UNIVERSITAT POLITÈCNICA DE VALÈNCIA

etsinf
Escola Tècnica Superior d'Enginyeria Informàtica

Escola Tècnica Superior d'Enginyeria Informàtica

Universitat Politècnica de València

Análisis comparativo del comportamiento de diferentes motores de búsqueda en el tratamiento de la investigación sobre Enfermedades Raras.

Trabajo Fin de Máster

Máster Universitario en Gestión de la Información

Autor: Cristina I. Font Julián

Tutor: José Antonio Ontalba y Ruipérez

Tutor: Enrique Orduña Malea

Curso Académico: 2015 - 2016

“La ciencia más útil es aquella cuyo fruto es el más comunicable”

Leonardo Da Vinci (1452-1519)

Agradecimientos

A mis padres y mi hermana, porque sin ellos no soy nadie.

A Lidia, Mari, Xaume y los *Informáticos*, por soportarme y ayudarme en todo momento.

A José Antonio y Enrique, por confiar y creer en mi incluso cuando yo no lo hago.

Resumen

Las Enfermedades Raras son aquellas que afectan a una pequeña proporción de la población, con los consiguientes problemas de investigación y financiación que existen. Para dar visibilidad a las mismas en internet, se quiere conocer el tratamiento que dan los buscadores a la información que se encuentra online.

Las búsquedas en Internet son esenciales para poder encontrar información, debido a esto se plantea conocer la visibilidad y presencia de los portales de asociaciones de Enfermedades Raras en la Web mediante los motores de búsqueda., analizando el sesgo que estos aplican sobre la información relacionada con Enfermedades Raras.

Para ello, se realiza un análisis acerca del interés sobre Enfermedades Raras, listando todas las asociaciones relacionadas extraídas de diversas fuentes de información, y elaborando un directorio unificado con 438 entradas. Se selecciona un top 50 de enfermedades y, tras cruzar los datos con el listado de asociaciones, se escogen cien distintas.

Se realiza la extracción de datos relativa a diversos indicadores para cada una de ellas, empleando varios programas creados para este proyecto, que automáticamente recogen los datos de los dos buscadores analizados, Google y Bing. Finalmente, se procesan y analizan para conocer el tratamiento que dan los motores de búsqueda a la información existente sobre Enfermedades Raras.

De los 100 portales analizados, únicamente el 25% obtiene buenos resultados. El impacto en general no es bueno, de media se sitúa en 5,25 puntos según el Factor de Impacto en la Web (WIF). Debido a la correlación que existe entre los diferentes resultados obtenidos, Google es el más indicado para encontrar información relativa a Enfermedades Raras. En función de los resultados obtenidos, se concluye que la presencia y visibilidad en la Web de las Enfermedades Raras debería ser mejorada.

Palabras clave: motores de búsqueda, enfermedades raras, visibilidad web.

Abstract

Rare Diseases are those who affect a small portion of the population, with the related problems in research and financial problems that this has. To give visibility to them on the Internet, it’s necessary to know how the search engines manage the online information.

Internet searches are essential to find information, because of this, knowing about the visibility and presence of Rare Diseases association’s portals in the web through search engines is required, analysing the bias applied by them to the Rare Diseases information.

To do this, an analysis of interest on Rare Diseases is made, listing all related associations drawn from various sources of information, and developed an unified directory with 438

entries. Selected a top 50 diseases and, after crossing the data with the list of associations, a hundred associations are chosen.

Data extraction is performed on several indicators, using different programs developed for this project, which automatically collect data from two search engines, Google and Bing. Finally, data is processed and analysed to determine how the search engines manage rare diseases information.

Of the hundred portals analysed, only 25% has good results. The overall impact is not good, the average stands at 5,25 points according to Web Impact Factor (WIF). Regarding the correlation between the results, Google is the best suited to find information on Rare Diseases. Depending on the results, it is concluded that the presence and visibility on the Web of Rare Disesase should be improved.

Keywords : search engines, rare diseases, web visibility.

Resumen

Les Malalties Rares són aquelles que afecten una menuda proporció de la població, amb els consegüents problemes d'investigació i finançament que existeixen. Per a donar visibilitat a les mateixes en Internet, es vol conéixer el tractament que donen els buscadors a la informació que es troba online.

Les busques en Internet són essencials per a trobar informació, per açò es planteja conéixer la visibilitat i presència dels portals d'associacions de Malalties Rares en la Web per mitjà dels motors de busca, analitzant el biaix que apliquen a la informació relacionada amb Malalties Rares.

Per a això, es realitza una anàlisi sobre l'interès sobre Malalties Rares, llistant totes les associacions relacionades extretes de diverses fonts d'informació, i elaborant un directori unificat amb 438 entrades. Es selecciona un top 50 de malalties i, creuant les dades amb el llistