades en base a sus similitudes fenotípicas. Estas similitudes utilizadas en la presente tesis, sirven para

4.1 Construcción del conjunto de datos

que posteriormente se construya una red de enfermedades que relaciona enfermedades con enfermedades en base a sus características fenotípicas.

4.1 Construcción del conjunto de datos

En esta sección se describirán conceptos básicos para la comprensión correcta de los resultados presentados en este capítulo.

Además se presenta una descripción general que ha permite caracterizar las enfermedades en base a los elementos de diagnóstico contenidos en el conjunto de datos médicos DISNET.

4.1.1 Snapshot

Una *snapshot* es una captura de conocimiento que se ha ejecutado en una fecha determinada para recuperar información sobre enfermedades de una de las fuentes biomédicas relevantes para esta tesis.

4.1.2 Enfermedad

Es de vital importancia describir que documentos son considerados como enfermedades dentro del conjunto de datos médicos. En la Figura 22 están las entidades que permiten identificar documentos como enfermedades.

Una entidad *document* es una enfermedad:

- i. En primer lugar, cuando tiene asociados términos médicos relevantes que han sido evaluados por el proceso de TVP como verdaderos elementos de diagnóstico. Es de recordar que estos términos son los que han encontrado en los textos de las secciones relevante –ver Capítulo 3 –. Esto se consigue por medio de la conjunción de las entidades *section*, *text*, *symptom* y especialmente por la entidad de relación *has_symptom* que es en donde se indica el resultado de la evaluación del proceso TVP para cada termino encontrado.
- ii. Y en segundo y último lugar, cuando la entidad relacionada *disease* tiene en su campo *relevant* el valor *true* –ver Figura 22–. Este campo indica según una verificación manual si un artículo se refiere a una enfermedad humana o no.

La verificación es un filtro que consiste en buscar y despreciar manualmente nombres de artículos irrelevantes como: “*Impaction (animals)*”,

4.1.2 Enfermedad

“*Equine shivers*”, “*Zoonosis*”, “*Histiocytoma (dog)*”, “*Elbow dysplasia*”, “*Explosion*”, entre otros nombres de enfermedades animales, eventos u otros tipos de artículos muy alejados de ser enfermedades humanas, las que son sujeto de estudio. Es importante decir que esta verificación por su naturaleza es limitada y en los trabajos futuros se harán esfuerzos para mejorarla. En el siguiente enlace¹³⁷ se encuentra una lista de aquellos artículos encontrados irrelevantes.

Figura 22. Entidades del conjunto de datos médicos que permiten determinar enfermedades DISNET. Extracto de relaciones del DMDS (*DISNET Medical Dataset*).

Este tipo de artículos irrelevantes se encuentran especialmente en Wikipedia y se deben a: 1) que en DBpedia están mal catalogados y 2) que tienen secciones relevantes durante el proceso de extracción de información pero su contenido no es médico –ver sección 3.1.1.1 –, por ejemplo, el artículo de la “*Explosion*” tiene la sección “*Causes*”¹³⁸ en la que incluso se encontró un término médico relevante y si no fuera por esta verificación manual sería considerada como una enfermedad.

En resumen, una enfermedad es aquella que tiene asociados al menos un término médico relevante (evaluado positivamente por el proceso TVP) y que tenga una evaluación positiva por la verificación manual. En la Figura 23 se muestra la consulta SQL que obtiene el número de enfermedades de la fuente PubMed.

¹³⁷ https://github.com/disnet-project/thesis/blob/master/no_diseases.tsv

¹³⁸ <https://en.wikipedia.org/wiki/Explosion>

4.1.3 Lista de enfermedades

Formar un léxico completo de enfermedades es un objetivo a alcanzar del proyecto DISNET y el trabajo de la presente tesis es el primer paso para conseguirlo.

Un léxico homogéneo es una tarea de enormes proporciones porque actualmente existen muchas fuentes biomédicas, cada una con sus propias listas de enfermedades, identificadores, jerarquías, agrupaciones, entre otras. De esta diversidad surge la complejidad de encontrar una única lista completa de enfermedades que sea capaz de integrar información de todos los aspectos de las enfermedades.

Algunos vocabularios como el UMLS, la *Disease Ontology* (DO) o MalaCards son ejemplos de esfuerzos por formar léxicos médicos homogéneos.

La lista de enfermedades en el conjunto de datos médicos DISNET es heterogénea parcialmente porque no es posible relacionar la totalidad de las enfermedades. Esto se debe, a que de algunas enfermedades tienen códigos asociados, es decir, identificadores que pertenecen a vocabularios biomédicos.

```
SELECT DISTINCT d.disease_id, d.name, COUNT(DISTINCT hsym.cui)
term_count
FROM disease d
    INNER JOIN has_disease hd ON hd.disease_id = d.disease_id
    INNER JOIN document doc ON doc.document_id = hd.document_id AND
doc.date = hd.date
    INNER JOIN has_source hs ON hs.document_id = doc.document_id AND
hs.date = doc.date
    INNER JOIN source s ON s.source_id = hs.source_id
    INNER JOIN has_section h on doc.document_id = h.document_id and
doc.date = h.date
    INNER JOIN has_text ht on h.document_id = ht.document_id and
h.date = ht.date and h.section_id = ht.section_id
    INNER JOIN text t on ht.text_id = t.text_id
    INNER JOIN has_symptom hsym on t.text_id = hsym.text_id
WHERE s.name COLLATE utf8_bin = 'pubmed'
    AND d.relevant = true
    AND hsym.validated = true
GROUP BY d.disease_id, d.name
ORDER BY d.name ASC;
```

Figura 23. Consulta SQL que obtiene la lista de enfermedades de la fuente PubMed.

Con los códigos se pueden hacer mapeos entre enfermedades que se permiten establecer relaciones con otros identificadores, por ejemplo, la correspondencia entre un identificador MeSH y uno de la *Disease Ontology* o entre un identificador DISNET (DISNET id).

4.2 Vista global

De las fuentes seleccionadas Mayo Clinic es la única que no tiene ningún tipo de códigos de enfermedades. Mientras que PubMed tiene identificadores de vocabularios como: MeSH, DO, ICD-10, SNOMED CT, CUI y OMIM. Finalmente, y considerando todas las *snapshots* de Wikipedia, se puede decir que esta fuente tiene identificadores de al menos 50 vocabularios distintos, y de los que más contribuyen con códigos son: MeSH (4.263), ICD-10 (3.627), ICD-9 (3.165), eMedicine (2.816), DiseasesDB (2.814), OMIM (2.082), MedlinePlus (1.401), entre otros.

Finalmente, con los identificadores de enfermedades es posible obtener la lista de enfermedades en DISNET, es decir, la suma de la información de todas las *snapshots* de cada fuente. En la Tabla 2 se muestran el número de enfermedades entre todas las *snapshots* de todas las fuentes, el número de enfermedades de todas las *snapshots* de cada fuente y el número de enfermedades entre Wikipedia y PubMed. Así como el número de identificadores de enfermedades y de vocabularios encada fuente.

Tabla 2. Números de enfermedades DISNET, identificadores (códigos de enfermedades), términos médicos y vocabularios.

SOURCES	DISEASES	IDENTIFIERS	VOCABULARIES
Todas las <i>snapshots</i>	7.160	35.203	76
Wikipedia (todas <i>snapshots</i>)	5.750	23.808	74
PubMed (todas <i>snapshots</i>)	2.211	15.682	5
Mayo Clinic (todas <i>snapshots</i>)	1.140	0	0
Wikipedia $\cap$ PubMed	1.024	2.771	3
Wikipedia (2020-03-01 <i>snapshot</i>)	5.563	21.920	65

Al sumar la información de todas las *snapshots* de todas las fuentes, el sistema cuenta con 7.160 enfermedades¹³⁹, 35.203 identificadores¹⁴⁰ de enfermedades provenientes de alrededor de 76 vocabularios¹⁴¹, ontologías o bases de datos biomédicas encontrados mediante el minado de Wikipedia, la DO y el vocabulario MeSH.

4.2 Vista global

Por lo tanto, la base de datos del sistema DISNET (DMDS, por sus siglas en inglés: *DISNET Medical Dataset*) contiene información recuperada de tres fuentes de información: Wikipedia, PubMed y Mayo Clinic. Y como ya se ha descrito anteriormente, el sistema se encuentra actualmente activo por lo que aún se siguen generando capturas de conocimiento de las fuentes, en particular de Wikipedia

¹³⁹ https://github.com/disnet-project/thesis/blob/master/all_disease_all_sn.tsv

¹⁴⁰ https://github.com/disnet-project/thesis/blob/master/all_codes_all_sn.tsv

¹⁴¹ https://github.com/disnet-project/thesis/blob/master/all_vocabularies_all_sn.tsv

4.2 Vista global

porque es el proveedor de información diferencial y principal de estudio de la presenta tesis.

De esta forma a principios de marzo de 2020 en la DMDS se han generado 51 *snapshots* de Wikipedia, del 1 de febrero de 2018 al 1 de marzo de 2020, de PubMed se ha generado una *snapshot*, la del 3 de abril de 2018 y de Mayo Clinic hay 38 *snapshots*, del 15 de agosto de 2018 al 1 de marzo de 2020.

Dentro del sistema es posible consultar para cada *snapshot* y fuente: el número total de artículos con términos médicos¹⁴², el número total de términos médicos encontrados¹⁴³, el número de textos procesados¹⁴⁴, el número total de códigos recuperados¹⁴⁵ y el número total de *semantic types* encontrados¹⁴⁶.

Solo para Wikipedia, a lo largo de todas las *snapshot* se ha obtenido de DBpedia una lista de 15.133 artículos catalogados como enfermedades¹⁴⁷. De las cuales solo 7.363 artículos son considerados porque al menos tienen asociado un texto relativo a conocimiento diagnóstico de la enfermedad, o tienen asociado al menos un código de un vocabulario biomédico. Y finalmente, quedan 5.750 artículos que se han considerados como enfermedades porque se encontraron en ellos al menos un concepto médico de diagnóstico relevante¹⁴⁸.

La *snapshot* de PubMed ha sido construida usando una lista de 2.355 términos MeSH relativos a enfermedades humanas¹⁴⁹. Y con estos términos MeSH se han recuperado 174.900 *abstracts*^{150, 151} de artículos científicos (alrededor del 0,663% de los aproximadamente 30 millones de artículos existentes en PubMed). Y en solo 112.515 *abstracts* se encontraron términos médicos relevantes que son referentes a 2.211 términos de MeSH enfermedades.

Finalmente, Mayo Clinic se inició a minar el 15 de agosto de 2018 y se recuperó información de 1.083 enfermedades humanas de 1.176 artículos. Como se ha mencionado anteriormente, de esta fuente solo se pueden recuperar contenido textual porque ninguno de los artículos tiene asociados ni identificadores propios ni externos.

Las configuraciones de todas las herramientas que han sido usadas para generar cada una de las *snapshots* de Wikipedia, PubMed y Mayo Clinic se encuentran disponibles en línea en el siguiente enlace¹⁵².

¹⁴² https://github.com/disnet-project/thesis/blob/master/real_disease_with_dke_sn_2020_03_01.tsv

¹⁴³ https://github.com/disnet-project/thesis/blob/master/all_dke.tsv

¹⁴⁴ https://github.com/disnet-project/thesis/blob/master/processed_text_wikipedia_sn_2020_03_01.tsv

¹⁴⁵ https://github.com/disnet-project/thesis/blob/master/codes_wikipedia_sn_2020_03_01.tsv

¹⁴⁶ https://github.com/disnet-project/thesis/blob/master/semantic_types_wikipedia_sn_2020_03_01.tsv

¹⁴⁷ https://github.com/disnet-project/thesis/blob/master/all_dbpedia_wikipedia_url_articles.tsv

¹⁴⁸ https://github.com/disnet-project/thesis/blob/master/all_real_disease_wikipedia.tsv

¹⁴⁹ https://github.com/disnet-project/thesis/blob/master/human_disease_mesh.tsv

¹⁵⁰ https://github.com/disnet-project/thesis/blob/master/retrieval_pubmed_texts_01.tsv.zip

¹⁵¹ https://github.com/disnet-project/thesis/blob/master/retrieval_pubmed_texts_02.tsv.zip

¹⁵² https://github.com/disnet-project/thesis/blob/master/configuration_tool.tsv

4.3 Validación del sistema

En esta sección se describen los resultados del proceso de validación ejecutado sobre el sistema de extracción y generación de conocimiento, con el fin de acreditar la relevancia del conocimiento diagnóstico (términos de diagnóstico médico válidos) generado a través del proceso de NLP, conformado por la herramienta MetaMap y el proceso de TVP –ver sección 3.2.1–.

Este proceso de validación solo se ha llevado a cabo con respecto a los datos de Wikipedia y PubMed, siendo suficiente para acreditar que la información extraída es relevante a efectos del presente trabajo de tesis.

Para llevar a cabo dicha validación, se seleccionaron, al azar, 100 artículos o resúmenes de artículos de cada una de las fuentes de información pública cuya única condición era que tuvieran, al menos, 20 términos médicos calificados como tal mediante el proceso de evaluación de TVP. Mediante esta única condición se asegura que el proceso de validación analiza artículos con una alta concentración de conocimiento médico.

La validación de Wikipedia se llevó a cabo sobre la *snapshot* del 1 de febrero de 2018, y la evaluación de PubMed sobre la *snapshot* del 3 de abril de 2018.

Hay que recordar la diferente naturaleza de cada una de las fuentes de información que se han sometido a la validación. Aunque se han elegido 100 artículos de cada una de ellas se debe destacar que mientras 100 artículos de Wikipedia equivalen a 100 enfermedades validadas, 100 artículos de PubMed no equivalen a 100 enfermedades, porque una misma enfermedad puede estar descrita en uno o en varios artículos. Así, validar un artículo de Wikipedia significa validar una enfermedad en su totalidad, y validar un artículo de PubMed supone validar una parte de la enfermedad a la que se refiere el mismo.

Estas *snapshots* se realizaron en diferentes momentos, y por lo tanto con diferentes configuraciones –estas últimas se pueden ver en línea en el siguiente enlace¹⁵².

Con la primera configuración realizada para llevar a cabo la validación, que incluía 11 tipos semánticos, se detectó que MetaMap no extraía todos los conceptos médicos relevantes, dejando fuera elementos como: “*Anxiety*”, “*Stress*”, “*Amnesia*”, “*Bulimia*”, entre otros.

Por lo tanto, se decidió actualizar esta lista inicial de *semantic types*, ampliándolos a 16 incluyéndose, a los primeros, los siguientes *semantic types*: “*Intellectual Product*”, “*Mental Process*”, “*Mental or Behavioral Dysfunction*”, “*Pathologic Function*”, “*Congenital Abnormality*”. El resultado de esta configuración se ha descrito

4.3.1 Diseño de la evaluación

en la sección 3.2.4.1 sobre la configuración de la herramienta NLP seleccionada, MetaMap.

4.3.1 Diseño de la evaluación

La evaluación se llevó a cabo mediante un minucioso análisis manual sobre los textos de cada enfermedad . Para cada enfermedad obtenida de Wikipedia o PubMed, se comparó:

- 1) La lista de términos médicos extraídos manualmente de los textos que describen la enfermedad.
- 2) La lista de términos médicos extraídos por la herramienta MetaMap de los mismos textos.
- 3) El valor (*TRUE = valid o FALSE = invalid*) resultante del proceso de TVP al filtrar cada término médico detectado por MetaMap (Rodríguez González et al., 2015).
- 4) El valor de la relevancia diagnóstica de cada término.

La integración de estos cuatro elementos de comparación se encuentran contenido en hojas de validación. En la Tabla 3 se muestra un ejemplo del formato de la hoja de validación para el contenido textual de la enfermedad “*Acute decompensated heart failure*”.

Tabla 3. Hoja de validación para el artículo Wikipedia referente a la enfermedad "Acute decompensated heart failure". Información recuperada el 1 de febrero de 2018. Abreviaciones de columnas: WIKIPEDIA (W), METAMAP (M), TVP (T), RELEVANT (R).

	WIKIPEDIA TERMS	METAMAP TERMS	DISNET VALIDATION			
			W	M	T	RELEVANT
1	acute, myocardial, infarction	Acute myocardial infarction	YES	YES	YES	FPCONTEXT
2	illness	Illness (finding)	YES	YES	YES	FPREAL
3	hyperthyroidism	Hyperthyroidism	YES	YES	YES	FPCONTEXT
4	anemia	Anemia	YES	YES	YES	FPCONTEXT
5	weightloss	Weight decreased	YES	YES	YES	YES
6	palpitations	Palpitations	YES	YES	YES	YES
7	nausea	Nausea	YES	YES	YES	YES
8	chest, pain	Chest pain NOS	YES	YES	YES	YES
9	exertional, dyspnoea	Dyspnea on exertion	YES	YES	YES	YES
10	pneumonia	Pneumonia	YES	YES	YES	FPCONTEXT
11	high, blood, pressure	Hypertensive disease	YES	YES	YES	FPCONTEXT
12	weakness	Weakness	YES	YES	YES	YES
13	pain	Pain	YES	YES	YES	FPREAL
14	heart, failure	Heart failure	YES	YES	YES	FPCONTEXT
15	paroxysmal, nocturnal, dyspnoea	Paroxysmal nocturnal dyspnea	YES	YES	YES	YES
16	orthopnoea	Orthopnea	YES	YES	YES	YES
17	difficulty, breathing	Dyspnea	YES	YES	YES	YES

4.3.1 Diseño de la evaluación

	WIKIPEDIA TERMS	METAMAP TERMS	DISNET VALIDATION			
			W	M	T	RELEVANT
18	heart, attack	Myocardial infarction, NOS	YES	YES	YES	FPCONTEXT
19	abnormal, heart, rhythms	Cardiac arrhythmia	YES	YES	YES	FPCONTEXT
20	bloating	Abdominal bloating	YES	YES	YES	YES
21	chest, pressure	Pressure in chest	YES	YES	YES	YES
22	low, urine, output	Oliguria	YES	YES	YES	YES
23	fatigue	Fatigue	YES	YES	YES	YES
24	jugular, venous, distension	Jugular venous engorgement	YES	YES	YES	YES
25	atrial, fibrillation	Electrocardiographic atrial fibrillation	YES	YES	NO	NO
26	left, ventricular, failure	Left-sided heart failure	YES	YES	NO	NO
27	sign, signs	Physical finding	YES	YES	NO	NO
28	excess, fluid	Fluid overload	YES	YES	NO	NO
29	chronic, heart, failure	Chronic heart failure	YES	YES	NO	NO
30	pressure	Pressure (finding)	YES	YES	NO	NO
31	acute, heart, failure	Acute heart failure	YES	YES	NO	NO
32	myocardial, infarction	Electrocardiogram: myocardial infarction (finding)	YES	YES	NO	NO
33	decompensation	Decompensation	YES	YES	NO	NO
34	gasping	Gasping for breath	YES	YES	NO	NO
35	symptom, symptoms	Symptom	YES	YES	NO	NO
36	confusion	Confusion	YES	YES	YES	YES
37	fluid, retention	Body fluid retention	YES	YES	YES	FPCONTEXT
38	memory, impairment	Memory impairment	YES	YES	YES	YES
39	sensitive	Hypersensitivity	YES	YES	NO	NO
40	anxiety	Anxiety	YES	YES	YES	YES
	loss of appetite		YES	NO	NO	FN
	waking up at night to urinate		YES	NO	NO	FN
	cerebral symptoms		YES	NO	NO	FN

Como se muestra en la tabla, han sido detectados 40 términos médicos relevantes (numerados del 1 al 40), encontrándose al pie de la tabla tres conceptos médicos no numerados no detectados por el proceso de validación de MetaMap sino mediante la validación manual.

Se observa también una columna adicional denominada *RELEVANT*, que sintetiza el resultado de dicha validación, clasificando cada término como relevante o no. Los posibles valores de esta columna se definen como:

1. **RELEVANT = YES.** Si (WIKIPEDIA = YES) & (METAMAP = YES) & (TVP = (YES o NO)). Los términos clasificados de esta forma son considerados conceptos médicos válidos y relevantes para el diagnóstico de una enfermedad.

Obtienen dicho resultado conceptos como: “*Abdominal pain*”, “*nauseas*”; “*weakness*”, *etc.*

2. **RELEVANT = NO.** Si (WIKIPEDIA = YES) & (METAMAP = YES) & (TVP = NO), es decir, se considera un concepto médico no específico, y por lo tanto demasiado genérico como para ser considerado útil en el diagnóstico de una enfermedad.

4.3.1 Diseño de la evaluación

Obtienen dicha clasificación términos tales como “*sign, signs*”; “*symptom, symptoms*”; “*pressure*”, etc.

3. RELEVANT = **FPREAL**. Si (WIKIPEDIA = YES) & (METAMAP = YES) & (TVP = YES). A Pesar de que ha sido detectado por MetaMap, por considerarlo un término médico, y evaluado por TVP, se trata de términos no relevantes para el presente trabajo de tesis por no ser un concepto concreto que sirva para llevar a cabo el diagnóstico de la enfermedad que lo recoge.

Conceptos como “*illness*” o “*pain*” son catalogados como FPREAL.

Por ejemplo, en un extracto del 1 de febrero de 2018 de la enfermedad “*Acute decompensated heart failure*” en Wikipedia:

“*Other cardiac symptoms of heart failure include chest pain/pressure and palpitations...*”.

En el ejemplo anterior, MetaMap ha detectado los términos “*Chest pain*” y “*Pain*” de forma individualizada, descartándose el término unitario “*pain*” por ser demasiado genérico, teniéndose en cuenta, sin embargo, el término “*Chest pain*”.

4. RELEVANT = **FPCONTEXT**. Si (WIKIPEDIA = YES) & (METAMAP = YES) & (TVP = YES). En este caso, y a diferencia del anterior, el término es válido y relevante como método de diagnóstico sin embargo, no es relevante de forma concreta para la enfermedad que lo contiene, siendo un término descontextualizado.

Por ejemplo, en otro extracto del 1 de febrero de 2018 relativo a la enfermedad “*Acute decompensated heart failure*” de Wikipedia, de la expresión

“*Other well recognized precipitating factors include anemia and hyperthyroidism...*”.

MetaMap ha detectado los términos “*Anemia*” e “*Hyperthyroidism*” como términos médicos y, verdaderamente lo son, sin embargo se descartan como términos médicos relevantes dentro de la enfermedad que los recoge, puesto que en esta concreta enfermedad son tenidos como factores de riesgo de dicha enfermedad, pero no como elementos de diagnóstico.

5. RELEVANT = **FN**. Si (WIKIPEDIA = YES) & (METAMAP = NO) & (TVP = NO). Estos términos no fueron detectados por MetaMap y, por lo tanto, no fueron incluidos en la lista de elementos relevantes,

4.3.1 Diseño de la evaluación

siendo detectados de forma manual e incluidos en el pie de tabla de cada una de las enfermedades validadas.

Tanto los catalogados como FPREAL como FPCONTEXT deben ser descartados, en el primero de los casos porque el término no es en absoluto un elemento de diagnóstico, y en el segundo de los casos porque, si bien es cierto se trata de un término médico y que serviría como elemento diagnóstico, no lo es en la enfermedad concreta que lo refiere.

Con el resultado obtenido mediante este proceso de validación (el valor de la columna *RELEVANT*), se han calculado: el número de verdaderos positivos (**TP**, por sus siglas en inglés), falsos positivos (**FP**, por sus siglas en inglés), verdaderos negativos (**TN**, por sus siglas en inglés) y falsos negativos (**FN**, por sus siglas en inglés) para calcular la precisión (*precision*), el índice de recuperación (*recall*) y los valores de la puntuación F1 (*F1 score*) como métricas para medir el rendimiento del sistema de extracción y generación de conocimiento desarrollado en esta tesis. Los valores medios de estos parámetros se representan en la Figura 24.

Ahora bien, los **TP** son todos los términos con (WIKIPEDIA = YES) & (METAMAP = YES) & (TVP = YES) & (RELEVANT = YES). Como se ha explicado anteriormente, los **FP** están compuestos por dos partes, siendo la FP total la suma de FP_REAL + FP_CONTEXT:

- i. **FP_REAL** = (WIKIPEDIA = YES) & (METAMAP = YES) & (TVP = YES) & (RELEVANT = FPREAL).
- ii. **FP_CONTEXT** = (WIKIPEDIA = YES) & (METAMAP = YES) & (TVP = YES) & (RELEVANT = FPCONTEXT).

Los **FN** también se componen de dos partes, es decir, FN_METAMAP + FN_TVP.

- i. **FN_METAMAP** = (WIKIPEDIA = YES) & (METAMAP = NO) & (TVP = NO) & (RELEVANT = FN). Estos son términos que MetaMap no ha encontrado.
- ii. **FN_TVP** = (WIKIPEDIA = YES) & (METAMAP = YES) & (TVP = NO) & (RELEVANT = YES). Estos son términos que el proceso TVP ha evaluado como falsos, pero que son relevantes.

Finalmente, el **TN** mide el proceso de TVP (WIKIPEDIA = YES) & (METAMAP = YES) & (TVP = NO) & (RELEVANT = NO). En la

Tabla 4 se muestran los valores obtenidos para Wikipedia y PubMed.

4.3.1 Diseño de la evaluación

Figura 24. Comparativa de las métricas de validación tanto en Wikipedia como en PubMed.

Tabla 4. Valores totales de la *snapshot* del 1 de febrero de 2018 de Wikipedia y de la *snapshot* del 3 de abril de 2018 de PubMed.

PARAMETROS	WIKIPEDIA	PUBMED
TP	(31,11%) 2.075	(31,20%) 724
FP	(11,41%) 761	(17,54%) 407
FPREAL	279	107
FPCONTEXT	482	300
TN	(35,78%) 2.386	(32,84%) 762
FN	(21,68%) 1.446	(18,40%) 427
FN_METAMAP	709	201
FN_TVP	737	226
TOTAL	(100%) 6.668	(100%) 2.320
PRECISIÓN	0.731	0.640

Los resultados detallados de cada enfermedad están disponibles en línea, para Wikipedia en el siguiente enlace¹⁵³ y para PubMed en el siguiente enlace¹⁵⁴, donde se incluyen: la lista de términos extraídos manualmente de los textos relevantes de los artículos, la concordancia con la lista de términos proporcionada por MetaMap, el resultado del proceso de TVP para cada término y el valor de la relevancia del término para una enfermedad.

Los resultados indican que el proceso de NLP (MetaMap + TVP) tiene una precisión de 0,731 (intervalo de confianza de [0,710 – 0,753], calculado a través de

¹⁵³ https://github.com/disnet-project/thesis/tree/master/wikipedia_validation_sheets

¹⁵⁴ https://github.com/disnet-project/thesis/tree/master/pubmed_validation_sheets

4.3.1 Diseño de la evaluación

un intervalo de puntuación de Wilson (Wilson, 1927) con corrección de continuidad y un nivel de confianza del 99%) para Wikipedia y de 0,640 (intervalo de confianza de: [0,606 – 0,680]) para PubMed (Figura 24), siendo el proceso de NLP suficientemente fiable.

Los resultados de los cálculos de estos parámetros para cada enfermedad pueden verse en línea en el siguiente enlace¹⁵⁵ para Wikipedia y para cada resumen en PubMed en el siguiente enlace¹⁵⁶.

Con respecto a los resultados obtenidos para los **FP**, se debe a la amplitud de los parámetros de configuración de MetaMap para la extracción de términos utilizada, extendiendo la lista de *semantic types* de los 11 iniciales a 16 con el fin de abarcar el mayor número de términos relevantes y no dejar fuera un número elevado de ellos, prefiriendo ignorar los no procedentes que no abarcar todo el contenido importante para el desarrollo del presente trabajo de tesis –ver sección 3.2.4.1–.

También es interesante observar la gran diferencia en las tasas de **FP** entre Wikipedia (11,41%) y PubMed (17,54%) debido, principalmente, a la diferencia de concreción de los artículos en las distintas fuentes de información; así, como ya se ha explicado en numerosas ocasiones, mientras los artículos de Wikipedia hacen referencia, cada uno de ellos, a una enfermedad en concreto, haciendo referencia a otras enfermedades solo de forma relacional, haciendo referencia excepcionalmente a diagnósticos diferenciales, de forma que apenas incluyen términos médicos irrelevantes, en el caso de PubMed sus artículos no son tan concretos, pues presentan casos reales que entrecruzan información de los signos y síntomas que presenta un paciente concreto que padece la enfermedad a la que se refiere, con las características médicas anteriores o concomitantes del propio paciente

En consecuencia, el mismo artículo de PubMed incluye síntomas de muchas enfermedades diferentes que, si bien son verdaderos términos médicos y por lo tanto son reconocidos por MetaMap, no son relevantes para la enfermedad que se está analizando.

En el caso de **TN**, también hay que tener en cuenta que la mayoría de los términos extraídos por MetaMap como relevantes han sido depurados por TVP, que se ha encargado de determinar qué términos son relevantes y cuáles no, de manera que la gran mayoría de los términos extraídos por MetaMap que no son relevantes para la enfermedad han sido clasificados de esta forma por el proceso de TVP (35,78% para Wikipedia y 32,84% para PubMed).

¹⁵⁵ https://github.com/disnet-project/thesis/blob/master/wikipedia_individual_validation_results.csv

¹⁵⁶ https://github.com/disnet-project/thesis/blob/master/pubmed_individual_validation_results.csv

4.3.1 Diseño de la evaluación

Además, se ha observado que la mayoría de los términos TN tanto en Wikipedia como en PubMed son constantes, tales como: “*indicated*”, “*syndrome*”, “*disease*”, “*illness*”, “*infected*”, “*sing*”, “*symptoms*”, “*used to*”, etc.

Por último, los **FN** son aquellos términos que son relevantes para la enfermedad en cuestión, pero que no han sido detectados por MetaMap; nótese que estos han sido extraídos manualmente para el proceso de validación. La gran mayoría de FN están formados por expresiones complejas del lenguaje, por lo que su detección es un reto para cualquier herramienta de NLP. Además se observa que la diferencia en la proporción de falsos negativos entre Wikipedia (21,68%) y PubMed (18,40%) es del 3,28%. Esta diferencia se debe principalmente a las formas de expresión utilizadas en ambas fuentes, siendo Wikipedia más discursiva, en contraposición al estilo científico de PubMed.

En síntesis, se concluye que existe una clara relación entre el rendimiento del sistema y la naturaleza de la fuente de datos subyacente. Concretamente, mientras que PubMed es una fuente exclusivamente médica, creada, escrita y editada por especialistas en la materia, Wikipedia es una fuente de información pública, escrita por cualquiera que tenga acceso a la web, de modo que los artículos en ella contenidos pueden ser escritos por estudiantes de medicina o sólo por usuarios con algún conocimiento en la materia, cuyas expresiones no pueden ser comparadas con la de los especialistas que redactan artículos médicos en PubMed.

También, teniendo en cuenta que la herramienta utilizada para la extracción de términos médicos (MetaMap) es una herramienta médica, no es de extrañar que muestre una mayor capacidad de reconocimiento de términos médicos, frente a términos más coloquiales formados por frases más complejas; así, hay términos como “*Swollen lymph glands under the jaw*” o “*sensation of swelling in the area of the larynx*” que MetaMap no puede reconocer.

Es cierto que los porcentajes de validación no parecen muy elevados, pero hay que tener en cuenta los siguientes hechos, en primer lugar, que no existe ningún otro sistema que extraiga y genere información fenotípica con un enfoque como el propuesto en este documento de tesis y, en segundo lugar, que el objetivo del documento no es clínico, sino puramente de investigación, y por lo tanto permite poner todo el conocimiento generado al alcance de otros investigadores y de la comunidad científica en general.

Como conclusión se puede decir que la información médica contenida en DIS-NET puede ser utilizada por cualquier tipo de persona, especialistas en medicina o legos en ella, siendo todos ellos responsables del uso que den a dichos datos.

En todo caso y a pesar de la compleja naturaleza de los textos objeto de validación, los porcentajes reflejan un buen trabajo de investigación.

4.4 Análisis descriptivo del conjunto de datos médicos

En esta sección se describe el análisis descriptivo del contenido del conjunto de datos médicos de DISNET (DMDS, por sus siglas en inglés). El análisis se centra principalmente en el contenido obtenido de Wikipedia, la fuente principal de estudio en esta tesis y de la que se tiene mayor información histórica: desde el 1 de febrero de 2018 al 1 de marzo de 2020.

El objetivo de este análisis descriptivo es estudiar si la evolución de la información contenida en los artículos médicos de Wikipedia es de utilidad en un contexto de investigación.

El estudio está enfocado en el análisis de los elementos de conocimiento médico que se extrajeron de los artículos sobre enfermedades de Wikipedia y los que se generaron a partir de ellos, los DKE (*Diagnostic knowledge elements*), que se utilizan para guiar un proceso de diagnóstico, apoyar la creación de sistemas de diagnóstico o analizar las similitudes entre las enfermedades.

Los DKE son considerados como conocimiento diagnóstico, es decir, cualquier elemento relacionado con una enfermedad que permita a los profesionales de salud, dentro del proceso de diagnóstico, determinar o descartar posibles enfermedades. Estrictamente, éstas incluyen a las manifestaciones fenotípicas, esto es, los hallazgos, signos y síntomas. Sin embargo, otros elementos como los procedimientos de diagnóstico, las pruebas y resultados de laboratorio, o dicho de otra forma, análisis de sangre, recuento celular, prueba de función hepática, entre otros, también se consideran conocimientos diagnósticos, ya que permiten guiar el proceso de diagnóstico.

En la Tabla 5 se reportan las características obtenidas de la *snapshot* del 1 de marzo de 2020 que son clave para este estudio propuesto; así mismo en la Tabla 9 se ofrece información similar para los demás puntos temporales de conocimiento minado.

Es importante recordar que para la buena comprensión de esta sección es necesario conocer sobre la estructura del conjunto de datos DISNET, los elementos que la componen y como han sido obtenidos. Todos los DKE encontrados en los textos son conceptos provenientes del vocabulario UMLS. Y para ello se debe remitir a la sección 3.1.1 y 3.2 del Capítulo 3.

Como se observa en la Tabla 5, existe una diferencia significativa entre el número de artículos clasificados en DBpedia (Auer et al., 2007) como enfermedades (15.133) y el número de los artículos de enfermedades que finalmente contenían DKE evaluados positivamente por el proceso TVP (Rodríguez González et al., 2015) y que, por lo tanto, son considerados como enfermedades (5.565 - 36,77%).

4.4 Análisis descriptivo del conjunto de datos médicos

Tabla 5. Valores de los parámetros clave que conforman la *snapshot* del 1 de marzo de 2020.

KEY PROPERTIES	COUNT
Articles categorized as disease in DBpedia (DDLE process)	15.133
Articles that contained diagnostic knowledge elements applying TVP evaluation	5.565
Diagnostic knowledge elements (DKE) found (no duplicates)	2.053
UMLS semantic types found (no duplicates)	16
Disease codes found (no duplicates)	18.566
External vocabularies found (no duplicates)	47
Number of texts (no duplicates)	32.535
Number of links found in the texts	204.934

Este fenómeno tiene dos explicaciones: **i)** los artículos no contenían DKE evaluados positivamente por el proceso TVP; o **ii)** los artículos están catalogados incorrectamente, es decir, que en DBpedia fueron catalogados con la clase “Disease” pero el contenido del artículo Wikipedia no contenía información sobre una enfermedad.

Se considera que un artículo catalogado en DBpedia como enfermedad es “genuino” sólo si tiene al menos, un código de un sistema externo de vocabulario/clasificación como, ICD, OMIM, entre otros. Los artículos que no devuelven ningún códigos han sido clasificados como artículos no relacionados con una enfermedad, lo que implica que en DBpedia se han catalogado incorrectamente. Más información sobre los artículos que han sido recuperados como enfermedades en DBpedia está disponible en línea¹⁵⁷.

Ahora en el contexto de los DKE, en la Tabla 6 se presentan las seis enfermedades con mayor y menos cantidad de DKE obtenidos entre sus textos. En primer lugar, se observa la enfermedad “*Kawasaki disease*” con 100 conceptos médicos. El documento Wikipedia que corresponde a esta enfermedad cuenta una cantidad media de contenido textual referente al diagnóstico, ya que la longitud en caracteres de su contenido textual es de 11.343, en contraste, con la longitud más grande de 282.438, que es perteneciente a la enfermedad “*POEMS syndrome*”, donde se obtuvieron 46 DKE, y la más corta pertenece a la enfermedad “*Digestive system neoplasm*” con 33 de longitud en donde se encontró un solo DKE.

En promedio el contenido textual es de 11.460,66 caracteres y el promedio de DKE es 11.48. Todo esto sugiere que sugiere que la longitud de los textos no tiene una relación directa con respecto al número de DKE. Esta situación se aborda con mayor profundidad en la sección 4.6.

Dentro de la misma línea de DKE, en la Tabla 7 se presentan los seis conceptos médicos cuya frecuencia de asociación a enfermedades es mayor y menor. En la cabeza de la lista se encuentra el concepto “*Pain (sosal)*” con una frecuencia de

¹⁵⁷ https://github.com/disnet-project/thesis/blob/master/dbpedia_wikipedia_url_articles_and_codes.tsv

4.4 Análisis descriptivo del conjunto de datos médicos

2.322 apariciones. Opuestamente está el concepto “*Gaenslen's test (sosy)*” que solo aparecen en una sola enfermedad, de hecho en la *snapshot* del 1 de marzo de 2020 hay 224 conceptos médicos que están asociados solo una vez.

Tabla 6. Artículos Wikipedia de enfermedad con mayor y menor cantidad DKE en enfermedades (*snapshot*: 1 de marzo de 2020).

METRIC	DISEASE	URL HTTP://EN.WIKIPE- DIA.ORG	UNIQUE DKE COUNT
Disease with more DKE (1st)	Kawasaki disease	/wiki/Kawasaki_disease	100
Disease with more DKE (2nd)	Heart failure	/wiki/Heart_failure	81
Disease with more DKE (3rd)	Cerebral palsy	/wiki/Cerebral_palsy	80
Disease with more DKE (4th)	Anorexia nervosa	/wiki/Anorexia_nervosa	79
Disease with more DKE (5th)	Hypoglycemia	/wiki/Hypoglycemia	78
Disease with more DKE (6th)	Crohn's disease	/wiki/Crohn's_disease	78
Disease with less DKE (1st)	Vocal loading	/wiki/Vocal_loading	1
Disease with less DKE (2nd)	Transient tachypnea of the newborn	/wiki/Transient_tachypnea_of_the_newborn	1
Disease with less DKE (3rd)	X-linked intellectual disability	/wiki/X-linked_intellectual_disability	1
Disease with less DKE (4th)	Priming (steam locomotive)	/wiki/Priming_(steam_locomotive)	1
Disease with less DKE (5th)	Skin tag	/wiki/Skin_tag	1
Disease with less DKE (6th)	Renal dysplasia-limb defects syndrome	/wiki/Renal_dysplasia-limb_defects_syndrome	1

Tabla 7. Frecuencia mayor y menor de DKE (total de apariciones del concepto; total de artículos de enfermedad en los que aparece el concepto) (*snapshot*: 1 de marzo de 2020).

METRIC	CONCEPT (SEMANTIC TYPE)	APPEARANCE COUNT
Term with more appearances (1st)	Pain (sosy)	2.322
Term with more appearances (2nd)	Neoplasms (neop)	1.956
Term with more appearances (3rd)	Lesion (fndg)	1.620
Term with more appearances (4th)	Magnetic resonance imaging (diap)	1.238
Term with more appearances (5th)	Convulsions (sosy)	1.101
Term with more appearances (6th)	Malnutrition (dsyn)	1.084
Term with less appearances (1st)	Mixed type deafness (fndg)	1
Term with less appearances (2nd)	Nocturnal myoclonus (sosy)	1
Term with less appearances (3rd)	Nevus spilus (dsyn)	1
Term with less appearances (4th)	Bleeding from ear (lbtr)	1
Term with less appearances (5th)	Albumin/Globulin ratio (dsyn)	1
Term with less appearances (6th)	Gaenslen's test (sosy)	1

Otra métrica es el número de *semantic types* UMLS encontrados. Estos tipos semánticos corresponden a una agrupación de DKE en familias homogéneas y

4.4 Análisis descriptivo del conjunto de datos médicos

relacionadas conceptualmente como ya se ha mencionado anteriormente en la sección 3.2.4. La distribución de los DKE agrupados por su *semantic types* está disponible en la Tabla 8 y la distribución de los DKE agrupados por secciones relevantes está disponible en la Figura 25.

Como se observa en la Tabla 8, los *semantic types* con mayor número de DKE son: dsyn, sosy, fndg, diap, patf y mobd, algo que era de esperar dado el tipo de información que se ha recuperado. Por otro lado, están los *semantic types*, comd y lbtr como los tipos semánticos menos frecuentes.

Tabla 8. Distribución de los DKE agrupados según los *semantic types* (*snapshot*: 1 de marzo de 2020).

SEMANTIC TYPE	DESCRIPTION	COUNT
acab	Acquired Abnormality	16
anab	Anatomical Abnormality	17
cgab	Congenital Abnormality	48
clna	Clinical Attribute	7
comd	Cell or Molecular Dysfunction	2
diap	Diagnostic Procedure	146
dsyn	Disease or Syndrome	623
fndg	Finding	399
lbpr	Laboratory Procedure	47
lbtr	Laboratory or Test Result	7
menp	Mental Process	15
mobd	Mental or Behavioral Dysfunction	138
neop	Neoplastic Process	25
patf	Pathologic Function	139
sosy	Sign or Symptom	445

Por otro lado, en la Figura 25 se observa claramente que hay dos secciones de las que se han obtenido la misma cantidad de DKE: “*Signs and symptoms*” y “*Diagnosis*”. Las siguientes sección en la que se obtuvieron mayor cantidad de DKE han sido: “*Causes*”, “*infobox*”, “*Cause*”, “*Symptoms*” y “*Presentation*” ordenados de mayor a menor. Esto es importante para trabajos futuros ya que permite enfocar los esfuerzos de extracción, por ejemplo, en las primeras *snapshots* en la sección “*infobox*” solo se encontraba información referente a identificadores de enfermedades, pero se cambió ese enfoque para incluir información de diversos aspectos de la enfermedad como: síntomas, durabilidad, fármacos, tratamientos, entre otros aspectos. Esto ha permitido que “*infobox*” ha sido más relevante para la extracción de conocimiento diagnóstico.

4.4.1 Evolución de la información

Figura 25. Distribución de los DKE agrupados por las secciones relevantes (snapshot: 1 de marzo de 2020).

4.4.1 Evolución de la información

En cuanto a la evolución de la información, se han calculado varias métricas para las diferentes *snapshots*: **a)** el número de artículos recuperados por DBpedia como enfermedades (DBpDis), **b)** el número de artículos de Wikipedia que contienen DKE aplicando la evaluación del proceso TVP (WRDArt), **c)** el número de DKE encontrados por MetaMap (sin aplicar la evaluación del proceso TVP pero eliminando los duplicados) (WRawDF), **d)** el número de textos (WTxt), **e)** el número de tipos semánticos encontrados (de nuevo: antes de aplicar el proceso TVP)

4.4.1 Evolución de la información

(WST), **f**) el número de códigos de enfermedades externos encontrados en Wikipedia (WExCd), **g**) el número de fuentes biomédicas externas encontradas en Wikipedia (WExtSrc) y **h**) el número de enlaces encontrados en los textos (WLink). La Tabla 9 muestra la evolución de estas métricas en las diferentes *snapshots*.

Tabla 9. Evolución de la información contenida en cada una de las *snapshots* en Wikipedia, del 1 de febrero de 2018 al 1 de marzo de 2020.

SNAPSHOT	DBpDis	WRDArt	WRawDF	WTxt	WST	WExCd	WExtSrc	WLink
2018-02-01	8.161	3.623	1.932	15	15.046	45	20.017	115.431
2018-02-15	8.161	3.629	1.934	15	15.030	45	20.058	115.747
2018-03-01	8.161	3.634	1.937	15	15.008	45	20.144	116.246
2018-03-15	8.161	3.642	1.942	15	15.023	45	20.309	117.091
2018-04-01	9.857	3.838	1.949	15	15.179	46	21.190	121.969
2018-04-15	9.858	3.843	1.950	15	15.197	46	21.242	122.465
2018-05-01	9.858	3.857	1.954	15	15.226	46	21.333	123.054
2018-05-15	9.858	3.867	1.955	15	15.264	46	21.424	123.402
2018-06-01	9.858	4.043	1.964	15	15.720	45	22.178	127.033
2018-06-15	9.858	4.056	1.968	16	15.724	45	22.378	128.046
2018-07-01	9.858	4.058	1.970	15	15.738	46	22.463	128.468
2018-07-15	9.858	4.065	1.972	15	15.742	45	22.525	128.800
2018-08-01	9.858	4.079	1.973	15	15.768	45	22.596	129.366
2018-08-15	9.858	4.083	1.974	15	15.774	45	22.650	129.751
2018-09-01	9.858	4.086	1.974	15	15.885	45	22.682	129.952
2018-09-15	9.858	4.087	1.973	15	15.895	45	22.664	129.968
2018-10-01	9.858	4.089	1.973	15	15.911	45	22.702	130.124
2018-10-15	9.858	4.093	1.973	15	15.928	45	22.763	130.505
2018-11-01	9.858	4.097	1.970	15	15.945	45	22.820	130.971
2018-11-15	9.858	4.102	1.972	15	15.959	45	22.889	131.535
2018-12-01	9.858	4.107	1.976	15	15.977	45	22.988	132.019
2018-12-15	9.858	4.115	1.978	15	15.994	45	23.110	132.746
2019-01-01	9.858	4.124	1.981	15	16.031	45	23.216	133.401
2019-01-15	9.858	4.135	1.981	15	16.055	45	23.274	133.742
2019-02-01	11.084	4.688	2.009	15	16.663	45	26.388	150.860
2019-02-15	11.084	4.702	2.009	15	16.664	45	26.604	152.174
2019-03-01	11.121	4.736	2.011	15	16.734	45	26.970	154.369
2019-03-15	11.125	4.742	2.013	15	16.764	45	27.010	154.779
2019-04-01	10.160	4.322	1.998	15	16.371	45	24.983	143.943
2019-04-15	11.169	4.765	2.016	15	16.827	45	27.333	156.795
2019-05-01	11.172	4.770	2.018	15	16.831	45	27.395	157.303
2019-05-15	11.177	4.771	2.019	15	17.641	44	27.427	157.666
2019-06-01	11.187	4.781	2.019	15	17.665	44	27.624	158.988
2019-06-15	11.228	4.783	2.019	15	17.678	45	27.674	159.309
2019-07-01	11.278	4.803	2.021	15	17.694	45	27.823	160.150
2019-07-15	11.285	4.801	2.021	15	17.689	45	27.877	160.515
2019-08-01	11.290	4.812	2.021	15	17.713	45	27.995	161.270
2019-08-15	11.625	5.103	2.041	15	18.331	45	29.743	187.457
2019-09-01	11.625	5.107	2.041	15	18.333	46	29.884	188.366
2019-09-15	11.659	5.129	2.043	15	18.346	46	29.986	188.874
2019-10-01	11.829	5.218	2.045	15	18.421	46	30.331	190.725
2019-10-15	11.835	5.226	2.045	15	18.432	46	30.369	190.976
2019-11-01	11.843	5.246	2.041	15	18.467	47	30.537	191.916
2019-11-15	11.846	5.251	2.042	15	18.478	47	30.667	192.836

4.4.1 Evolución de la información

SNAPSHOT	DBpDis	WRDArt	WRawDF	WTxt	WST	WExCd	WExtSrc	WLink
2019-12-01	11.863	5.127	2.040	15	18.218	47	30.071	189.742
2019-12-15	15.133	5.472	2.048	15	18.306	47	31.637	198.251
2020-01-01	15.133	5.401	2.037	15	18.379	47	30.423	193.941
2020-01-15	15.133	5.551	2.053	15	18.546	47	32.299	202.846
2020-02-01	15.133	5.558	2.055	15	18.557	47	32.460	204.066
2020-02-15	15.133	5.507	2.053	15	18.430	47	32.063	202.019
2020-03-01	15.133	5.565	2.053	15	18.566	47	32.535	204.934

El tiempo juega un papel importante porque permite observar la evolución del conocimiento médico almacenado en Wikipedia. Como se puede observar en la Tabla 9, hay una progresión en la mayoría de los elementos. Sólo los *semantic types* (WST) son los que han permanecido sin cambios significativos, como es de esperar, ya que este número sólo cambia al encontrar nuevos DKE de otros diferentes tipos o con que un concepto médico sea pertenezca a dos *semantic types*.

También se observa que todas las métricas (excepto la WExCd, que presenta un comportamiento irregular) están aumentando monótonamente, lo que sugiere que las modificaciones en los artículos de enfermedad de Wikipedia dan lugar a la inclusión de nueva información que es capturada por el proceso de extracción y generación de conocimiento diagnóstico descrito en detalle en la sección 3.2.

Desde un punto de vista global, todas estas métricas descritas hasta el momento apoyan la viabilidad de usar Wikipedia como fuente de información médica.

Por lo tanto, se ha obtenido información de 7.810 artículos que están catalogados como enfermedades y que contienen información de diagnóstico, y de los cuales 5.563 eran enfermedades relevantes, es decir, con al menos un término médico (después de aplicar el proceso TVP). Si bien es digno de mención, este número debe compararse con los obtenidos en otras fuentes médicas: $\approx 4,500$ para MeSH¹⁵⁸; $\approx 7,000$ para OMIM¹⁵⁹; y $\approx 10,500$ para DisGeNET¹⁶⁰.

Si bien a primera vista DisGeNET incluye información sobre el doble de las enfermedades de Wikipedia, hay que tener en cuenta dos aspectos adicionales. En primer lugar, Wikipedia es un sistema vivo, actualizado frecuentemente por sus usuarios, y cuya base de conocimientos aumenta constantemente, como se muestra en el Tabla 9. En segundo lugar, los diferentes sitios web dependen de diferentes vocabularios: una misma enfermedad puede entonces clasificarse de manera diferente, dividirse en diferentes subtipos, fusionarse, y así sucesivamente. Por lo tanto, el elevado número encontrado para DisGeNET no implica necesariamente

¹⁵⁸ <https://midas.ctb.upm.es/gitlab/disnet/paperWikipedia/blob/master/getDiseasesMesh.sparql>

¹⁵⁹ <https://www.omim.org/statistics/entry>

¹⁶⁰ <https://midas.ctb.upm.es/gitlab/disnet/paperWikipedia/blob/master/getDiseasesDisGeNET.sparql>

4.4.1 Evolución de la información

un mayor cuerpo de información, al menos en lo que se refiere a la información DKE.

En la Figura 26 se muestra un gráfico que compara la primera *snapshot* generada con la última. Con ella se confirma que durante estos poco más de dos años de extraer y generar conocimiento diagnóstico se ha encontrado un aumento de la información contenida en los artículos catalogados como enfermedades y, sobre todo, que cada nueva contribución tiene suficiente contenido textual médico para permitir, identificar un número creciente de DKE.

Figura 26. Comparación entre la primera *snapshot* (1 de febrero de 2018) y la última (1 de marzo de 2020).

A pesar de toda la información descrita anteriormente, hay algunas cuestiones que se deben mencionar. En primer lugar, se han encontrado muchos artículos que fueron catalogados como enfermedades en DBpedia, pero que fueron descartados después de aplicar el método de filtrado (no hay códigos externos en DBpedia ni en Wikipedia).

Algunos de los artículos están relacionados con el ámbito médico (por ejemplo: “*famous outbreaks*”, “*health crisis*”, etc.), pero verdaderamente no son “Diseases”. Algunos ejemplos de esos artículos son: “*1924 Los Angeles pneumonic plague outbreak*”, “*1852–60 cholera pandemic*”; “*1863–75 cholera pandemic*” y “*2013 Swansea measles epidemic*”.

4.4.1 Evolución de la información

Sin embargo, muchos otros ni siquiera están relacionados con la información médica de ninguna manera (por ejemplo: “2008 Western Australian gas crisis”, “2010 in film”, “2003 Wimbledon Championships – Women's Singles”).

Otro inconveniente es la detección de artículos de Wikipedia potencialmente relevantes que no han dado ningún resultado en el proceso de extracción y generación de conocimiento. La razón radica en el hecho de que se han limitado las secciones de texto a “*Signs and symptoms*”, “*Causes*”, “*Diagnosis*” and “*Presentation*”, pero éstas están vacías en algunos artículos (por ejemplo: “*Hereditary sensory and autonomic neuropathy*”, “*17q21.31 microdeletion syndrome*”, “*2,4 Dienoyl-CoA reductase deficiency*”).

Por otra parte, se han encontrado artículos sobre enfermedades que no han sido estructurados utilizando las secciones relevantes que se han usado para recuperar contenido textual médico (algunos de estos artículos son: “*Bleb (medicine)*”, “*Oligodactyly*” o “*Meteoropathy*”).

También hay que poner atención en aquellos artículos referentes a enfermedades que ya no están catalogadas como tales, por ejemplo, el artículo sobre la “*Female Hysteria*”.

En la Tabla 10 se muestra una comparación detallada entre el número de enfermedades en todas las *snapshots* disponibles, comparando una *snapshot* con su predecesora y se pueden encontrar más detalles en línea¹⁶¹. En la primera columna se compara la lista de enfermedades y se indica si el número de enfermedades ha aumentado (+) o disminuido (-). La segunda columna indica el porcentaje de superposición entre las enfermedades comunes de ambas *snapshots* (cuántas enfermedades hay en ambas *snapshots*). Y la última columna muestra el porcentaje de variación de las enfermedades, es decir, el porcentaje de enfermedades que han tenido un cambio en algunos de sus elementos de diagnóstico.

Tabla 10. Comparación entre todas las *snapshots* de enfermedades en Wikipedia. Considerando las *snapshots* que comprenden del 1 de febrero de 2018 al 1 de marzo de 2020.

SNAPSHOT	NO. DISEASE	% OVERLAPPING	% DISEASE VARIATION
2018-02-15	+6	99,83	0,97
2018-03-01	+8	99,78	1,65
2018-03-15	+15	99,59	1,96
2018-04-01	+202	94,74	2,09
2018-04-15	+10	99,74	1,43
2018-05-01	+21	99,46	1,56
2018-05-15	+12	99,69	5,91
2018-06-01	+179	95,57	5,23
2018-06-15	+17	99,58	6,34
2018-07-01	+8	99,80	7,26
2018-07-15	+9	99,78	1,36

¹⁶¹ https://github.com/disnet-project/thesis/blob/master/variation_between_wikipedia_sn.tsv

4.4.1 Evolución de la información

SNAPSHOT	NO. DISEASE	% OVERLAPPING	% DISEASE VARIATION
2018-08-01	+19	99,53	1,43
2018-08-15	+6	99,85	1,47
2018-09-01	+7	99,83	0,76
2018-09-15	+5	99,88	1,1
2018-10-01	+4	99,90	1,3
2018-10-15	+9	99,78	0,83
2018-11-01	+12	99,71	1,44
2018-11-15	+9	99,78	1,2
2018-12-01	+10	99,76	1,78
2018-12-15	+9	99,78	1,46
2019-01-01	+12	99,71	1,75
2019-01-15	+11	99,73	1,12
2019-02-01	+558	88,10	1,65
2019-02-15	+24	99,49	2,33
2019-03-01	+41	99,13	2,83
2019-03-15	+12	99,75	1,54
2019-04-01	+13	99,70	1,81
2019-04-15	+447	90,62	1,37
2019-05-01	+10	99,79	1,58
2019-05-15	+5	99,90	1,34
2019-06-01	+16	99,67	1,62
2019-06-15	+5	99,90	0,94
2019-07-01	+21	99,56	1,82
2019-07-15	+3	99,94	1,15
2019-08-01	+13	99,73	1,5
2019-08-15	+317	93,79	20,43
2019-09-01	+6	99,88	1,67
2019-09-15	+22	99,57	0,76
2019-10-01	+91	98,26	1,7
2019-10-15	+8	99,85	0,75
2019-11-01	+27	99,49	1,63
2019-11-15	+12	99,77	1,18
2019-12-01	+16	99,69	1,76
2019-12-15	+420	92,32	2,04
2020-01-01	+89	98,35	29,31
2020-01-15	+154	97,23	28,87
2020-02-01	+10	99,82	1,3
2020-02-15	+5	99,91	1,13
2020-03-01	+60	98,92	1,65

También en la Tabla 10 se observa que el porcentaje de variación de la enfermedad (columna % DISEASE VARIATION) tiene valores muy altos en las *snapshots* del 15 de mayo de 2018, 1 de junio de 2018, 15 de junio de 2018, 1 de julio de 2018, 15 de agosto de 2019, 1 de enero de 2020, 15 de enero de 2020. Estos tipos de aumentos se deben a que en esos intervalos de tiempo varios artículos de enfermedad fueron editados para añadir contenido fenotípico. Los picos más altos se han registrado durante el primer mes del año 2020.

También se observan altos valores en el número de enfermedades (columna NO. DISEASE) en las instantáneas del 1 de abril de 2018, 1 de junio de 2018 y 1

4.4.1 Evolución de la información

de febrero de 2019, 15 de abril de 2019, 15 de agosto de 2019, 1 de octubre de 2019, 15 de diciembre de 2019, 1 de enero de 2020, 15 de enero de 2020, 1 de marzo de 2020.

Esta situación se puede deber a varios factores asociados a la DBpedia, ya que es el medio por el cual se identifican nuevos artículos de enfermedades; nuevos artículos que no fueron catalogados como enfermedades; inconsistencia en los resultados del servicio del servidor HTTP/SPARQL.

También factores asociados a la Wikipedia, como que los artículos se editen y añadan tanto contenido fenotípico como identificadores externos de enfermedades, haciendo que un artículo empiece a pasar desapercibido. Por último, el aumento del número de enfermedades en la instantánea del 1 de febrero de 2019 se debe a la inclusión de los resultados de DBpedia-Live¹⁶².

Sin embargo, incluso teniendo en cuenta estos inconvenientes, los resultados son prometedores: ya que aun así, si del total de artículos obtenidos de Wikipedia (15.133) se descartan aquellos artículos que no son considerados como enfermedad, se han obtenido conceptos médicos en 5.565 (36,77%) documentos, lo que representa un número realmente alto.

Además, las cifras que en esta sección se han presentado y analizado también proporcionan un buen punto de referencia para uno de los objetivos específicos establecidos: obtener la mayor cantidad de información diagnóstica posible de los artículos de enfermedad de Wikipedia.

Adicionalmente, en la Tabla 11 se muestra información sobre los cambios en la lista de enfermedades en Mayo Clinic. El objetivo de la Tabla 11 es hacer una comparación con la información obtenida de Wikipedia. Y es que si se observa la lista de enfermedades Mayo Clinic en comparación con Wikipedia tiene cambios muy tenues con muy poca variación. Esto se debe a que Mayo Clinic tiene el control total sobre las ediciones de sus artículos y depende completamente de que su equipo de médicos y investigadores realicen modificaciones.

Tabla 11. Comparación entre todas las snapshots de enfermedades en Mayo Clinic. Considerando las *snapshots* que comprenden del 15 de agosto de 2018 al 1 de marzo de 2020.

SNAPSHOT	NO. DISEASE	% OVERLAPPING	% DISEASE VARIATION
2018-09-01	+0	100,00	1,11
2018-09-15	+1	99,91	0,55
2018-10-01	+2	99,82	1,01
2018-10-15	+0	100,00	0,74
2018-11-01	+2	99,82	0,55
2018-11-15	+3	99,72	0,28
2018-12-01	+3	99,73	0,64

¹⁶² <https://wiki.dbpedia.org/online-access/DBpediaLive>

4.5 Análisis descriptivo de identificadores de enfermedades

SNAPSHOT	NO. DISEASE	% OVERLAPPING	% DISEASE VARIATION
2018-12-15	+14	98,73	1,1
2019-01-01	+13	98,83	0,73
2019-01-15	+1	99,91	0,45
2019-02-01	+2	99,82	0,9
2019-02-15	+1	99,91	0,63
2019-03-01	+1	99,91	0,54
2019-03-15	+0	100,00	0,54
2019-04-01	+1	99,91	0,18
2019-04-15	+0	100,00	0,09
2019-05-01	+1	99,91	0,45
2019-05-15	+0	100,00	0,71
2019-06-01	+0	100,00	1,7
2019-06-15	+0	100,00	0,27
2019-07-01	+0	100,00	0,71
2019-07-15	+1	99,89	0,11
2019-08-01	+237	78,84	0,57
2019-08-15	+2	99,82	0,63
2019-09-01	+1	99,91	0,71
2019-09-15	+1	99,91	0,8
2019-10-01	+0	100,00	0
2019-10-15	+0	100,00	0
2019-11-01	+2	99,82	1,78
2019-11-15	+0	100,00	0,44
2019-12-01	+2	99,82	0,53
2019-12-15	+3	99,73	0,89
2020-01-01	+3	99,73	0,27
2020-01-15	+0	100,00	0,53
2020-02-01	+1	99,91	1,24
2020-02-15	+1	99,91	1,06
2020-03-01	+4	99,65	0,89

4.5 Análisis descriptivo de identificadores de enfermedades

En esta sección se presenta un análisis descriptivo referente solo a los códigos de enfermedades. Los códigos permiten identificar o mapear desordenes humanos en DISNET (DISNET id) con enfermedades de otros vocabularios o recursos biomédicos como OMIM, DO, UMLS, entre otros. Estos identificadores también permiten caracterizar las enfermedades ya que son elementos que sirven para clasificar, jerarquizar y organizar los desordenes que se encuentran en el conjunto de datos DISNET. Al igual que en la sección anterior 4.4 en esta sección se hace uso de todos los datos, es decir, los que comprenden del 1 de febrero de 2018 al 1 de marzo de 2020.

La capacidad de relacionar enfermedades de distintas bases de datos para complementar y mejorar el conocimiento de las enfermedades es uno de los objetivos específicos del proyecto DISNET y para ello es necesario conocer la distribución de los datos.

4.5 Análisis descriptivo de identificadores de enfermedades

Como se ha presentado en la Tabla 2 de la sección 4.2 en el conjunto de datos médicos (DMDS) el número total de los códigos identificadores de enfermedades es de 35.203 tomando en cuenta la información de todas las *snapshots* de todas las fuentes. En la Figura 27 se muestra un diagrama con la distribución de los vocabularios que mayor cantidad de códigos de enfermedades tienen en el conjunto de datos.

Figura 27. Número de códigos de enfermedades recuperados según cada vocabulario biomédico registrado en el conjunto de datos (considerando todas las *snapshots* de todas las fuentes).

También en la Figura 27 se aprecia que de SNOMED CT poco más de 8.000 códigos de enfermedades que suponen el 22,49% del total. La segunda y tercera posición la ocupan los vocabularios: MeSH e ICD-10 con 13,62% y 12,61% respectivamente. Después se encuentran cinco vocabularios que tienen casi la misma cantidad de códigos de enfermedades: ICD-9-CM, eMedicine, UMLS, DiseasesDB y OMIM. Finalmente, el resto de los vocabularios como el AO Foundation aportan menos de 30 códigos.

En la presente versión del sistema los métodos de obtención de códigos de enfermedades solo son dos –ver sección 3.1–: i) el proceso de extracción de información de Wikipedia que usa *Web scraping*, y ii) el proceso de extracción de información de PubMed que usa los descriptores MeSH referentes a enfermedades humanas y su relación con la *Disease Ontology* (DO).

4.5 Análisis descriptivo de identificadores de enfermedades

Figura 28. Códigos de enfermedades recuperados mediante el proceso de extracción de conocimiento de PubMed.

En la Figura 28 se muestran todos los vocabularios que se han obtenido códigos al extraer conocimiento de PubMed y la Figura 29 se muestran los 15 vocabularios biomédicos con mayor cantidad de códigos obtenidos por el proceso de extracción de conocimiento de Wikipedia, siendo MeSH el que más tiene. Además, se observa en la Figura 28 que de la extracción de PubMed se han recuperado el 100% de los códigos SNOMED CT y UMLS. Finalmente, en la Figura 30 se muestra una gráfica que compara los códigos de los vocabularios en PubMed y Wikipedia.

De Wikipedia solo se han extraído identificadores de enfermedades pero por minar PubMed y mediante la *Disease Ontology*, además se han extraído 7.683 identificadores de sinónimos de enfermedades¹⁶³, de los cuales 5.467 tienen códigos de enfermedades asociados¹⁶⁴. También por minar PubMed se han obtenido 22.653 identificadores de descriptores MeSH que están asociados a los artículos PubMed relevantes de esta tesis¹⁶⁵.

¹⁶³ https://github.com/disnet-project/thesis/blob/master/all_synonyms.tsv

¹⁶⁴ https://github.com/disnet-project/thesis/blob/master/synonym_codes.tsv

¹⁶⁵ https://github.com/disnet-project/thesis/blob/master/mesh_term_in_pubmed_papers.tsv.zip

4.5 Análisis descriptivo de identificadores de enfermedades

Figura 29. Códigos de enfermedades recuperadas mediante el proceso de extracción de conocimiento de Wikipedia.

Figura 30. Comparación entre el número de códigos de los vocabularios obtenidos por la extracción de conocimiento de PubMed y Wikipedia.

4.5 Análisis descriptivo de identificadores de enfermedades

Ahora bien, es importante conocer la distribución de las enfermedades que cuentan con identificadores, con el fin de que en tareas futuras se realicen esfuerzos que permitan completar la información existente.

En la Figura 31 en el cuadro a) se muestra que dentro del DMDS se encuentran 5.868 enfermedades que al menos tienen un identificador de las 7.160 enfermedades totales que existen. Es decir, al rededor 1.548 enfermedades no tienen asociado ningún identificador. En el grafico se han contemplado todas las fuentes y todas sus *snapshots*.

Figura 31. a) Se muestran todas las enfermedades que tienen al menos un identificador (5.868) dentro del conjunto de datos en contraste con el total de enfermedades, considerando en ambos todas las *snapshots*. b) Enfermedades con al menos un identificador en Wikipedia y en PubMed, considerando todas las *snapshots* para ambas fuentes.

En el cuadro b) de la Figura 31 se presenta una comparación entre las enfermedades que tienen al menos un identificador asociado tanto para Wikipedia como para PubMed. Se han considerado todas las *snapshots* de ambas fuentes. Se observa que hay 1.024 enfermedades que se pueden mapear entre las dos fuentes mediante 2.771 identificadores asociados.

Ahora en la Figura 32 se presenta una comparativa entre la primera y última *snapshot* de Wikipedia. Se observa que en poco más de dos años de recuperar conocimiento médico de esta fuente se ha mantenido información de 3.187 enfermedades, de 128 ya no se recuperó información, y hubo 1.330 nuevas enfermedades agregadas.

4.5 Análisis descriptivo de identificadores de enfermedades

Figura 32. Comparación del número de enfermedades con al menos un identificador asociado en la primera y última *snapshot* de Wikipedia.

Las razones por las que ya no se tiene información de 128 enfermedades pueden ser: por causa de una redirección¹⁶⁶, porque se eliminó el artículo o por una reestructuración de contenido, en la cual se eliminaron códigos de enfermedades o secciones relevantes para esta tesis, acción que hace que no se consideren.

Con la Tabla 12 se observan las enfermedades con mayor y menor número de códigos identificadores tanto de PubMed como de Wikipedia. Las seis primeras enfermedades con mayor número de códigos en PubMed superan a las de Wikipedia ya que tienen entre 45 y 29 códigos asociados, en comparación con la de Wikipedia que solo tienen entre 29 y 20.

Tabla 12. Enfermedades con mayor y menor número de códigos en PubMed y Wikipedia.

SOURCE	METRIC	DISEASE	CODE COUNT
PubMed	Disease with more codes (1st)	Coronary artery disease	45
PubMed	Disease with more codes (2nd)	Sarcoma, Kaposi	42
PubMed	Disease with more codes (3rd)	Joint Diseases	36
PubMed	Disease with more codes (4th)	Arthritis, Infectious	34
PubMed	Disease with more codes (5th)	Tongue Neoplasms	33
PubMed	Disease with more codes (6th)	Myopia	29
PubMed	Disease with less codes (1st)	Erysipeloid	1
PubMed	Disease with less codes (2nd)	Bone resorption	1
PubMed	Disease with less codes (3rd)	Mucositis	1
PubMed	Disease with less codes (4th)	Ichthyosis	1
PubMed	Disease with less codes (5th)	Lupus nephritis	1
PubMed	Disease with less codes (6th)	Cytomegalovirus retinitis	1
Wikipedia	Disease with more codes (1st)	Limb-girdle muscular dystrophy	29
Wikipedia	Disease with more codes (2nd)	Kallmann syndrome	29
Wikipedia	Disease with more codes (3rd)	Tuberous sclerosis	24
Wikipedia	Disease with more codes (4th)	Craniosynostosis	21
Wikipedia	Disease with more codes (5th)	Varicose veins	21
Wikipedia	Disease with more codes (6th)	Methylmalonyl-CoA mutase deficiency	20
Wikipedia	Disease with less codes (1st)	Otic polyp	1
Wikipedia	Disease with less codes (2nd)	Cryptomenorrhea	1

¹⁶⁶ https://en.wikipedia.org/wiki/Wikipedia:How_to_make_a_redirect

4.5 Análisis descriptivo de identificadores de enfermedades

SOURCE	METRIC	DISEASE	CODE COUNT
Wikipedia	Disease with less codes (3rd)	Knee pain	1
Wikipedia	Disease with less codes (4th)	Dentin hypersensitivity	1
Wikipedia	Disease with less codes (5th)	Ring chromosome 20 syndrome	1
Wikipedia	Disease with less codes (6th)	Anorectal varices	1

Ahora bien, en la Figura 33 se muestra las enfermedades que tienen códigos de los vocabularios que más identificadores aportan. Claramente se observa que hay 1.147 enfermedades que tienen códigos de las tres distintas fuentes (MeSH, SNOMED CT, ICD-10).

Figura 33. Enfermedades según los vocabularios con mayor número de códigos, considerando todas las *snapshots*.

A pesar de que Mayo Clinic no tiene códigos de enfermedades asociados, realizando un mapeo entre enfermedades según sus nombres se obtiene como resultado la Figura 34, que muestra que solo se han podido mapear 612 enfermedades.

Finalmente, se analizan los mapeos que se pueden realizar al relacionar los códigos de enfermedades contenidos en el conjunto de datos de DISNET (DISNET id) entre los códigos disponibles en la fuente de información DisGeNET¹⁶⁷. El mejor proceso para establecer correspondencias entre enfermedades es mediante el uso de los identificadores CUI (*Concept Unique Identifiers*) de UMLS, ya que las enfermedades en DisGeNET son identificadas con esos tipos de códigos.

En DisGeNET, hay 22.201 CUI diferentes y 65.643 códigos de 9 vocabularios diferentes. El número de mapeos es 97.502. La distribución de códigos y mapeos

¹⁶⁷ https://www.disgenet.org/static/disgenet_ap1/files/downloads/disease_mappings.tsv.gz

4.5 Análisis descriptivo de identificadores de enfermedades

a lo largo de los diferentes vocabularios se muestra en la Figura 35 y la Figura 36 respectivamente.

Figura 34. Mapeo por nombre entre las enfermedades en Wikipedia y Mayo Clinic.

En la Figura 37, se ve la distribución de la frecuencia que un identificador DISNET mapea un número de CUIs (“DISNET – DisGeNET”) y de las veces que un CUI mapea un número de identificadores DISNET (“DisGeNET – DISNET”). Se producen ambas situaciones y se observa que, la frecuencia con que un CUI mapea varios identificadores DISNET no es alta, ya que la correspondencia es que un CUI mapea a un identificador DISNET con una media es de 1,27.

Figura 35. Frecuencia de mapeo de códigos entre DISNET y DisGeNET y provenientes de cada vocabulario.

4.5 Análisis descriptivo de identificadores de enfermedades

Figura 36. Frecuencia de los códigos procedentes de cada vocabulario en DisGeNET en su versión 6.

Figura 37. Distribuciones de (DISNET id – CUI) y (CUI - DISNET id) de acuerdo a los mapeos hechos con DisGeNET.

4.5 Análisis descriptivo de identificadores de enfermedades

También se observa que un CUI que mapea 11 identificadores DISNET. Dicho CUI es el C0023794, cuyo nombre en DisGeNET (concepto UMLS) es “*Lipoidosis*”.

En la Tabla 13 se muestran las correspondencias de las diferentes enfermedades en DISNET con el CUI C0023794. Además, se muestran los vocabularios y los códigos con los que se ha hecho el mapeo, en este sentido, es de hacer notar que solo hay un código que viene del vocabulario MeSH. Esto se debe a que una de las enfermedades DISNET mapeadas tiene asociado dos códigos: el i) 272.7¹⁶⁸ de ICD-9-CM (común con las demás enfermedades) y el ii) D008064¹⁶⁹ de MeSH.

Para esta enfermedad, el ICD-9-CM se mantiene en un nivel jerárquico más alto y no baja en la clasificación. Los sinónimos aproximados (ver sección *Approximate Synonyms* del código¹⁶⁸) que convierten ICD-9-CM en ICD-10-CM incluyen las diferentes enfermedades que se encuentran en DISNET, lo que indica que podrían ser clasificadas con mayor profundidad. Incluso, hay CUI que podrían ser más precisos que el general.

Tabla 13. Mapeo del CUI C0023794 en DisGeNet con identificadores DISNET.

DISNET ID	DISEASE	VOCABULARY	CODE	CONCEPT
DIS005105	Cerebrotendineous xanthomatosis	ICD-9-CM	272.7	Lipidoses
DIS003639	Fabry disease	ICD-9-CM	272.7	Lipidoses
DIS003696	Gaucher's disease	ICD-9-CM	272.7	Lipidoses
DIS005506	I-cell disease	ICD-9-CM	272.7	Lipidoses
DIS004083	Lipid storage disorder	ICD-9-CM	272.7	Lipidoses
DIS004083	Lipid storage disorder	MeSH	D008064	Lipoidosis
DIS002526	Lysosomal acid lipase deficiency	ICD-9-CM	272.7	Lipidoses
DIS007879	Mucopolipidosis	ICD-9-CM	272.7	Lipidoses
DIS010670	Niemann–Pick disease	ICD-9-CM	272.7	Lipidoses
DIS003462	Niemann–Pick disease, SMPD1-associated	ICD-9-CM	272.7	Lipidoses
DIS007567	Niemann–Pick disease, type C	ICD-9-CM	272.7	Lipidoses
DIS004849	Pseudo-Hurler polydystrophy	ICD-9-CM	272.7	Lipidoses

Asimismo, el identificador DISNET que tiene un número mayor (35) de CUIs asociados es DIS006978, cuyo nombre es “*Pseudocholinesterase deficiency*”. En la Tabla 14 se muestran los diferentes CUIs, conceptos, vocabularios, códigos y nombres relacionados con esta enfermedad DISNET. Este caso es el opuesto al anterior, teniendo DISNET una categorización más general que el UMLS. El verdadero CUI para el DISNET id DIS006978 sería el C0268379, que está en la cima de la jerarquía.

¹⁶⁸ <http://www.icd9data.com/2015/Volume1/240-279/270-279/272/272.7.htm>

¹⁶⁹ <https://meshb.nlm.nih.gov/record/ui?ui=D008064>

4.5 Análisis descriptivo de identificadores de enfermedades

Tabla 14. Mapeos del identificador DISNET DIS006978 con los CUI de DisGeNet.

CUI	DisGeNet DISEASE (UMLS CONCEPT)	VOCABULARY	CODE	VOCABULARY NAME
C0268379	Pseudocholinesterase deficiency	MeSH	C537417	Pseudocholinesterase deficiency
C1283400	Butyrylcholinesterase deficiency	MeSH	C537417	Pseudocholinesterase deficiency
C1867468	Apnea, Postanesthetic	MeSH	C537417	Pseudocholinesterase deficiency
C1622434	Suxamethonium sensitivity	MeSH	C537417	Pseudocholinesterase deficiency
C1867467	Acholinesterasemia	OMIM	177400	PSEUDOCHOLINESTERASE DEFICIENCY
C4017559	BCHE, QUANTITATIVE H VARIANT PHENOTYPE	OMIM	177400	PSEUDOCHOLINESTERASE DEFICIENCY
C3889589	BCHE, K VARIANT PHENOTYPE	OMIM	177400	PSEUDOCHOLINESTERASE DEFICIENCY
C1867470	Hypocholinesterasemia, Fluoride-Resistant, Japanese Type	OMIM	177400	PSEUDOCHOLINESTERASE DEFICIENCY
C4017558	BCHE, QUANTITATIVE J VARIANT PHENOTYPE	OMIM	177400	PSEUDOCHOLINESTERASE DEFICIENCY
C1283400	Butyrylcholinesterase deficiency	OMIM	177400	PSEUDOCHOLINESTERASE DEFICIENCY
C3889605	BCHE CYNTHIANA PHENOTYPE	OMIM	177400	PSEUDOCHOLINESTERASE DEFICIENCY
C3889592	BCHE, H VARIANT PHENOTYPE	OMIM	177400	PSEUDOCHOLINESTERASE DEFICIENCY
C4017548	BCHE*FS117 PHENOTYPE	OMIM	177400	PSEUDOCHOLINESTERASE DEFICIENCY
C4017543	BCHE, DIBUCAINE-RESISTANT I PHENOTYPE	OMIM	177400	PSEUDOCHOLINESTERASE DEFICIENCY
C1867469	Butyrylcholinesterase Deficiency, Fluoride-Resistant, Japanese Type	OMIM	177400	PSEUDOCHOLINESTERASE DEFICIENCY
C4017553	CHE*390V PHENOTYPE	OMIM	177400	PSEUDOCHOLINESTERASE DEFICIENCY
C4017549	BCHE, FLUORIDE-RESISTANT I PHENOTYPE	OMIM	177400	PSEUDOCHOLINESTERASE DEFICIENCY
C4017554	BCHE*390V PHENOTYPE	OMIM	177400	PSEUDOCHOLINESTERASE DEFICIENCY
C4017546	BCHE ANN ARBOR PHENOTYPE	OMIM	177400	PSEUDOCHOLINESTERASE DEFICIENCY
C1867468	Apnea, Postanesthetic	OMIM	177400	PSEUDOCHOLINESTERASE DEFICIENCY
C4017544	CHE*70G PHENOTYPE	OMIM	177400	PSEUDOCHOLINESTERASE DEFICIENCY
C4017552	BCHE, FLUORIDE-RESISTANT II PHENOTYPE	OMIM	177400	PSEUDOCHOLINESTERASE DEFICIENCY
C4017545	BCHE*70G PHENOTYPE	OMIM	177400	PSEUDOCHOLINESTERASE DEFICIENCY
C0268379	Pseudocholinesterase deficiency	OMIM	177400	PSEUDOCHOLINESTERASE DEFICIENCY
C4017557	BCHE*539T PHENOTYPE	OMIM	177400	PSEUDOCHOLINESTERASE DEFICIENCY
C4017560	BCHE*FS126 PHENOTYPE	OMIM	177400	PSEUDOCHOLINESTERASE DEFICIENCY
C4017550	CHE*243M PHENOTYPE	OMIM	177400	PSEUDOCHOLINESTERASE DEFICIENCY

4.6 Análisis de la evolución histórica del contenido de enfermedades Wikipedia

CUI	DisGeNet DISEASE (UMLS CONCEPT)	VOCABULARY	CODE	VOCABULARY NAME
C3889591	BCHE, J VARIANT PHENOTYPE	OMIM	177400	PSEUDOCHOLINESTERASE DEFICIENCY
C3889586	BCHE, FLUORIDE 1 PHENOTYPE	OMIM	177400	PSEUDOCHOLINESTERASE DEFICIENCY
C4017555	BCHE, QUANTITATIVE K POLYMORPHISM PHENOTYPE	OMIM	177400	PSEUDOCHOLINESTERASE DEFICIENCY
C4017551	BCHE*243M PHENOTYPE	OMIM	177400	PSEUDOCHOLINESTERASE DEFICIENCY
C4017556	CHE*539T PHENOTYPE	OMIM	177400	PSEUDOCHOLINESTERASE DEFICIENCY
C4016356	APNEA, POSTANESTHETIC, DUE TO BCHE, ATYPICAL-1	OMIM	177400	PSEUDOCHOLINESTERASE DEFICIENCY
C4017547	CHE*FS117 PHENOTYPE	OMIM	177400	PSEUDOCHOLINESTERASE DEFICIENCY
C1622434	Suxamethonium sensitivity	OMIM	177400	PSEUDOCHOLINESTERASE DEFICIENCY
C3889588	BCHE, FLUORIDE 2 PHENOTYPE	OMIM	177400	PSEUDOCHOLINESTERASE DEFICIENCY
C3889604	BCHE, SILENT 1 PHENOTYPE	OMIM	177400	PSEUDOCHOLINESTERASE DEFICIENCY
C3889606	BCHE JOHANNESBURG PHENOTYPE	OMIM	177400	PSEUDOCHOLINESTERASE DEFICIENCY
C3889596	BCHE NEWFOUNDLAND PHENOTYPE	OMIM	177400	PSEUDOCHOLINESTERASE DEFICIENCY

4.6 Análisis de la evolución histórica del contenido de enfermedades Wikipedia

En esta sección se describe el análisis que se ha llevado a cabo con el fin de observar los cambios producidos en el contenido médico de los artículos de enfermedades contenidos en Wikipedia –en su edición en inglés– en una franja de tiempo determinada de 18 meses que comprende información del 1 de febrero de 2018 al 1 de julio de 2019. En este sentido, es importante aclarar que aunque en el conjunto de datos médicos se tiene información hasta el 1 de marzo de 2020 (51 *snapshots* de Wikipedia), en esta sección el análisis evolutivo se realiza con información de un menor número de *snapshots* ya que la cantidad de datos acumulados fue suficiente para arrojar resultados de valor, los cuales se describirán en los siguientes párrafos de esta sección.

Observar la evolución de la información, como va cambiando a lo largo del tiempo el contenido de los artículos médicos de Wikipedia permite caracterizar mejor las enfermedades ya que se muestran cuáles son los elementos clave que agregan valor al contenido y a cuales se les debe poner más atención para mejorar.

A pesar de que las fuentes tenidas en cuentas para el desarrollo de este trabajo de tesis han sido, no solo Wikipedia, sino también PubMed y Mayo Clinic, el

4.6.1 Diseño del análisis

análisis que se describe en la presente sección solo se ha llevado a cabo en relación con Wikipedia, en primer lugar, porque es Wikipedia la fuente de información diferencial en este trabajo de tesis, anteponiéndola sobre las demás, en segundo lugar, y en relación con lo anterior, porque es la fuente de información de la que más contenido se ha extraído, pues el análisis de los artículos de contenido médico contenidos en Wikipedia es el objetivo principal de la presente sección. En tercer lugar, Wikipedia representa una fuente de información que evoluciona en cada artículo por si misma mientras que los artículos de PubMed son elementos independientes que no pueden ser vinculados para analizar una evolución temporal.

4.6.1 Diseño del análisis

Como se ha mencionado con anterioridad, para realizar esta caracterización particular que pretende analizar la evolución del contenido de los artículos de Wikipedia sobre enfermedades, se ha centrado esta sección en la medición y observación de tres variables: **i)** el número de caracteres que contiene el artículo, **ii)** el número de referencias que contiene el artículo y **iii)** el número de DKE (*Diagnostic Knowledge Elements*). Además, el presente análisis se centra tanto en las variables individuales como en las conjuntas.

El número de caracteres es una medida que permite reconocer fácilmente cuando un artículo ha sido actualizado o modificado. Nótese que las modificaciones pueden implicar del añadir más texto con nueva información, la eliminación de texto en modo de depuración para lograr una versión más sintética y, finalmente, cambios en la forma de expresar las ideas.

Mientras el número de caracteres se centra en la cantidad, las referencias y los DKE están más relacionadas con la calidad; las primeras (las referencias) cuantifican el nivel de influencia en la literatura científica, las segundas (DKE) la densidad de los conceptos médicos. Obsérvese que la relación entre estas tres métricas está lejos de ser trivial, ya que un incremento en una de ellas no implica necesariamente un incremento en las otras dos.

Adicionalmente, todos los resultados obtenidos por este proceso de análisis han sido sintetizados y publicados en línea, para facilitar futuros estudios. Estos son: **a)** la variación en el número de caracteres por enfermedad¹⁷⁰, **b)** la variación en el número de referencias por enfermedad¹⁷¹ y **c)** la variación en el número de términos médicos por enfermedad¹⁷². Cada conjunto de datos está estructurado como una matriz, en la que cada fila es una enfermedad y cada columna es el número

¹⁷⁰ https://github.com/disnet-project/thesis/blob/master/wikipedia_diseases_characters.csv

¹⁷¹ https://github.com/disnet-project/thesis/blob/master/wikipedia_diseases_references.csv

¹⁷² https://github.com/disnet-project/thesis/blob/master/wikipedia_diseases_dkes.csv

4.6.1 Diseño del análisis

de caracteres, referencias o DKE (según el conjunto de datos –ver sección 3.2.2–) en cada una de las 36 DISNET *snapshots*.

Particularidades a considerar:

- Es importante señalar que el presente análisis se realiza sobre la información médica recuperada de los artículos de enfermedades de Wikipedia de forma quincenal durante un periodo de 18 meses, mediante el uso de la propia API de Wikipedia –ver sección 3.3–. Los 18 meses se comprenden del 1 de febrero de 2018 al 1 de julio de 2019.
- También en este punto, vale la pena discutir dos situaciones específicas que no pueden ser procesadas correctamente por el sistema desarrollado. En primer lugar, Wikipedia ofrece a sus usuarios tanto la posibilidad de borrar un artículo completo, con todas sus revisiones, como la posibilidad de borrar selectivamente una o más revisiones de un artículo. De esta forma, el proceso de extracción desarrollado no es capaz de recuperar ningún tipo de información histórica, pues esta ha sido eliminada.

Un ejemplo de esto se puede encontrar en el artículo de Wikipedia “*Pancreatic mucinous cystic neoplasm*”. Si bien la información se recuperó correctamente desde el 1 de febrero de 2018 hasta el 1 de noviembre de 2018. A partir de noviembre de 2018 el artículo dejó de estar disponible debido al motivo “*Unambiguous copyright infringement*”¹⁷³.

- En segundo lugar, no todas las enfermedades estuvieron disponibles como objeto de análisis a lo largo de esos 18 meses como consecuencia de la naturaleza evolutiva tanto del sistema DISNET como de Wikipedia, ya que ambas han ido ampliando sus respectivas listas de enfermedades: la primera gracias a mejoras internas, y la segunda como consecuencia de adiciones. En consecuencia, para algunas de éstas enfermedades puede faltar información de *snapshots*.

Para resolver estos inconvenientes y tener una cantidad homogénea de información a analizar, sólo se han considerado los artículos de aquellas enfermedades de las que se han extraído información en cada una de las 36 *snapshots* (dos cada mes durante 18 meses).

Así, tras la aplicación de este filtro la lista de enfermedades pasó de contener 4.917 a 3.508, que han sido las tenidas en cuentas para el desarrollo del análisis propuesto en esta sección.

¹⁷³ https://en.wikipedia.org/wiki/Special:Log?type=delete&user=&page=Pancreatic_mucinous_cystic_neoplasm&wpdate=&tagfilter=&subtype=

4.6.2 Desarrollo del análisis

En la Figura 38, en el cuadro de la izquierda, se representa en primer lugar un histograma del número de *snapshots* en función del porcentaje de cambio en el número de caracteres (línea negra), términos (línea azul) y referencias (línea roja), con respecto a la *snapshot* anterior de la misma enfermedad. En otras palabras, dadas dos *snapshots* consecutivas en el tiempo t y $t + 1$, se calculó el cambio relativo como: $v_m(t + 1) = 100 \cdot (m_{t+1} - m_t)/m_t$, siendo la métrica que se está analizando.

Se dibujan dos ideas interesantes. En primer lugar, las tres distribuciones se caracterizan por una larga cola, ya que, por ejemplo, es posible encontrar un número significativo de *snapshots* que tienen modificaciones en sus métricas del 40% o más.

Esto es de esperar en el caso de las referencias y los términos, ya que es posible que se añada contenido a las *snapshots* que inicialmente sólo incluían textos muy cortos, y por lo tanto también sólo un puñado de términos.

No obstante, se observa el mismo comportamiento en cuanto al número de caracteres se refiere; por ello, las modificaciones pueden ir con frecuencia más allá de simples correcciones gramaticales, y pueden incluir profundas mejoras en el texto.

Figura 38. Evolución de las *snapshots* de Wikipedia. (Izquierda) Distribución de la probabilidad de la variación del número de caracteres (línea negra), DKE (línea roja) y referencias (línea azul) en dos *snapshots* consecutivas, expresada en porcentaje. (Derecha) Distribución de probabilidad acumulada de las mismas tres métricas.

La gran prevalencia de pequeñas correcciones gramaticales también queda excluida por la alta correlación entre los cambios en el número de caracteres y las frases ($r^2 = 0,914$); por lo tanto, los grandes cambios textuales no son consecuencia de las reformulaciones.

4.6.2 Desarrollo del análisis

La dinámica observada de las tres métricas se confirma además en el cuadro derecho de la Figura 38, que representa la distribución de probabilidad acumulativa de la fracción de *snapshots* que han presenciado un cambio determinado en las tres métricas. En términos matemáticos, esta probabilidad se define como $P_m(v) = P(m \geq v)$, donde m es la métrica considerada, y v la variación que se está evaluando. Se aprecia también que, especialmente por el número de referencias, la mayoría de los cambios suponen una variación superior al 50%.

Volviendo al cuadro izquierdo de la Figura 38, mientras que la mayoría de las modificaciones implican un aumento del contenido, es interesante observar que algunas de ellas muestran una reducción sustancial. Concretamente, se observa una reducción en el 32,17% de las *snapshots* en el caso del número de caracteres, 25,16% en el caso de los términos y 15,60% en el caso de las referencias. Esto puede deberse a divisiones, por ejemplo, cuando una misma enfermedad se divide en dos conceptos separados; pero también, a una importante reestructuración en la que se elimina la información redundante –obsérvese que la primera hipótesis no puede probarse, ya que tales divisiones no se registran en la historia del artículo–.

Por último, cabe señalar que, si bien el cuadro izquierdo de la Figura 38 parece sugerir que la evolución del número de términos y referencias está fuertemente correlacionada, en realidad no es así (*Pearson's* $r^2 = 0,262$). La similitud en la forma de ambas distribuciones se debe al número relativamente bajo de términos y referencias, y por lo tanto sobre el hecho de que algunos porcentajes de variación están prohibidos – para ilustrarlo, se requiere que al menos 100 términos tengan una variación entre -1% y +1%.

En la Figura 39 se estudia más a fondo la evolución del contenido de las *snapshots*, mostrando cómo el número de las tres métricas ha evolucionado a través del tiempo. Concretamente, el eje X representa el tiempo (cada punto es una *snapshot* disponible), mientras que el eje Y indica el incremento (o decremento) en las tres métricas con respecto a la *snapshot* inicial.

A favor de una mejor comprensión, y debido al elevado número de enfermedades, éstas se representan como distribuciones de probabilidad, en las que cada línea indica un decil (véase la leyenda en la esquina inferior derecha). Además, en interés de dilucidar, no se representan aquellas enfermedades que han sufrido cambios en el número de caracteres superiores a 1.000. Como es de esperar, el tamaño de la información disponible tiende a aumentar con el tiempo, y sólo menos del 10% de las enfermedades reducen su texto.

Finalmente se analiza cómo se relacionan las tres diferentes métricas entre ellas. Para ello, la Figura 40 representa cuatro gráficos de densidad, relacionando el número de términos, caracteres, referencias y referencias PMID (PubMed ID) para un mismo artículo de Wikipedia (los colores oscuros indican densidades más

4.6.2 Desarrollo del análisis

altas, por lo tanto un mayor número de *snapshots*, y las líneas punteadas azules los mejores ajustes lineales). Se observa una modesta correlación en el primer caso (*Pearson's* $r^2 = 0,5685$), y una menos definida en el segundo (*Pearson's* $r^2 = 0,5093$).

Figura 39. Evolución del número de caracteres, DKE y referencias, a lo largo de todas las *snapshots*. Cada parte de color representa un decil de la distribución (véase la leyenda de la derecha). Todos los valores están normalizados de tal manera que el valor inicial es cero.

Por consiguiente, puede concluirse que, cuanto más largo es el texto, mayor es el número de términos médicos en él incluidos; por otra parte, es posible que los términos no siempre estén respaldados por las correspondientes referencias.

Esto puede indicar que algunos términos comunes, como “*fever*”, no necesitan estar respaldados por evidencias científicas para ser comúnmente aceptados por la comunidad. En el cuarto cuadro de la Figura 40 se describe, por último, la relación entre el número de referencias incluidas en cada artículo y *snapshots*, y el número de referencias con un código PMID; nótese que la presencia de dicho código indica que la referencia es un documento médico y, por lo tanto, se supone que su contenido es fiable. Más de la mitad de las referencias son de documentos médicos ($r^2 = 0,822$), lo que sugiere una buena calidad en los materiales de referencia externos.

Tras este análisis estadístico global, ahora se describe un estudio detallado de algunas enfermedades cuyas *snapshots* presentan comportamientos de interés.

4.6.2 Desarrollo del análisis

Figura 40. Relaciones entre las métricas consideradas. Los cuatro cuadros informan de los gráficos de densidad que representan las relaciones entre el número de términos, caracteres, referencias y referencias PMID para un mismo artículo de Wikipedia. Las sombras más oscuras indican un mayor número de snapshots. La línea punteada azul indica el mejor ajuste lineal, junto con la pendiente y r^2 correspondientes. La línea punteada roja del cuarto panel corresponde a la diagonal principal, es decir, un número de referencias del PMID igual al número de referencias.

En primer lugar, la Tabla 15 presenta las seis enfermedades con mayor y menor contenido, en términos del número de caracteres promediados en todas las *snapshots*. No es de extrañar que en la parte superior del listado se encuentren artículos como “*Drowning*”, “*Alzheimer's disease*” y “*HIV/AIDS*”, ya que son de gran interés tanto para los lectores como para los editores: atraen a un gran número de espectadores (más de 10^5 visualizaciones por mes) y de modificaciones.

Tabla 15. Artículos de enfermedad con el mayor y el menor promedio de caracteres.

MÉTRICA	ARTICULO DE ENFERMEDAD	CARACTERES MIN Y MAX	NO. DE CAMBIOS
1st highest character average	Drowning	161,160 - 167,190	21
2nd highest character average	HPV-positive oropharyngeal cancer	150,073 - 162,228	17
3rd highest character average	Alzheimer's disease	152,351 - 166,992	28
4th highest character average	Major depressive disorder	150,271 - 159,412	28
5th highest character average	HIV/AIDS	147,001 - 151,266	27
6th highest character average	Addiction	141,740 - 157,435	27
6th lowest character average	Aortoenteric fistula	1,017 - 1,218	2
5th lowest character average	Serositis	1,125 - 1,305	2
4th lowest character average	Presbylarynx	1,191 - 1,191	0
3rd lowest character average	Omental infarction	1,142 - 1,249	3
2nd lowest character average	Saber shin	1,172 - 1,193	2

4.6.2 Desarrollo del análisis

MÉTRICA	ARTICULO DE ENFERMEDAD	CARACTERES MIN Y MAX	NO. DE CAMBIOS
1st lowest character average	Katz syndrome	940 - 1,097	1

Se aprecia también que existe una fuerte correlación entre el tamaño de un artículo y el número de veces que éste ha sido editado; y que ambos valores están probablemente impulsados por el interés público por esa enfermedad.

Esto se confirma además en la Figura 41, en la que se informa de la distribución del número de modificaciones en función del tamaño del artículo (es decir, el número de caracteres). Cuando se realiza un ajuste cuadrático para la mediana del número de cambios en función de la mediana del tamaño del artículo, el coeficiente cuadrático es casi cero (concretamente, $1,030 \cdot 10^{-8}$), lo que indica que los cambios son atraídos linealmente por el tamaño.

Figura 41. Relación entre el tamaño de cada artículo y el número de veces que se ha modificado. Todos los artículos se han dividido en diez grupos, que corresponden a los diez deciles según el número de caracteres - la mediana de cada grupo se indica en el eje X. Cada cuadro representa la distribución del número de cambios para ese decil.

En la misma línea, la Figura 42 presenta dos paneles, con la evolución a través del tiempo del número de caracteres de las 6 enfermedades con mayor y menor cantidad respectivamente, como se indica en la Tabla 15.

Una vez más se puede apreciar que las enfermedades populares se editan con frecuencia, lo que conduce a una evolución (más o menos) suave. Por el contrario, los artículos de contenido limitado presentan cambios abruptos, seguidos de largos períodos de inactividad.

4.6.2 Desarrollo del análisis

Además del problema de la popularidad, esta dinámica irregular puede deberse también a la concentración de la mayoría de las modificaciones en unos pocos usuarios.

Figura 42. Comparación entre la evolución de los artículos sobre enfermedades en Wikipedia con una cantidad alta y baja de contenido textual. a),b).

4.6.2 Desarrollo del análisis

Concretamente, si se consideran únicamente los pequeños artículos de la Tabla 15, los usuarios registrados en Wikipedia con mayor número de contribuciones son el “Doc James”¹⁷⁴, médico de urgencias canadiense, que editó el artículo sobre el “*Katz syndrome*” el 13 de febrero de 2019, añadiendo una nueva sección y una imagen; y “Ozzie10aaaaa”¹⁷⁵, que añadió, por ejemplo, dos identificadores de enfermedades en el artículo sobre la “*Aortoenteric fistula*”¹⁷⁶ el 2 de junio de 2018.

El historial de ambos usuarios incluye un gran número de colaboraciones, y reconocimientos por parte del WikiProject Medicine¹⁷⁷ del que son miembros. Dicho proyecto fue iniciado en 2004 por el Dr. Jacob de Wolff¹⁷⁸, que tiene como objetivo gestionar y ayudar en la conservación de los artículos médicos de Wikipedia.

En la lista de colaboradores de los artículos más grandes también se encuentra el “Doc James”¹⁷⁴, junto con una lista mucho más numerosa de editores como “Keith D”¹⁷⁹, “TylerDurden8823”¹⁸⁰ o “Boghog”¹⁸¹.

Esta relación entre el tamaño de un artículo y el número correspondiente de autores colaboradores también se confirma en la Figura 43, en la que se indica la distribución del número de autores en función del tamaño de cada artículo. En el decil más bajo, es decir, para los artículos con menos de 6.813 caracteres, la mediana del número de autores es sólo 5. Como en el caso de la Figura 41, un ajuste

Figura 43. Relación entre el tamaño de cada artículo y el número de autores que han modificado ese artículo. Como se ha hecho en la Figura 41, los artículos se han dividido en diez deciles según el número de caracteres - la mediana de cada grupo se indica en el eje X. Cada recuadro representa la distribución del número de autores para ese decil.

¹⁷⁴ https://en.wikipedia.org/wiki/User:Doc_James

¹⁷⁵ <https://en.wikipedia.org/wiki/User:Ozzie10aaaa>

¹⁷⁶ https://en.wikipedia.org/w/index.php?title=Aortoenteric_fistula&oldid=844077228

¹⁷⁷ https://en.wikipedia.org/wiki/Wikipedia:WikiProject_Medicine

¹⁷⁸ <https://en.wikipedia.org/wiki/User:Jfdwolff>

¹⁷⁹ https://en.wikipedia.org/wiki/User:Keith_D

¹⁸⁰ <https://en.wikipedia.org/wiki/User:TylerDurden8823>

¹⁸¹ <https://en.wikipedia.org/wiki/User:Boghog>

4.6.2 Desarrollo del análisis

cuadrático presenta un coeficiente cuadrático insignificante ($3,702 \cdot 10^{-9}$), lo que sugiere una relación lineal entre las dos métricas.

Sin embargo, las tendencias descritas anteriormente no son universales, como puede apreciarse en la Figura 44, que representa la evolución temporal de las 12 enfermedades con mayor variabilidad (en términos de número de caracteres). Estos artículos tienen entre 5 y 17 cambios, con variaciones en el tamaño del contenido que van del 40% al 80%. Cinco de ellos tuvieron, en algún momento, más de 20.000 caracteres, por lo que pueden considerarse como artículos de gran tamaño y alta variabilidad. Sin embargo, una variabilidad tan grande no es la norma: en todo el conjunto de datos hay una variabilidad media del 10%; además, 3.322 artículos sobre enfermedades muestran una variabilidad menor que la media (frente a 186, para los que la variabilidad es mayor), lo que indica la presencia de una distribución asimétrica con una larga cola derecha.

Figura 44. Evolución a través del tiempo del número de caracteres de los 12 artículos de enfermedad con mayor variabilidad.

Como último punto, se analizan aquellos artículos que son más notables en cuanto al número de términos médicos que contienen. La Tabla 16 muestra el número de DKE para los seis artículos de enfermedades con mayor contenido textual, según las Tabla 15 y Tabla 17 muestra los seis artículos de enfermedades

4.7 Similitud en base a características fenotípicas

con el mayor número de DKE. Se aprecia que no se observa una fuerte correlación entre la longitud y el número de DKE, como también se ha destacado anteriormente en la Figura 40. Esto es fácil de explicar, ya que los artículos con el texto más largo pueden tener ese contenido distribuido en diferentes secciones, no todas ellas relevantes para el diagnóstico. Por otra parte, los artículos cortos, como los enumerados en la Tabla 17, pueden tener una parte sustancial de ellos dedicada al diagnóstico y, por lo tanto, pueden incluir un gran número de términos.

Tabla 16. Número de DKE de los artículos sobre enfermedades con mayor contenido textual (número de caracteres).

MÉTRICA	ARTICULO DE ENFERMEDAD	DKE MIN Y MAX
Disease article with highest character average (1st)	Drowning	16 – 17
Disease article with highest character average (2nd)	HPV-positive oropharyngeal cancer	10 – 11
Disease article with highest character average (3rd)	Alzheimer's disease	40 – 41
Disease article with highest character average (4th)	Major depressive disorder	53 – 54
Disease article with highest character average (5th)	HIV/AIDS	28 – 30
Disease article with highest character average (6th)	Addiction	10 – 11

Tabla 17. Artículos de enfermedades con el mayor número de DKE.

MÉTRICA	ARTICULO DE ENFERMEDAD	DKE MIN Y MAX
Disease article with highest character average (1st)	Drowning	16 – 17
Disease article with highest character average (2nd)	HPV-positive oropharyngeal cancer	10 – 11
Disease article with highest character average (3rd)	Alzheimer's disease	40 – 41
Disease article with highest character average (4th)	Major depressive disorder	53 – 54
Disease article with highest character average (5th)	HIV/AIDS	28 – 30
Disease article with highest character average (6th)	Addiction	10 – 11

4.7 Similitud en base a características fenotípicas

En esta sección se presenta un análisis que comprende los cálculos de similitudes entre enfermedades. Este análisis permite caracterizar las enfermedades del conjunto de datos médico de DISNET según el cálculo de diferentes coeficientes de similitud en base a sus características de diagnóstico.

En los trabajos relacionados se describieron algunos estudios de investigación que habían desarrollado sus propios cálculos de similitud, así como había y otros que usaban métodos ya establecidos como lo son: Coseno, Jaccard o Dice (Ljubesic et al., 2008; Chen et al., 2015) con buenos resultados. Un ejemplo sin duda es el trabajo de (Zhou et al., 2014b) el cual abrió en su momento un nuevo horizonte de conocimiento por medio de la creación de una red de enfermedades relacionadas en base a la similitud fenotípica según el coeficiente Coseno. Demostrando que algunas enfermedades con alto índice de similitud fenotípica comparten también características a un nivel biológico, así como otros hallazgos relevantes.

4.7 Similitud en base a características fenotípicas

En esta tesis se han calculado las similitudes según los métodos establecidos más comunes y referentes (Pesquita et al., 2008; Zhou et al., 2014c; Cheng et al., 2016; Peng et al., 2016): Coseno, Jaccard y Dice porque han sido usados en trabajos muy relevantes en el ámbito de la salud.

A continuación se muestran las fórmulas que calculan cada coeficiente de similitud:

- El coeficiente Coseno se calcula de la siguiente manera:

$$sim_{Cosine}(A, B) = \frac{|A \cap B|}{|A|^{\frac{1}{2}} + |B|^{\frac{1}{2}}} = \frac{\sum_{i=1}^N A_i B_i}{\sqrt{\sum_{i=1}^N A_i^2} \sqrt{\sum_{i=1}^N B_i^2}}$$

- El coeficiente Jaccard se calcula de la siguiente manera:

$$sim_{Jaccard}(A, B) = \frac{|A \cap B|}{|A \cup B|} = \frac{\sum_{i=1}^N \min(A_i B_i)}{\sum_{i=1}^N \max(A_i B_i)}$$

- El coeficiente Sørensen–Dice se calcula de la siguiente manera:

$$sim_{Dice}(A, B) = \frac{2|A \cap B|}{|A| + |B|} = \frac{2 \sum_{i=1}^N \min(A_i B_i)}{\sum_{i=1}^N (A_i + B_i)}$$

Para calcular las similitudes fenotípicas se han seguido los siguientes pasos:

- Se obtienen según una fuente y una *snapshot* la relación entre los elementos de diagnóstico y las enfermedades, tal y como se muestra en la Tabla 18. Estas relaciones se han formado considerando solo con aquellas enfermedades con al menos cinco DKE asociados con el fin de asegurar que entre enfermedades se tengan cinco DKE en común y así las similitudes altas no sean por compartir un solo DKE. Estos casos generan mucho ruido para el análisis.

Tabla 18. Ejemplo del resultado de consultar los elementos diagnósticos de cada enfermedad. Los datos presentados en esta tabla corresponden a la *snapshot* de Wikipedia del 1 de marzo de 2020.

SOURCE	SNAPSHOT	DISNET ID	NAME	CUI	NAME	SEMANTIC TYPES
Wikipedia	2020-03-01	DIS000004	Otic polyp	C0013456	Earache	sosy
Wikipedia	2020-03-01	DIS000004	Otic polyp	C0013604	Edema	fndg
Wikipedia	2020-03-01	DIS000004	Otic polyp	C0018777	Conductive hearing loss	dsyn
Wikipedia	2020-03-01	DIS000004	Otic polyp	C0019080	Hemorrhage	patf
Wikipedia	2020-03-01	DIS000004	Otic polyp	C0155540	Ear discharge	fndg
Wikipedia	2020-03-01	DIS000004	Otic polyp	C0221198	Lesion	fndg

4.7 Similitud en base a características fenotípicas

- ii. Ahora, para cada enfermedad del conjunto de datos resultantes del paso anterior se calcula el número de elementos de diagnóstico. En la Tabla 19 se muestra un ejemplo de los datos obtenidos.

Tabla 19. Ejemplo del conjunto de datos resultantes de sumar el número de DKE para cada una de las enfermedades. Los datos presentados en esta tabla corresponden a la *snapshot* de Wikipedia del 1 de marzo de 2020.

SOURCE	SNAPSHOT	DISNET ID	CUI COUNT (DKE COUNT)
Wikipedia	2020-03-01	DIS000004	6
Wikipedia	2020-03-01	DIS000005	16
Wikipedia	2020-03-01	DIS000020	41
Wikipedia	2020-03-01	DIS000025	8

- iii. A continuación, se realizan las combinaciones de enfermedades con DKE coincidentes sin agrupar posibles, omitiendo la relación de una enfermedad con si misma como parte del proceso. En la Tabla 20 se muestra un ejemplo de esta combinación entre enfermedades

Tabla 20. Ejemplo del conjunto de datos resultante de realizar todas las posibles asociaciones enfermedad–enfermedad. Los datos presentados en esta tabla corresponden a la *snapshot* de Wikipedia del 1 de marzo de 2020.

SOURCE	SNAPSHOT	DISNET ID 1	DISNET ID 2
Wikipedia	2020-03-01	DIS000004	DIS000020
Wikipedia	2020-03-01	DIS000004	DIS000072
Wikipedia	2020-03-01	DIS000004	DIS000072
Wikipedia	2020-03-01	DIS000004	DIS000086

- iv. Finalmente, por cada relación enfermedad–enfermedad se calculan los tres diferentes coeficientes de similitud relevantes usando las características fenotípicas en común. El resultado se muestra en la Tabla 21.

Tabla 21. Ejemplo del conjunto de datos resultante tras calcular tres índices de similitud para las relaciones enfermedad–enfermedad. Los datos presentados en esta tabla corresponden a la *snapshot* de Wikipedia del 1 de marzo de 2020.

SOURCE	SNAPSHOT	DISNET ID 1	DISNET ID 2	COSINE	JACCARD	DICE	AVG
Wikipedia	2020-03-01	DIS007698	DIS013214	1	1	1	1
Wikipedia	2020-03-01	DIS000206	DIS002838	0.92582	0.857143	0.923077	0.902013
Wikipedia	2020-03-01	DIS001442	DIS005374	0.877328	0.78125	0.877193	0.845257
Wikipedia	2020-03-01	DIS006035	DIS007263	0.856349	0.733333	0.846154	0.811945

Es por tanto, que este último conjunto de datos presentado en la Tabla 21 es el que ha usado para realizar el análisis de similitudes propuesto en esta sección.

4.7 Similitud en base a características fenotípicas

Por cada relación enfermedad–enfermedad además de calcular los tres diferentes coeficientes de similitud se ha calculado un promedio de las tres.

Se han calculado similitudes fenotípicas para la *snapshot* de Wikipedia del 1 de marzo de 2020.

En la Figura 45 se presenta una gráfica en la que se muestran para dicha *snapshot*, como se han caracterizado las enfermedades mediante la distribución de los valores de los índices de similitud. Es decir, se han distribuido los índices cuyo valor se encuentre entre 0,8 y 1 considerados como los más altos, seguidos de los índices con valor entre 0,8 y 0,6; después entre 0,4 y 0,6 y así sucesivamente.

Figura 45. Distribución de los coeficientes de similitud Coseno, Jaccard, Dice. La figura ha sido creada considerando la información de la *snapshot* de Wikipedia del 1 de marzo de 2020.

Lo primero que se puede observar en la Figura 45 es que hay una tendencia en las agrupaciones, ya que, a mayor valor de similitud, la cantidad de pares de enfermedades que pertenecen a dicho grupo es menor. Por ejemplo, considerando los valores de coseno, se observa que solo hay 11 pares enfermedad–enfermedad que tienen coeficientes muy altos. Por otro lado, la cantidad de pares enfermedad–enfermedad con coeficientes muy bajos es de poco más de tres millones. En este sentido, todos los coeficientes de similitud considerados como muy bajos (entre 0 y 0,2) son muy abundantes tal y como se puede ver en la distribución.

4.7 Similitud en base a características fenotípicas

Otra cuestión es que el promedio calculado tampoco ofrece mejoras en los valores de las similitudes, ya que se aprecia que se suavizan los índices, permitiendo que se pueda cubrir algún índice con un valor alto y relevante.

Dentro del mismo contexto de la Figura 45 se observa que el índice de similitud cuyos valores son más altos es el coeficiente Coseno. Y es que a pesar de la gran cantidad de valores muy bajos, es el coeficiente que ha ofrecido una distribución más homogénea. El siguiente coeficiente con valores más altos es Dice, dejando en el tercer lugar al coeficiente Jaccard.

En la Tabla 22 se muestran los valores promediados de los coeficientes de similitud calculados. En la que se puede confirmar que los resultados con valores más altos se han obtenido con el coeficiente Coseno. Los valores altos no son mejores pero son interesantes ya que con ellos se identifican pares enfermedad–enfermedad. Estas parejas de enfermedades permiten verificar a nivel biológico si esas mismas enfermedades comparten más que características fenotípicas, por ejemplo, si comparten características biológicas. Y estos tipos de hallazgos pueden apoyar al reposicionamiento de fármacos u otros campos de la medicina.

Tabla 22. Medias los valores de los coeficientes de similitud calculados. Considerando la *snapshot* de Wikipedia del 1 de marzo de 2020.

AVG COSINE	AVG JACCARD	AVG DICE
0.111	0.055	0.102

En la Tabla 23 se muestran los once pares enfermedad–enfermedad con coeficientes Coseno muy altos, tal y como también se muestra en la Figura 45. Lo primero que se debe observar en la Tabla 23 son los valores de la columna “*SIMILAR DKE COUNT*”, los cuales indican el número de DKE coincidentes entre las dos enfermedades. Por ejemplo, en la primera fila, la enfermedad con identificador DIS007698 tiene el coeficiente de similitud Coseno más alto, pero también sucede lo mismo para los coeficientes de Jaccard y Dice. Estos coeficientes tan altos se deben a que tanto la enfermedad DIS007698 como la DIS013214 tienen asociados cinco DKE y en ambas los cinco son los mismos. Estos identificadores se refieren a las enfermedades: “*Uterine serous carcinoma*” y “*Uterine clear-cell carcinoma*”.

Tabla 23. Información detallada de las once relaciones enfermedad–enfermedad cuyos coeficientes de similitud Coseno son muy altos. Considerando la *snapshot* de Wikipedia del 1 de marzo de 2020.

DISNET ID 1	DKE COUNT	DISNET ID 1	DKE COUNT	SIMILAR DKE COUNT	COSINE	JACCARD	DICE	PROM
DIS007698	5	DIS013214	5	5	1	1	1	1
DIS000206	7	DIS002838	6	6	0.92582	0.857143	0.923077	0.902013
DIS001442	28	DIS005374	29	25	0.877328	0.78125	0.877193	0.845257

4.7 Similitud en base a características fenotípicas

DISNET ID 1	DKE COUNT	DISNET ID 1	DKE COUNT	SIMILAR DKE COUNT	COSINE	JACCARD	DICE	PROM
DIS006035	15	DIS007263	11	11	0.856349	0.733333	0.846154	0.811945
DIS008676	9	DIS009369	10	8	0.843274	0.727273	0.842105	0.804217
DIS004121	8	DIS014151	9	7	0.824958	0.7	0.823529	0.782829
DIS003632	9	DIS004362	6	6	0.816497	0.666667	0.8	0.761054
DIS003463	21	DIS004560	14	14	0.816497	0.666667	0.8	0.761054
DIS010463	5	DIS013214	5	4	0.8	0.666667	0.8	0.755556
DIS007698	5	DIS010463	5	4	0.8	0.666667	0.8	0.755556
DIS005503	5	DIS013336	5	4	0.8	0.666667	0.8	0.755556

Tanto “*Uterine serous carcinoma*” como “*Uterine clear-cell carcinoma*” tienen asociados dentro del conjunto de datos médico de DISNET los mismos códigos de enfermedades: el código 182 de ICD-9-CM y el código C54.1 de ICD-10. El alto índice de similitud indica que estas enfermedades en Wikipedia son consideradas iguales tanto en el vocabulario ICD-9-CM como en el ICD-10.

Otra relación interesante en la Tabla 23 es la de las enfermedades con identificadores: DIS001442–DIS005374. Se observa que ambas tienen gran cantidad de DKE asociados y también tienen muchos DKE en común, 25. Estas enfermedades son: “*Soy allergy*” y “*Milk allergy*”.

Finalmente, en la Tabla 23 se observa que los coeficientes con los que se han obtenido valores de similitud más altos han sido Coseno y Dice.

En la Figura 46 se presentan enfermedades relacionadas por sus altos números de DKE en común, por ejemplo, “*Brain damage*” y “*Head injury*” tienen en común 42 DKE. Esta relación se muestra en la Figura 47.

El grafo de la Figura 46 se puede ver como las enfermedades se definen y se caracterizan en base a sus elementos comunes.

A modo de comparación, en la Figura 48 se presenta para la *snapshot* de Pub-Med del 3 de abril de 2018, la distribución que ha permitido caracterizar las enfermedades según el valor de los diferentes coeficientes de similitud. También en la Figura 49 se presenta para la *snapshot* de Mayo Clinic del 1 de marzo de 2020, la distribución que ha permitido caracterizar las enfermedades según el valor de los diferentes coeficientes de similitud.

4.7 Similitud en base a características fenotípicas

Figura 46. Grafo de enfermedades relacionadas según el alto número de DKE que tienen en común. El rango de DKE en común es entre 20 y 42 elementos. El grosor de las líneas de relación depende del número de DKE así como el color del mismo. El grafo se desarrolló usando información de la *snapshot* de Wikipedia del 1 de marzo de 2020.

Figura 47. Extracto de la Figura 46 en el que se aprecia la única relación ("*Head injury*"–"*Brain damage*") que se hace con más de 40 DKE. También se aprecia una de las enfermedades más relacionadas, la "*Dementia*".

4.7 Similitud en base a características fenotípicas

Figura 48. Distribución de los coeficientes de similitud Coseno, Jaccard, Dice. La figura ha sido creada considerando la información de la *snapshot* de PubMed del 3 de abril de 2018.

Figura 49. Distribución de los coeficientes de similitud Coseno, Jaccard, Dice. La figura ha sido creada considerando la información de la *snapshot* de Mayo Clinic del 1 de marzo de 2020.

4.7 Similitud en base a características fenotípicas

Es interesante observar que la caracterización realizada para PubMed mediante el coeficiente Coseno carece de valores altos, tal y como se aprecia en la Figura 48. Apenas cuenta con tres relaciones enfermedad–enfermedad dentro del rango de los valores más altos, sin embargo para los rangos entre 0,60 y 0,40 no hay ningún registro. A diferencia de los cálculos para los coeficientes Jaccard y Dice.

En contraste con la Figura 49 referente a Mayo Clinic, se observa que a pesar de contar con mucha menos cantidad de información hay resultados interesantes. Incluso en rangos de valores de coeficientes muy altos supera tanto a PubMed como a Wikipedia ya que en los tres índices calculados se encuentran entre 59 y 70 relaciones enfermedad–enfermedad. Esta comparación permite proponer que Mayo Clinic a pesar de contener poca información fenotípica es suficientemente buena como para relacionar enfermedades mediante el común de sus características fenotípicas. Finalmente, para ambas fuentes el coeficiente con valores más bajos es Jaccard y muy iguales se encuentra el coeficiente Coseno y Dice.

Figura 50. Distribución de los coeficientes de similitud Coseno, Jaccard, Dice considerando solo los DKE pertenecientes a los *semantic types*: "sosy" y "fndg". La figura ha sido creada considerando la información de la *snapshot* de Wikipedia del 1 de marzo de 2020.

4.8 Conclusiones

Lo primero que se puede observar en la Figura 50 es que el número de parejas enfermedad–enfermedad se ha visto disminuido a casi un tercio con respecto a los anteriores conjuntos formados usando los DKE de todos los *semantic types*. También se observa que, los valores altos en las parejas se han incrementado. En este sentido se ve una mejora porque hay mayor cantidad de parejas enfermedad–enfermedad que se pueden analizar para verificar si a nivel biológico también comparten algo más. Así que esto sugiere que la selección de DKE es un buen método para encontrar relaciones de enfermedades relevantes tanto en el nivel fenotípico como en el nivel biológico.

4.8 Conclusiones

En esta sección se han presentado los resultados derivados del análisis de similitudes y como estos permiten caracterizar enfermedades mediante la implementación de diferentes análisis sobre el conjunto de datos médicos de DISNET, con el fin de dilucidar su contenido. También en esta sección se ha presentado la validación del sistema encargado de extraer y generar con el tiempo conocimiento relativo al diagnóstico de enfermedades humanas.

La particularidad de contemplar el tiempo en los datos es algo único que se ha usado para definir las enfermedades en una dimensión diferente que el propio contenido. Además ha permitido presentar un análisis de la evolución del contenido de los artículos sobre enfermedades en la versión en inglés de Wikipedia, con especial atención a la forma en que esa evolución puede afectar a la recopilación automática de conocimientos médicos y su utilización en las tareas posteriores de la ciencia de los datos. En consecuencia, se consideraron tres parámetros de caracterización, que son: la longitud total del texto, el número de términos médicos (DKE) y el número de referencias.

Se han extraído algunas conclusiones generales, como, que el conocimiento médico a nivel mundial aumenta con el tiempo, tanto en lo que respecta a los caracteres como a los términos y las referencias; se observó que sólo el 10% de las enfermedades han experimentado una reducción general de su contenido. El tamaño de los artículos también se correlaciona con la popularidad de las enfermedades correspondientes, de manera que los “temas candentes” se visualizan y editan con mayor frecuencia. Así también, cuanto más largo es el texto, mayor es el número de DKE, aunque estos últimos no siempre están respaldados por las correspondientes referencias.

Además se ponen de relieve la importancia de la dimensión temporal en cualquier análisis basado en la Wikipedia, y cómo se debe prestar especial atención a los artículos “pequeños”. Los artículos pequeños rara vez se actualizan, pero cuando esto sucede, pueden introducirse cambios importantes en el contenido. Por

4.8 Conclusiones

lo tanto, es importante utilizar conjuntos de datos recientes, ya que la cantidad de información puede duplicarse en cuestión de días. Los artículos pequeños también son pequeños porque se refieren a temas especializados y, por lo tanto, reciben la atención de un número restringido de editores y espectadores. Esto implica un mayor riesgo de información sesgada o errónea, ya que menos editores colaboran en el proceso de redacción y menos lectores pueden dar aviso de correcciones.

Si los artículos pequeños son potencialmente menos fiables, los largos también presentan inconvenientes. Si bien su comportamiento suele ser más estable, pueden sufrir importantes cambios bruscos. Un ejemplo interesante es el de la “*Cerebellar hipoplasia*”, cuyo contenido se ha reducido a la mitad entre dos instantáneas consecutivas –ver Figura 42–. Además, los textos extensos del artículo no implican la presencia de un gran número de términos médicos, ni de referencias, lo que puede limitar su utilidad en las tareas de extracción automática de datos.

También se ha discutido sobre los identificadores de enfermedades, elementos clave que facilitan la definición de enfermedades tanto dentro del conjunto de datos DISNET como en vocabularios y bases de datos externas. Por minar Wikipedia y PubMed se han obtenido números códigos de enfermedades de diversas fuentes y vocabularios biomédicos. Por minar Mayo Clinic no se han conseguido ningún tipo de códigos de enfermedad.

El no contar con códigos de enfermedad supone una dificultad al caracterizar enfermedades porque aísla y no permite relacionar toda la información de una fuente con otras. Sin duda Mayo Clinic contiene información de calidad pero sus hallazgos se ven opacados por la falta de mapeos de enfermedades.

En este sentido, se abren muchas oportunidades para desarrollar nuevos métodos de caracterización que permitan relacionar enfermedades que sean alternativos al uso de códigos de enfermedades. Algunas iniciativas han desarrollado métodos de asociación según el nombre de la enfermedad (Espe, 2018).

Se ha discutido sobre la lista de enfermedades, como los identificadores son elementos vitales para una hacer que una lista sea lo más homogénea posible. Mientras más homogénea sea mayor será la capacidad de caracterizar, clasificar, integrar y relacionar información sobre diferentes aspectos de la enfermedad.

En el contexto de los identificadores de enfermedades, se han observado los mapeos realizados entre enfermedades de Wikipedia y PubMed. Las enfermedades mapeadas entre estas dos fuentes y también los mapeos realizados con la fuente de información DisGeNet han arrojado luz en cuanto a la necesidad de mejorar la lista de enfermedades mediante la implementación de jerarquías o clasificaciones.

Finalmente, se ha discutido la validación del sistema, sugiriendo que es lo suficientemente bueno como para ser utilizado para extraer enfermedades y términos

4.8 Conclusiones

relevantes para el diagnóstico. Al mismo tiempo, la evaluación también reveló que podrían introducirse mejoras para aumentar la fiabilidad del sistema.

Capítulo 5 Verificación de hipótesis

En el presente capítulo se describen como los experimentos de caracterización de enfermedades desarrollados en el Capítulo 4 validan las hipótesis planteadas en este trabajo de tesis, así como el cumplimiento de los objetivos establecidos en la misma.

Las hipótesis expuestas al comienzo de la investigación de esta tesis planteaban la posibilidad de permitir caracterizar enfermedades humanas en base a sus elementos de diagnóstico. Concretamente, las hipótesis establecidas son las siguientes:

- **Hipótesis 1 (H1):** Es posible extraer y generar información fenotípica proveniente de diversas fuentes de información pública, por medio de la implementación de un enfoque basado en el procesamiento de lenguaje natural (NLP *Natural-Language Processing*, por sus siglas en inglés).
- **Hipótesis 2 (H2):** La información fenotípica contenida en los artículos de Wikipedia clasificados como enfermedades es de calidad.
- **Hipótesis 3 (H3):** A medida que pasa el tiempo estas fuentes contienen más y mejor información que permiten enriquecer el conocimiento sobre las enfermedades.
- **Hipótesis 4 (H4):** La cantidad de elementos de conocimiento diagnóstico (DKE, por sus siglas en inglés) obtenidos de textos con contenido diagnóstico de los artículos de enfermedades en la versión inglesa de Wikipedia están directamente relacionados con el número de referencias científicas y el número de caracteres totales.
- **Hipótesis (H5):** Los elementos de conocimiento de diagnóstico (DKE, por sus siglas en inglés) obtenidos de fuentes textuales de conocimiento públicas son suficientemente relevantes para caracterizar enfermedades en base a sus características fenotípicas.

En el primer experimento se diseñó y desarrollo una plataforma web basada en micro servicios que de forma automática, dos veces por mes establece conexión con las fuentes de información públicas seleccionadas y recupera su contenido textual referente a las características de diagnóstico de enfermedades.

Después de almacenar toda la información diagnóstica en bruto (textos) y mediante un proceso de NLP que contempla la herramienta MetaMap y el proceso de evaluación TVP –ver sección 3.2– se obtienen y evalúan los conceptos UMLS que según el origen de los textos son conceptos médicos que permiten caracterizar

enfermedades y consecuentemente identificar o descartar el padecimiento en cuestión. Dicho proceso ha poblado gradualmente con cada extracción de información el conjunto de datos de DISNET.

Posteriormente, para verificar la precisión con la que el sistema web desarrollado obtiene conocimiento relevante de diagnóstico se ejecutó sobre los datos un proceso de validación que obtuvo una precisión del 0,73 para Wikipedia y 0,64 para PubMed. Un porcentaje que se considera como bueno para ser utilizado para extraer enfermedades y términos relevantes para el diagnóstico y que también se puede mejorar.

Con la generación del conjunto de datos y su posterior validación se permite la validación de las hipótesis **H1** y **H2**.

A continuación se presenta otro experimento en el cual se realizaron dos tipos de análisis que han permitido caracterizar enfermedades según:

- Un análisis descriptivo: se consideró todo el contenido del conjunto de datos médicos DISNET, es decir, todas las enfermedades, su información diagnóstica y demás datos recolectados (códigos de enfermedades, textos, enlaces, entre otros) provenientes de las fuentes de conocimiento relevantes para esta tesis. Como se encuentra organizadas, agrupadas, número de enfermedades, de DKE, de secciones, entre otros.
- Y un segundo análisis que observa cómo ha ido evolucionando el contenido médico de los artículos de enfermedades en Wikipedia según tres métricas clave: **i)** la variación del número de DKE obtenido en cada documento de enfermedad, **ii)** la variación del número de caracteres en cada documento de enfermedad y **iii)** la variación del número de referencias científicas (PubMed) en cada documento de enfermedad.

Por medio del primer análisis del experimento descrito anteriormente se ha podido verificar también la hipótesis **H2** ya que el análisis ha permitido ver en detalle, como con el paso del tiempo los elementos que caracterizan las enfermedades en las fuentes: textos, secciones relevantes, códigos de enfermedades, entre otros, se han ido incrementado y mejorando, para que consecuentemente a su vez se haya ido depurando el contenido. También los editores de artículos de enfermedades en Wikipedia y la misma organización por medio de cada vez más rigurosas reglas de edición permiten que la calidad en el contenido de este tipo de artículos mejore.

Además por medio del segundo análisis del experimento, por el que se han caracterizado las enfermedades según la evolución de las métricas: número de referencias, número de caracteres y número de DKE, se vuelve a comprobar la hipótesis **H2** ya que como se presenta en el cuarto cuadro de la Figura 39, se

describe la relación entre el número de referencias incluidas en cada artículo y *snapshots*, y el número de referencias que incluyen un código PMID; es de mencionar que la presencia de dicho código indica que la referencia es un documento médico y, por lo tanto, la fiabilidad del artículo es mayor al referenciar este tipo de contenido. Se observó que más de la mitad de las referencias son de documentos médicos ($r^2 = 0,822$), sugiriendo una buena calidad en los materiales de referencia externos.

En el segundo análisis elaborado se caracterizaron las enfermedades mediante la observación del comportamiento del conocimiento médico en Wikipedia. Se analizó como cambia con el tiempo, tanto en lo que respecta a la métrica de los caracteres como a los términos y las referencias; se observó que sólo el 10% de las enfermedades han experimentado una reducción general de su contenido, lo que significa que con este experimento se permite validar la hipótesis **H3**.

Pero también por este segundo análisis se rechaza la hipótesis **H4**, ya que en la Figura 39 se analiza cómo se relacionan las tres diferentes métricas descritas anteriormente entre ellas. La Figura 39 representa cuatro gráficos de densidad, que caracterizan las enfermedades relacionando el número de términos médicos, caracteres, referencias regulares y referencias PMID (PubMed ID) para un mismo documento de enfermedad en Wikipedia (los colores oscuros indican densidades más altas, por lo tanto un mayor número de *snapshots*, y las líneas punteadas azules los mejores ajustes lineales). Se observa una modesta correlación en el primer caso (*Pearson's* $r^2 = 0,5685$), y una menos definida en el segundo (*Pearson's* $r^2 = 0,5093$).

Así que, lo que se puede concluir por las caracterizaciones realizadas es que, cuanto más largo es el texto, mayor es el número de términos médicos en él incluidos; pero por otra parte, cabe la posibilidad de que los DKE no siempre estén respaldados por las correspondientes referencias.

Finalmente, por medio del experimento sobre el cálculo de los coeficientes de similitudes de los pares enfermedad-enfermedad se ha comprobado la hipótesis **H5**. Ya que por medio de la revisión de las distribuciones de los índices de similitud se puede observar cómo la información extraída permite obtener relaciones significativas (con una mayor significancia en determinadas fuentes y según el método de similitud) en como las enfermedades se han definido en base a sus elementos fenotípicos comunes.

Capítulo 6 Conclusiones y trabajos futuros

La presente tesis doctoral se ha centrado en la caracterización de las enfermedades en base a sus características fenotípicas contenidas en un conjunto de datos médicos creado a partir de minar fuentes de información biomédica textuales públicas. En esta sección se describen las conclusiones y las líneas futuras de investigación posibles para el presente trabajo doctoral.

6.1 Conclusiones

La investigación sobre la que se ha centrado esta tesis ha permitido demostrar en primer lugar el objetivo principal así como cada uno de los objetivos específicos. A continuación se describe como se han conseguido los objetivos –sección 1.4– y por lo tanto como se han verificado las hipótesis –sección 1.3–.

En primer lugar, dentro del contexto del primer objetivo (**O1**), la presente tesis doctoral se ha centrado en un exhaustivo análisis de la literatura actual sobre los diversos aspectos que se tratan en esta tesis. Los aspectos analizados han permitido obtener información sobre las características más relevantes de los diferentes fuentes de conocimiento biomédico usados en la actualidad. Con el fin de implementarlas para la construcción de un conjunto de datos único que se ha usado para comprobar las hipótesis que han dirigido esta investigación.

También se analizaron las características de las diferentes enfoques de extracción y generación de conocimiento biomédico. Como forma muy particular se ha puesto atención en los trabajos que han caracterizado enfermedades relacionándolas en base a las similitudes de diferentes aspectos. Finalmente, se han estudiado las redes de enfermedades complejas ya que la presente tesis es parte del proyecto DISNET –ver sección 3.2.5–, el cual contempla el desarrollo de una red de enfermedades multicapa compleja para apoyar el estudio del reposicionamiento de fármacos. El estudio de las redes de enfermedades ha permitido ver que el conjunto de datos creado para el desarrollo de esta tesis es viable para la construcción de redes de enfermedades.

En segundo lugar, dentro del contexto de los objetivos **O2** y **O3**, se ha desarrollado un método único de extracción y generación de conocimiento basado en la obtención de elementos de conocimiento de diagnóstico a partir de piezas de textos biomédicos. Este enfoque de extracción ha permitido no solo extraer información biomédica de diferentes fuentes sino que también con diferentes estructuras de datos cada una. Se ha usado como método de NLP un enfoque ya existente (Rodríguez González et al., 2015) que ha permitido obtener elementos de conocimiento diagnóstico relevantes para la caracterización y definición de las distintas enfermedades contenidas en el conjunto de datos.

6.1 Conclusiones

Al alcanzar los tres primeros objetivos particulares de esta tesis se ha podido asegurar la comprobación de la hipótesis **H1**.

En tercer lugar, dentro del contexto del objetivo **O4**, se ha verificado que la información resultante del método de extracción y generación de conocimiento diagnóstico es buena para ser considerada para su uso en otros trabajos de investigación científica. En este sentido se realizó un proceso de validación del conocimiento diagnóstico obtenido con resultados que permiten proponer que el conjunto de datos médicos de DISNET es lo suficientemente bueno para poder ser reutilizado. El proceso de validación asegura la verificación de la hipótesis **H2**.

Hasta este punto es importante hacer un paréntesis que permita aclarar que el enfoque adoptado en este trabajo doctoral para construir un conjunto de datos médicos referente al diagnóstico de enfermedades es válido, pero también es mejorable.

Se pueden hacer diferentes cambios para incrementar la calidad del presente trabajo. Por ejemplo, incrementar el número de fuentes biomédicas referentes al diagnóstico de enfermedades y también de otros aspectos de la enfermedad. Otra opción de mejora, sería hacer uso de más de una herramienta de procesamiento de lenguaje natural para tratar de mejorar los resultados relativos a extracción de conocimiento de los textos. Finalmente, para incluir a la comunidad investigadora y del campo de la salud, se podría crear un sitio web que permita a todo aquel interesado la opción de validar la información contenida en el conjunto de datos médicos de DISNET.

En tercer lugar, dentro del contexto del objetivo **O5**, se plantea el desarrollo de una API que proporciona acceso libre a todo aquel usuario registrado en el sitio web del proyecto DISNET (disnet.ctb.upm.es) para consumir de forma estructurada toda la información diagnóstica extraída y generada. Con este objetivo se aporta relevancia a la verificación de la hipótesis **H2**, porque tener un medio estructurado de compartición de conocimiento es fundamental para fomentar el estudio de las enfermedades.

Por último, y dentro del contexto de los objetivos **O6**, **O7** y **O8**, el presente trabajo ha desarrollado un serie de análisis que han permitido definir y caracterizar el conjunto de enfermedades pertenecientes al conjunto de datos médicos de DISNET.

Todos los experimentos han sido centrados sobre Wikipedia, la fuente de libre acceso cuyas características han permitido observar como se comporta el contenido de los artículos de enfermedad en un intervalo definido de tiempo. Y por medio de este experimento se ha logrado verificar la hipótesis **H3** ya que los resultados ponen en evidencia que con el tiempo el contenido médico en Wikipedia se incrementa. La hipótesis **H4** es la que por medio del análisis evolutivo del

6.1 Conclusiones

contenido en Wikipedia se ha rechazado porque el experimento sugirió que el número de DKE solo está relacionado directamente con la longitud del contenido médico.

Otros análisis que se han realizado han permitido ver el estado del conocimiento médico contenido en Wikipedia en una franja temporal de apenas dos años. También se ha podido estudiar el potencial de Wikipedia con respecto a obtención de elementos de conocimiento diagnóstico. Esto puede indicar que también Wikipedia puede servir para obtener información de otros aspectos de las enfermedades, por ejemplo, información referente a la prevalencia, a los fármacos, o a los tratamientos, entre otras.

Al realizar comparaciones entre Wikipedia, PubMed y Mayo Clinic se puede concluir que las tres fuentes son muy buenas proveedoras de conocimiento diagnóstico. Claramente y dada la cantidad de información en Wikipedia se sugiere que la contribución de muchos profesionales e investigadores impacta más positivamente, que si la actualización de la información está a cargo de un grupo cerrado.

En este sentido se observa que el contenido de Mayo Clinic con respecto al de Wikipedia tiene poca variabilidad y por lo tanto el conocimiento es más estático, pero de buena calidad. En este punto esta fuente tiene un gran inconveniente y es que ninguno de sus enfermedades cuenta con identificador que permita relacionar y conectar sus datos con otras fuentes o vocabularios biomédicos.

Por otro lado, PubMed es una referencia de fuente de información biomédica pero presenta ciertas desventajas. Primero porque está compuesto por millones de documentos científicos que son de gran calidad, pero si se quiere obtener información de una sola enfermedad es necesario recuperar y tratar grandes cantidades de artículos, con los recursos computacionales asociados que esto conlleva. A esto hay que sumar que no siempre se tiene libre acceso al contenido completo de los artículos y aunque se tuviera, este adolece de una estructura común que permita un procesamiento y extracción de información eficiente.

Finalmente, con la lista de enfermedades y sus DKE se han realizado experimentos que han permitido relacionar aquellas enfermedades que comparten elementos de diagnóstico. Las relaciones enfermedad-enfermedad se han construido mediante tres coeficientes de similitud diferentes: Coseno, Jaccard y Dice. Estos cálculos de similitud han permitido caracterizar las enfermedades y agruparlas en base a sus coeficientes de similitud. Estos coeficientes permiten observar cuáles son las características que poseen estos grupos en la definición de las propias enfermedades. En base a estos experimentos desarrollados se ha podido verificar la hipótesis **H5**.

6.2 Líneas futuras

Esta tesis doctoral siguió un enfoque determinado para extraer y generar conocimiento diagnóstico a partir de fuentes de información públicas. Y esto ha permitido primero, validar que es un enfoque valido y segundo, que se pueden mejorar y por lo tanto a continuación se describen las líneas futuras de investigación y desarrollo que se podrían seguir para dar más valor a los construido en esta tesis.

Lo primero que se puede abordar, es la integración de conocimiento fenotípico de muchas más fuentes de información biomédica. En este sentido, no solo incorporar información fenotípica sino contemplar la idea de obtener información de muchos más aspectos de las enfermedades, como información sobre fármacos, tratamientos, prevalencias, genes, entre otros. También, se debe considerar que las fuentes de información que se integren deben cubrir como requisito tener asociados a sus enfermedades, identificadores propios como externos. Y mucho más importante será la depuración de la lista de enfermedades para conseguir una lista mucho más homogénea y con muchos más mecanismos capaces de conectar con datos en vocabularios externos. La implementación de jerarquías, clasificaciones tanto internas como exteriores, con el fin de no desaprovechar el conocimiento existente.

También otra línea de investigación es desarrollar o implementar nuevos enfoques de NLP, que permitan mejorar la obtención de DKE de textos. Y no limitarse a textos en inglés, es verdad que existe mucha cantidad importante de contenido médico en este lenguaje, pero también en otros se puede encontrar información relevante en otros idiomas como el chino o el español.

También desarrollar otros tipos de análisis que permitan identificar que tan bien se definen las enfermedades con respecto a los DKE.

Alineado a los objetivos del proyecto DISNET se debe seguir con la construcción de la capa de información biológica y para ello, se debe también construir una red de enfermedades–DKE que permita identificar las relaciones enfermedad–enfermedad con mayor índice de similitud para analizar en la capa biológica si esas enfermedades comparten más que DKE.

Capítulo 7 Bibliografía

Aibar E. 2017. La ciencia de la Wikipedia. Available at <https://metode.es/revistas-metode/article-revistas/la-ciencia-de-la-wikipedia.html> (accessed February 18, 2018).

Al Tamime R, Giordano R, Hall W. 2018. Humans and bots in controversial environments: A closer look at their interactions. *Qatar Foundation Annual Research Conference Proceedings 2018:ICTPD425*. DOI: 10.5339/qfarc.2018.IC-TPD425.

Al-Safadi L, Alomran R, Almutairi F. 2013. Evaluation of Metamap Performance in Radiographic Images Retrieval. *Research Journal of Applied Sciences, Engineering and Technology* 6:4231–4236. DOI: 10.19026/rjaset.6.3538.

Amberger J, Bocchini C, Hamosh A. 2011. A new face and new challenges for Online Mendelian Inheritance in Man (OMIM®). *Human Mutation* 32:564–567. DOI: 10.1002/humu.21466.

Amberger JS, Bocchini CA, Schiettecatte F, Scott AF, Hamosh A. 2015. OMIM.org: Online Mendelian Inheritance in Man (OMIM®), an online catalog of human genes and genetic disorders. *Nucleic Acids Research* 43:D789–D798. DOI: 10.1093/nar/gku1205.

Aronson AR. 2001a. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. *Proceedings. AMIA Symposium*:17–21.

Aronson AR. 2001b. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. *Proceedings. AMIA Symposium*:17–21.

Aronson AR. 2006. Metamap: Mapping text to the umls metathesaurus. *Bethesda, MD: NLM, NIH, DHHS*:1–26.

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G. 2000. Gene Ontology: tool for the unification of biology. *Nature genetics* 25:25–29. DOI: 10.1038/75556.

Auer S, Bizer C, Kobilarov G, Lehmann J, Cyganiak R, Ives Z. 2007. DBpedia: A Nucleus for a Web of Open Data. In: *The Semantic Web*. Lecture Notes in Computer Science. Springer, Berlin, Heidelberg, 722–735. DOI: 10.1007/978-3-540-76298-0_52.

Azer SA. 2015. Is Wikipedia a reliable learning resource for medical students? Evaluating respiratory topics. *Advances in Physiology Education* 39:5–14. DOI: 10.1152/advan.00110.2014.

Azzam A, Bresler D, Leon A, Maggio L, Whitaker E, Heilman J, Orlowitz J, Swisher V, Raspberry L, Otoide K, Trotter F, Ross W, McCue JD. 2017. Why Medical Schools Should Embrace Wikipedia: Final-Year Medical Student Contributions to Wikipedia Articles for Academic Credit at One School. *Academic Medicine* 92:194–200. DOI: 10.1097/ACM.0000000000001381.

Barrett DJ. 2008. *MediaWiki: Wikipedia and Beyond*. O'Reilly Media, Inc.

Bauer-Mehren A, Bundschuh M, Rautschka M, Mayer MA, Sanz F, Furlong LI. 2011. Gene-Disease Network Analysis Reveals Functional Modules in Mendelian, Complex and Environmental Diseases. *PLOS ONE* 6:e20284. DOI: 10.1371/journal.pone.0020284.

Bodenreider O. 2004. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Research* 32:D267–D270. DOI: 10.1093/nar/gkh061.

Buneman P. 2009. Curated Databases. In: Agosti M, Borbinha J, Kapidakis S, Papatheodorou C, Tsakonas G eds. *Research and Advanced Technology for Digital Libraries*. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, 2–2. DOI: 10.1007/978-3-642-04346-8_2.

Canese K, Weis S. 2013. *PubMed: The Bibliographic Database*. National Center for Biotechnology Information (US).

Chen X, Gururaj AE, Ozyurt B, Liu R, Soysal E, Cohen T, Tiriyaki F, Li Y, Zong N, Jiang M, Rogith D, Salimi M, Kim H, Rocca-Serra P, Gonzalez-Beltran A, Farcas C, Johnson T, Margolis R, Alter G, Sansone S-A, Fore IM, Ohno-Machado L, Grethe JS, Xu H. 2018. DataMed – an open source discovery index for finding biomedical datasets. *Journal of the American Medical Informatics Association* 25:300–308. DOI: 10.1093/jamia/ocx121.

Chen Y, Zhang X, Zhang G, Xu R. 2015. Comparative analysis of a novel disease phenotype network based on clinical manifestations. *Journal of Biomedical Informatics* 53:113–120. DOI: 10.1016/j.jbi.2014.09.007.

Cheng L, Jiang Y, Wang Z, Shi H, Sun J, Yang H, Zhang S, Hu Y, Zhou M. 2016. DisSim: an online system for exploring significant similar diseases and exhibiting potential therapeutic drugs. *Scientific Reports* 6. DOI: 10.1038/srep30024.

CMS. 2019.HCPCS - General Information. Available at <https://www.cms.gov/Medicare/Coding/MedHCPCSGenInfo/index.html?redirect=/MedHCPCSGenInfo/> (accessed November 6, 2019).

Cohen N. 2013. Editing Wikipedia Pages for Med School Credit. *The New York Times*.

Cousyn C, Bouchard K, Bouchard B, Gaboury S. 2018. Automated Extraction of Symptoms Related to Rare Diseases from Scientific Publications. In: *Proceedings of the 4th EAI International Conference on Smart Objects and Technologies for Social Good*. Goodtechs '18. New York, NY, USA: ACM, 13–18. DOI: 10.1145/3284869.3284892.

Cui L, Sahoo SS, Lhatoo SD, Garg G, Rai P, Bozorgi A, Zhang G-Q. 2014. Complex epilepsy phenotype extraction from narrative clinical discharge summaries. *Journal of biomedical informatics* 51:272–279. DOI: 10.1016/j.jbi.2014.06.006.

Deng L, Ye D, Zhao J, Zhang J. 2019. MultiSourceDSim: an integrated approach for exploring disease similarity. *BMC Medical Informatics and Decision Making* 19:269. DOI: 10.1186/s12911-019-0968-8.

Densen P. 2011. Challenges and Opportunities Facing Medical Education. *Transactions of the American Clinical and Climatological Association* 122:48–58.

Dias G, Oliveira JL, Vicente F-J, Martín-Sánchez F. 2005. Integration of Genetic and Medical Information Through a Web Crawler System. In: Oliveira JL, Maojo V, Martín-Sánchez F, Pereira AS eds. *Biological and Medical Data Analysis*. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 78–88. DOI: https://doi.org/10.1007/11573067_9.

Donnelly K. 2006. SNOMED-CT: The advanced terminology and coding system for eHealth. *Studies in Health Technology and Informatics* 121:279–290.

van Driel MA, Bruggeman J, Vriend G, Brunner HG, Leunissen JAM. 2006. A text-mining analysis of the human genome. *European journal of human genetics: EJHG* 14:535–542. DOI: 10.1038/sj.ejhg.5201585.

Elsevier. Visión panorámica del enfermo desde el punto de vista del estudiante de Medicina. Available at <https://www.elsevier.com/es-es/connect/estudiantes-de-ciencias-de-la-salud/vision-panoramica-del-enfermo-desde-el-punto-de-vista-del-estudiante-de-medicina> (accessed July 17, 2019).

Eppig JT, Blake JA, Bult CJ, Kadin JA, Richardson JE. 2015. The Mouse Genome Database (MGD): facilitating mouse as a model for human biology and disease. *Nucleic Acids Research* 43:D726–D736. DOI: 10.1093/nar/gku967.

Espe S. 2018. Malacards: The Human Disease Database. *Journal of the Medical Library Association: JMLA* 106:140–141. DOI: 10.5195/jmla.2018.253.

Fabregat A, Jupe S, Matthews L, Sidiropoulos K, Gillespie M, Garapati P, Haw R, Jassal B, Korninger F, May B, Milacic M, Roca CD, Rothfels K, Sevilla C, Shamovsky V, Shorser S, Varusai T, Viteri G, Weiser J, Wu G, Stein L, Hermjakob H, D'Eustachio P. 2018. The Reactome Pathway Knowledgebase. *Nucleic Acids Research* 46:D649–D655. DOI: 10.1093/nar/gkx1132.

Fiorini N, Leaman R, Lipman DJ, Lu Z. 2018. How user intelligence is improving PubMed. *Nature Biotechnology*. DOI: 10.1038/nbt.4267.

Friedlin J, McDonald CJ. 2010. An evaluation of medical knowledge contained in Wikipedia and its use in the LOINC database. *Journal of the American Medical Informatics Association: JAMIA* 17:283–287. DOI: 10.1136/jamia.2009.001180.

Goh K-I, Cusick ME, Valle D, Childs B, Vidal M, Barabási A-L. 2007. The human disease network. *Proceedings of the National Academy of Sciences* 104:8685–8690. DOI: 10.1073/pnas.0701361104.

Gooch P, Roudsari A. 2011. A tool for enhancing MetaMap performance when annotating clinical guideline documents with UMLS concepts. In: Bled, Slovenia,.

Groza T, Köhler S, Moldenhauer D, Vasilevsky N, Baynam G, Zemojtel T, Schriml LM, Kibbe WA, Schofield PN, Beck T, Vasant D, Brookes AJ, Zankl A, Washington NL, Mungall CJ, Lewis SE, Haendel MA, Parkinson H, Robinson PN. 2015. The Human Phenotype Ontology: Semantic Unification of Common and Rare Disease. *American Journal of Human Genetics* 97:111–124. DOI: 10.1016/j.ajhg.2015.05.020.

Guo D, Li M, Yu Y, Li Y, Duan G, Wu F, Wang J. 2018. Disease Inference with Symptom Extraction and Bidirectional Recurrent Neural Network. In: *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. 864–868. DOI: 10.1109/BIBM.2018.8621182.

Gutiérrez-Sacristán A, Grosdidier S, Valverde O, Torrens M, Bravo À, Piñero J, Sanz F, Furlong LI. 2015. PsyGeNET: a knowledge platform on psychiatric disorders and their genes. *Bioinformatics* 31:3075–3077. DOI: 10.1093/bioinformatics/btv301.

Halu A, Domenico MD, Arenas A, Sharma A. 2019. The multiplex network of human diseases. *npj Systems Biology and Applications* 5:15. DOI: 10.1038/s41540-019-0092-5.

Hamosh A, Scott AF, Amberger JS, Bocchini CA, McKusick VA. 2005. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and

genetic disorders. *Nucleic Acids Research* 33:D514–D517. DOI: 10.1093/nar/gki033.

Hassan M, Coulet A, Toussaint Y. 2014. Learning Subgraph Patterns from Text for Extracting Disease-symptom Relationships. In: *Proceedings of the 1st International Conference on Interactions Between Data Mining and Natural Language Processing - Volume 1202*. DMNLP’14. Aachen, Germany, Germany: CEUR-WS.org, 81–96.

Hassan M, Makkaoui O, Coulet A, Toussaint Y. 2015. Extracting Disease-Symptom Relationships by Learning Syntactic Patterns from Dependency Graphs. In: 184.

Heilman JM, West AG. 2015. Wikipedia and Medicine: Quantifying Readership, Editors, and the Significance of Natural Language. *Journal of Medical Internet Research* 17. DOI: 10.2196/jmir.4069.

Herrero Jaén S. 2016. Formalización del concepto de salud a través de la lógica: impacto del lenguaje formal en las ciencias de la salud. *Ene* 10:0–0.

Himmelstein DS, Lizee A, Hessler C, Brueggeman L, Chen SL, Hadley D, Green A, Khankhanian P, Baranzini SE. 2017. Systematic integration of biomedical knowledge prioritizes drugs for repurposing. *eLife* 6. DOI: 10.7554/eLife.26726.

Hoehndorf R, Schofield PN, Gkoutos GV. 2011. PhenomeNET: a whole-phenome approach to disease gene discovery. *Nucleic Acids Research* 39:e119–e119. DOI: 10.1093/nar/gkr538.

Hoehndorf R, Schofield PN, Gkoutos GV. 2015. Analysis of the human diseasesome using phenotype similarity between common, genetic, and infectious diseases. *Scientific Reports* 5:srep10888. DOI: 10.1038/srep10888.

Huete García A, Díaz Velázquez E, Díaz García E, Sola Bautista A, Petisco Rodríguez E, Lara Gonzalo P. 2009. Estudio ENSERio I. Available at <https://enfermedades-raras.org/index.php/estudio-enserio/1281-estudio-enserio> (accessed April 23, 2019).

Humphreys G. 2012. WHO | Coming together to combat rare diseases. Available at <https://www.who.int/bulletin/volumes/90/6/12-020612/en/> (accessed April 23, 2019).

Hwang S, Kim CY, Yang S, Kim E, Hart T, Marcotte EM, Lee I. 2019. HumanNet v2: human gene networks for disease research. *Nucleic Acids Research* 47:D573–D580. DOI: 10.1093/nar/gky1126.

Information NC for B, Pike USNL of M 8600 R, MD B, Usa 20894. 2009. *Metathesaurus*. National Library of Medicine (US).

Jia J, An Z, Ming Y, Guo Y, Li W, Li X, Liang Y, Guo D, Tai J, Chen G, Jin Y, Liu Z, Ni X, Shi T. 2018. PedAM: a database for Pediatric Disease Annotation and Medicine. *Nucleic Acids Research* 46:D977–D983. DOI: 10.1093/nar/gkx1049.

Kanehisa M, Goto S. 2000. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research* 28:27–30. DOI: 10.1093/nar/28.1.27.

Karczewski KJ, Weisburd B, Thoma B, Solomonson M, Ruderfer DM, Kavanagh D, Hamamsy T, Lek M, Samocha KE, Cummings BB, Birnbaum D, Daly MJ, The Exome Aggregation Consortium, MacArthur DG. 2017. ExAC browser: displaying reference data information from over 60 000 exomes | *Nucleic Acids Research* | Oxford Academic. *Nucleic Acids Research* 45:D840–D845. DOI: <https://doi.org/10.1093/nar/gkw971>.

Kessler G. 2016. Are there really 10,000 diseases and just 500 ‘cures’? Available at <https://www.washingtonpost.com/news/fact-checker/wp/2016/11/17/are-there-really-10000-diseases-and-500-cures/> (accessed March 29, 2019).

Kibbe WA, Arze C, Felix V, Mitra E, Bolton E, Fu G, Mungall CJ, Binder JX, Malone J, Vasant D, Parkinson H, Schriml LM. 2015. Disease Ontology 2015 update: an expanded and updated database of human diseases for linking biomedical knowledge through disease data. *Nucleic Acids Research* 43:D1071–D1078. DOI: 10.1093/nar/gku1011.

Köhler S, Vasilevsky NA, Engelstad M, Foster E, McMurry J, Aymé S, Baynam G, Bello SM, Boerkoel CF, Boycott KM, Brudno M, Buske OJ, Chinnery PF, Cipriani V, Connell LE, Dawkins HJS, DeMare LE, Devereau AD, de Vries BBA, Firth HV, Freson K, Greene D, Hamosh A, Helbig I, Hum C, Jähn JA, James R, Krause R, F. Laulederkind SJ, Lochmüller H, Lyon GJ, Ogishima S, Olry A, Ouwehand WH, Pontikos N, Rath A, Schaefer F, Scott RH, Segal M, Sergouniotis PI, Sever R, Smith CL, Straub V, Thompson R, Turner C, Turro E, Veltman MWM, Vulliamy T, Yu J, von Ziegenweidt J, Zankl A, Züchner S, Zemojtel T, Jacobsen JOB, Groza T, Smedley D, Mungall CJ, Haendel M, Robinson PN. 2017. The Human Phenotype Ontology in 2017. *Nucleic Acids Research* 45:D865–D876. DOI: 10.1093/nar/gkw1039.

Ku L-W, Li W-L, Chang T-C. 2013. Disease Detection and Symptom Tracking by Retrieving Information from the Web. In: *AAAI Spring Symposium Series*.

Kumar R. 2010. Wiki Installation and Customization.

Landrum MJ, Lee JM, Riley GR, Jang W, Rubinstein WS, Church DM, Maglott DR. 2014. ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Research* 42:D980–D985. DOI: 10.1093/nar/gkt1113.

Leaman R, Islamaj Dogan R, Lu Z. 2013. DNorm: disease name normalization with pairwise learning to rank. *Bioinformatics (Oxford, England)* 29:2909–2917. DOI: 10.1093/bioinformatics/btt474.

Lehmann J, Isele R, Jakob M, Jentzsch A, Kontokostas D, Mendes PN, Hellmann S, Morsey M, van Kleef P, Auer S, Bizer C. 2015. DBpedia – A large-scale, multilingual knowledge base extracted from Wikipedia. *Semantic Web* 6:167–195. DOI: 10.3233/SW-140134.

Levenshtein VI. 1966. Binary codes capable of correcting deletions, insertions, and reversals. In: *Soviet physics doklady*. 707–710.

Li Y, Agarwal P. 2009. A Pathway-Based View of Human Diseases and Disease Relationships. *PLOS ONE* 4:e4346. DOI: 10.1371/journal.pone.0004346.

Liberzon A, Subramanian A, Pinchback R, Thorvaldsdóttir H, Tamayo P, Mesirov JP. 2011. Molecular signatures database (MSigDB) 3.0. *Bioinformatics* 27:1739–1740. DOI: 10.1093/bioinformatics/btr260.

Lipscomb CE. 2000. Medical Subject Headings (MeSH). *Bulletin of the Medical Library Association* 88:265–266.

Ljubescic N, Boras D, Bakaric N, Njavro J. 2008. Comparing Measures of Semantic Similarity. In: *ITI 2008 - 30th International Conference on Information Technology Interfaces*. 675–682. DOI: 10.1109/ITI.2008.4588492.

Lo Surdo P, Calderone A, Iannuccelli M, Licata L, Peluso D, Castagnoli L, Cesareni G, Perfetto L. 2018. DISNOR: a disease network open resource. *Nucleic Acids Research* 46:D527–D534. DOI: 10.1093/nar/gkx876.

Lopes P, Oliveira JL. 2013. An innovative portal for rare genetic diseases research: The semantic Diseasecard. *Journal of Biomedical Informatics* 46:1108–1115. DOI: 10.1016/j.jbi.2013.08.006.

Lowe HJ, Barnett GO. 1994. Understanding and using the medical subject headings (MeSH) vocabulary to perform literature searches. *JAMA* 271:1103–1108.

Maglott D, Ostell J, Pruitt KD, Tatusova T. 2005. Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Research* 33:D54–D58. DOI: 10.1093/nar/gki031.

Maiella S, Rath A, Angin C, Mousson F, Kremp O. 2013. Orphanet et son réseau: où trouver une information validée sur les maladies rares. *Revue Neurologique* 169:S3–S8. DOI: 10.1016/S0035-3787(13)70052-3.

Matheson D, Matheson C. 2017. Open Medicine Journal Wikipedia as Informal Self-Education for Clinical Decision-Making in Medical Practice. *Open Medicine Journal* 4:1–25. DOI: 10.2174/1874220301704010015.

McInnes BT, Pedersen T, Carlis J. 2007. Using UMLS Concept Unique Identifiers (CUIs) for Word Sense Disambiguation in the Biomedical Domain. *AMIA Annual Symposium Proceedings* 2007:533–537.

Menche J, Sharma A, Kitsak M, Ghiassian SD, Vidal M, Loscalzo J, Barabási A-L. 2015. Uncovering disease-disease relationships through the incomplete interactome. *Science* 347:1257601. DOI: 10.1126/science.1257601.

Mering C von, Huynen M, Jaeggi D, Schmidt S, Bork P, Snel B. 2003. STRING: a database of predicted functional associations between proteins. *Nucleic Acids Research* 31:258–261. DOI: 10.1093/nar/gkg034.

Miller N, Lacroix E-M, Backus JEB. 2000. MEDLINEplus: building and maintaining the National Library of Medicine’s consumer health Web service. *Bulletin of the Medical Library Association* 88:11–17.

Mouriño-García MA, Pérez-Rodríguez R, Anido-Rifón LE. 2017. A Bag of Concepts Approach for Biomedical Document Classification Using Wikipedia Knowledge. *Methods of Information in Medicine* 56:370–376. DOI: 10.3414/ME17-01-0028.

NCBI Resource Coordinators. 2016. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research* 44:D7-19. DOI: 10.1093/nar/gkv1290.

Oliveira JL, Dias G, Oliveira I, Rocha P, Hermosilla I, Vicente J, Spiteri I, Martín-Sánchez F, Pereira AS. 2004. DiseaseCard: A Web-Based Tool for the Collaborative Integration of Genetic and Medical Information. In: Barreiro JM, Martín-Sánchez F, Maojo V, Sanz F eds. *Biological and Medical Data Analysis*. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 409–417. DOI: https://doi.org/10.1007/978-3-540-30547-7_41.

OrphaNet. 2012. Orphanet: About rare diseases. Available at https://www.orpha.net/consor/cgi-bin/Education_AboutRareDiseases.php?lng=EN (accessed April 23, 2019).

Osborne JD, Gyawali B, Solorio T. 2014. Evaluation of YTEX and MetaMap for clinical concept recognition. *arXiv Cornell University*.

OXFORD. 2019.cure | Definition of cure in English by Oxford Dictionaries. Available at <https://en.oxforddictionaries.com/definition/cure> (accessed April 24, 2019).

Park A, Hartzler AL, Huh J, McDonald DW, Pratt W. 2015. Automatically Detecting Failures in Natural Language Processing Tools for Online Community Text. *Journal of Medical Internet Research* 17. DOI: 10.2196/jmir.4612.

Peng J, Xue H, Shao Y, Shang X, Wang Y, Chen J. 2016. Measuring phenotype semantic similarity using Human Phenotype Ontology. In: *2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. 763–766. DOI: 10.1109/BIBM.2016.7822617.

Perfetto L, Briganti L, Calderone A, Perpetuini AC, Iannuccelli M, Langone F, Licata L, Marinkovic M, Mattioni A, Pavlidou T, Peluso D, Petrilli LL, Pirrò S, Posca D, Santonico E, Silvestri A, Spada F, Castagnoli L, Cesareni G. 2016. SIGNOR: a database of causal relationships between biological entities. *Nucleic Acids Research* 44:D548–D554. DOI: 10.1093/nar/gkv1048.

Pesquita C, Faria D, Bastos H, Ferreira AE, Falcão AO, Couto FM. 2008. Metrics for GO based protein semantic similarity: a systematic evaluation. *BMC Bioinformatics* 9:S4. DOI: 10.1186/1471-2105-9-S5-S4.

Piñero J, Bravo À, Queralt-Rosinach N, Gutiérrez-Sacristán A, Deu-Pons J, Centeno E, García-García J, Sanz F, Furlong LI. 2017. DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. *Nucleic Acids Research* 45:D833–D839. DOI: 10.1093/nar/gkw943.

Piñero J, Queralt-Rosinach N, Bravo À, Deu-Pons J, Bauer-Mehren A, Baron M, Sanz F, Furlong LI. 2015. DisGeNET: a discovery platform for the dynamical exploration of human diseases and their genes. *Database* 2015. DOI: 10.1093/database/bav028.

Pizzorno J. 2016. Can We Say “Cure”? *Integrative Medicine: A Clinician’s Journal* 15:8–12.

Pletscher-Frankild S, Pallejà A, Tsafou K, Binder JX, Jensen LJ. 2015. DISEASES: Text mining and data integration of disease–gene associations. *Methods* 74:83–89. DOI: 10.1016/j.ymeth.2014.11.020.

Povey S, Lovering R, Bruford E, Wright M, Lush M, Wain H. 2001. The HUGO Gene Nomenclature Committee (HGNC). *Human Genetics* 109:678–680. DOI: 10.1007/s00439-001-0615-0.

Prieto Santamaría L. 2018.Extracción, cálculo de similitudes y análisis de características biológicas para la creación de redes humanas complejas. *Available at* <http://oa.upm.es/52568/> (accessed November 6, 2019).

pubmeddev. 2019.Home - PubMed - NCBI. *Available at* <https://www.ncbi.nlm.nih.gov/pubmed/> (accessed February 16, 2018).

RAE A. 2019.«Diccionario de la lengua española» - Edición del Tricentenario. *Available at* <https://dle.rae.es/srv/search?m=30&w=curar> (accessed April 24, 2019).

Ramos EM, Hoffman D, Junkins HA, Maglott D, Phan L, Sherry ST, Feolo M, Hindorff LA. 2014. Phenotype-Genotype Integrator (PheGenI): synthesizing genome-wide association study (GWAS) data with existing genomic resources. *European journal of human genetics: EJHG* 22:144–147. DOI: 10.1038/ejhg.2013.96.

Rappaport N, Nativ N, Stelzer G, Twik M, Guan-Golan Y, Iny Stein T, Bahir I, Belinky F, Morrey CP, Safran M, Lancet D. 2013. MalaCards: an integrated compendium for diseases and their annotation. *Database* 2013. DOI: 10.1093/database/bat018.

Rappaport N, Twik M, Nativ N, Stelzer G, Bahir I, Stein TI, Safran M, Lancet D. 2014. MalaCards: A Comprehensive Automatically-Mined Database of Human Diseases. *Current Protocols in Bioinformatics* 47:1.24.1-1.24.19. DOI: 10.1002/0471250953.bi0124s47.

Rappaport N, Twik M, Plaschkes I, Nudel R, Iny Stein T, Levitt J, Gershoni M, Morrey CP, Safran M, Lancet D. 2017. MalaCards: an amalgamated human disease compendium with diverse clinical and genetic annotation and structured search. *Nucleic Acids Research* 45:D877–D887. DOI: 10.1093/nar/gkw1012.

Rath A, Olry A, Dhombres F, Brandt MM, Urbero B, Ayme S. 2012. Representation of rare diseases in health information systems: The orphanet approach to serve a wide range of end users. *Human Mutation* 33:803–808. DOI: 10.1002/humu.22078.

Reátegui R, Ratté S. 2018. Comparison of MetaMap and cTAKES for entity extraction in clinical notes. *BMC Medical Informatics and Decision Making* 18:74. DOI: 10.1186/s12911-018-0654-2.

Rebhan M, Chalifa-Caspi V, Prilusky J, Lancet D. 1997. GeneCards: integrating information about genes, proteins and diseases. *Trends in Genetics* 13:163. DOI: 10.1016/S0168-9525(97)01103-7.

Riedel S. 2005. Edward Jenner and the history of smallpox and vaccination. *Proceedings (Baylor University. Medical Center)* 18:21–25.

Robinson PN, Köhler S, Bauer S, Seelow D, Horn D, Mundlos S. 2008. The Human Phenotype Ontology: A Tool for Annotating and Analyzing Human Hereditary Disease. *American Journal of Human Genetics* 83:610–615. DOI: 10.1016/j.ajhg.2008.09.017.

Rodríguez González A, Costumero Moreno R, Martínez Romero M, Wilkinson MD, Menasalvas Ruiz E. 2015. Diagnostic Knowledge Extraction from MedlinePlus: An Application for Infectious Diseases. *Current Bioinformatics* 375:79–87.

Rodríguez González A, Costumero Moreno R, Martínez Romero M, Wilkinson MD, Menasalvas Ruiz E. 2018. Extracting diagnostic knowledge from MedLine Plus: a comparison between MetaMap and cTAKES Approaches. *Current Bioinformatics* 13:573–582. DOI: 10.2174/1574893612666170727094502.

Rodríguez-González A, Alor-Hernández G. 2013. An approach for solving multi-level diagnosis in high sensitivity medical diagnosis systems through the application of semantic technologies. *Computers in Biology and Medicine* 43:51–62. DOI: 10.1016/j.compbiomed.2012.10.007.

Rouillard AD, Gundersen GW, Fernandez NF, Wang Z, Monteiro CD, McDermott MG, Ma’ayan A. 2016. The harmonizome: a collection of processed datasets gathered to serve and mine knowledge about genes and proteins. *Database* 2016. DOI: 10.1093/database/baw100.

Safran M, Solomon I, Shmueli O, Lapidot M, Shen-Orr S, Adato A, Ben-Dor U, Esterman N, Rosen N, Peter I, Olender T, Chalifa-Caspi V, Lancet D. 2002. GeneCardsTM 2002: towards a complete, object-oriented, human gene compendium. *Bioinformatics* 18:1542–1543. DOI: 10.1093/bioinformatics/18.11.1542.

Savova GK, Masanz JJ, Ogren PV, Zheng J, Sohn S, Kipper-Schuler KC, Chute CG. 2010. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. *Journal of the American Medical Informatics Association: JAMIA* 17:507–513. DOI: 10.1136/jamia.2009.001560.

Sayers E. 2010. *A General Introduction to the E-utilities*. National Center for Biotechnology Information (US).

Sayers E. 2018. *E-utilities Quick Start*. National Center for Biotechnology Information (US).

Schlegel DR, Crowner C, Elkin PL. 2015. Automatically Expanding the Synonym Set of SNOMED CT using Wikipedia. *Studies in Health Technology and Informatics* 216:619–623.

Schriml LM. 2009. Symptomontologywiki. Available at http://symptomontologywiki.igs.umaryland.edu/mediawiki/index.php/Main_Page (accessed February 3, 2020).

Schriml LM, Arze C, Nadendla S, Chang Y-WW, Mazaitis M, Felix V, Feng G, Kibbe WA. 2012. Disease Ontology: a backbone for disease semantic integration. *Nucleic Acids Research* 40:D940–D946. DOI: 10.1093/nar/gkr972.

Schriml LM, Mittraka E, Munro J, Tauber B, Schor M, Nickle L, Felix V, Jeng L, Bearer C, Lichenstein R, Bisordi K, Campion N, Hyman B, Kurland D, Oates CP, Kibbey S, Sreekumar P, Le C, Giglio M, Greene C. 2019. Human Disease Ontology 2018 update: classification, content and workflow expansion. *Nucleic Acids Research* 47:D955–D962. DOI: 10.1093/nar/gky1032.

Shafee T, Masukume G, Kipersztok L, Das D, Häggström M, Heilman J. 2017. Evolution of Wikipedia’s medical content: past, present and future. *J Epidemiol Community Health*:jech-2016-208601. DOI: 10.1136/jech-2016-208601.

Sherry ST, Ward M-H, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K. 2001. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Research* 29:308–311. DOI: 10.1093/nar/29.1.308.

Siva N. 2008. 1000 Genomes project. *Nature Biotechnology* 26:256–256. DOI: 10.1038/nbt0308-256b.

Stenson PD, Mort M, Ball EV, Shaw K, Phillips AD, Cooper DN. 2014. The Human Gene Mutation Database: building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine. *Human Genetics* 133:1–9. DOI: 10.1007/s00439-013-1358-4.

Sudeshna P, Bhanumathi S, Hamlin MRA. 2017. Identifying symptoms and treatment for heart disease from biomedical literature using text data mining. In: *2017 International Conference on Computation of Power, Energy Information and Commuincation (ICCPEIC)*. 170–174. DOI: 10.1109/ICCPEIC.2017.8290359.

The Gene Ontology C, Acencio ML, Lægreid A, Kuiper M, Among O. 2019. The Gene Ontology Resource: 20 years and still GOing strong. *Nucleic Acids Research*. DOI: 10.1093/nar/gky1055.

Traverso-Ribón I, Vidal M-E. 2018. GARUM: A Semantic Similarity Measure Based on Machine Learning and Entity Characteristics. In: Hartmann S, Ma H,

Hameurlain A, Pernul G, Wagner RR eds. *Database and Expert Systems Applications*. Lecture Notes in Computer Science. Cham: Springer International Publishing, 169–183. DOI: 10.1007/978-3-319-98809-2_11.

Trueba-Gómez R, Estrada-Lorenzo J-M. 2010. La base de datos PubMed y la búsqueda de información científica. *Seminarios de la Fundación Española de Reumatología* 11:49–63. DOI: 10.1016/j.semreu.2010.02.005.

Tsumoto S, Kimura T, Iwata H, Hirano S. 2017. Mining Text for Disease Diagnosis. *Procedia Computer Science* 122:1133–1140. DOI: 10.1016/j.procs.2017.11.483.

Upton F, DeGette D. 2015. Can we find cures for 7,000 diseases? Available at <https://www.cnn.com/2015/01/13/opinion/upton-degette-cure-diseases/index.html> (accessed June 12, 2018).

Valle EPG del, García GL, Santamaría LP, Zanin M, Ruiz EM, González AR. 2018. Evaluating Wikipedia as a Source of Information for Disease Understanding. In: *2018 IEEE 31st International Symposium on Computer-Based Medical Systems (CBMS)*. 399–404. DOI: 10.1109/CBMS.2018.00076.

Wakap SN, Lambert DM, Olry A, Rodwell C, Gueydan C, Lanneau V, Murphy D, Cam YL, Rath A. 2019. Estimating cumulative point prevalence of rare diseases: analysis of the Orphanet database. *European Journal of Human Genetics*:1–9. DOI: 10.1038/s41431-019-0508-0.

Wan Q, Dingerdissen H, Fan Y, Gulzar N, Pan Y, Wu T-J, Yan C, Zhang H, Mazumder R. 2015. BioXpress: an integrated RNA-seq-derived gene expression database for pan-cancer analysis. *Database: The Journal of Biological Databases and Curation* 2015. DOI: 10.1093/database/bav019.

Wang X, Gulbahce N, Yu H. 2011. Network-based methods for human disease gene prediction. *Briefings in Functional Genomics* 10:280–293. DOI: 10.1093/bfpg/elr024.

Weinreich SS, Mangon R, Sikkens JJ, Teeuw ME, Cornel MC. 2008. [Orphanet: a European database for rare diseases]. *Nederlands tijdschrift voor geneeskunde* 152:518–519.

Westergaard D, Stærfeldt H-H, Tønsberg C, Jensen LJ, Brunak S. 2018. A comprehensive and quantitative comparison of text-mining in 15 million full-text articles versus their corresponding abstracts. *PLOS Computational Biology* 14:e1005962. DOI: 10.1371/journal.pcbi.1005962.

WHO C. 1946. WHO remains firmly committed to the principles set out in the preamble to the Constitution. Available at <https://www.who.int/about/who-we-are/constitution> (accessed April 24, 2019).

Williamson PO, Minter CIJ. 2019. Exploring PubMed as a reliable resource for scholarly communications services. *Journal of the Medical Library Association: JMLA* 107:16–29. DOI: 10.5195/jmla.2019.433.

Wilson EB. 1927. Probable Inference, the Law of Succession, and Statistical Inference. *Journal of the American Statistical Association* 22:209–212. DOI: 10.2307/2276774.

Wishart DS, Knox C, Guo AC, Shrivastava S, Hassanali M, Stothard P, Chang Z, Woolsey J. 2006. DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Research* 34:D668–D672. DOI: 10.1093/nar/gkj067.

World Health Organization. 1993. *The ICD-10 classification of mental and behavioural disorders: diagnostic criteria for research*. Geneva: Geneva: World Health Organization.

Wu CH, Apweiler R, Bairoch A, Natale DA, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, Magrane M, Martin MJ, Mazumder R, O'Donovan C, Redaschi N, Suzek B. 2006. The Universal Protein Resource (UniProt): an expanding universe of protein information. *Nucleic Acids Research* 34:D187–D191. DOI: 10.1093/nar/gkj161.

Xia E, Sun W, Mei J, Xu E, Wang K, Qin Y. 2018. Mining Disease-Symptom Relation from Massive Biomedical Literature and Its Application in Severe Disease Diagnosis. *AMIA Annual Symposium Proceedings* 2018:1118–1126.

Xu R, Li L, Wang Q. 2013. Towards building a disease-phenotype knowledge base: extracting disease-manifestation relationship from literature. *Bioinformatics (Oxford, England)* 29:2186–2194. DOI: 10.1093/bioinformatics/btt359.

Yang H, Lee HJ. 2018. Research Trend Visualization by MeSH Terms from PubMed. *International Journal of Environmental Research and Public Health* 15. DOI: 10.3390/ijerph15061113.

Yang K, Wang R, Liu G, Shu Z, Wang N, Zhang R, Yu J, Chen J, Li X, Zhou X. 2019. HerGePred: Heterogeneous Network Embedding Representation for Disease Gene Prediction. *IEEE Journal of Biomedical and Health Informatics* 23:1805–1815. DOI: 10.1109/JBHI.2018.2870728.

Capítulo 7 Bibliografía

Zhang M, Zhu C, Jacomy A, Lu LJ, Jegga AG. 2011. The Orphan Disease Networks. *The American Journal of Human Genetics* 88:755–766. DOI: 10.1016/j.ajhg.2011.05.006.

Zhou X, Li Y, Peng Y, Hu J, Zhang R, He L, Wang Y, Jiang L, Yan S, Li P, Xie Q, Liu B. 2014a. Clinical phenotype network: the underlying mechanism for personalized diagnosis and treatment of traditional Chinese medicine. *Frontiers of Medicine* 8:337–346. DOI: 10.1007/s11684-014-0349-8.

Zhou X, Menche J, Barabási A-L, Sharma A. 2014b. Human symptoms-disease network. *Nature Communications* 5:4212. DOI: 10.1038/ncomms5212.

Zhou X, Menche J, Barabási A-L, Sharma A. 2014c. Human symptoms-disease network. *Nature Communications* 5:4212. DOI: 10.1038/ncomms5212.

Zou D, Ma L, Yu J, Zhang Z. 2015. Biological Databases for Human Research. *Genomics, Proteomics & Bioinformatics* 13:55–63. DOI: 10.1016/j.gpb.2015.01.006.

Publicaciones

7.1 Publicaciones en congresos

Lagunes García G, Prieto Santamaría L, García del Valle EP, Zanin M, Menasalvas Ruiz E, Rodríguez González A. 2019. Wikipedia Disease Articles: An Analysis of their Content and Evolution. In: *2019 IEEE 32nd International Symposium on Computer-Based Medical Systems (CBMS)*. 664–671. DOI: 10.1109/CBMS.2019.00136.

Valle EPG del, García GL, Ruiz EM, Santamaría LP, Zanin M, Rodríguez-González A. 2019. Completing Missing MeSH Code Mappings in UMLS Through Alternative Expert-Curated Sources. In: *2019 IEEE 32nd International Symposium on Computer-Based Medical Systems (CBMS)*. 174–179. DOI: 10.1109/CBMS.2019.00044.

Valle EPG del, García GL, Santamaría LP, Zanin M, Ruiz EM, González AR. 2018. Evaluating Wikipedia as a Source of Information for Disease Understanding. In: *2018 IEEE 31st International Symposium on Computer-Based Medical Systems (CBMS)*. 399–404. DOI: 10.1109/CBMS.2018.00076.

7.2 Publicaciones en revistas de alto impacto

Lagunes-García G, Rodríguez-González A, Prieto-Santamaría L, Valle EPG del, Zanin M, Menasalvas-Ruiz E. 2020b. DISNET: a framework for extracting phenotypic disease information from public sources. *PeerJ* 8:e8580. DOI: 10.7717/peerj.8580. (Factor de impacto en 2018: 2.35 - Q1 en *Agricultural and Biological Sciences (miscellaneous)*).

Lagunes-García G, Rodríguez-González A, Prieto-Santamaría L, García del Valle EP, Zanin M, Menasalvas-Ruiz E. 2020a. How Wikipedia disease information evolve over time? An analysis of disease-based articles changes. *Information Processing & Management* 57:102225. DOI: 10.1016/j.ipm.2020.102225. (Factor de impacto en 2018: 3.892 - Q1 en *Computer Science Applications & Information Systems*).

García del Valle EP, Lagunes García G, Prieto Santamaría L, Zanin M, Menasalvas Ruiz E, Rodríguez-González A. 2019. Disease networks and their contribution to disease understanding: A review of their evolution, techniques and data sources. *Journal of Biomedical Informatics* 94:103206. DOI: 10.1016/j.jbi.2019.103206. (Factor de impacto en 2018: 2.950 - Q1 en *Computer Science Applications & Health Informatics*).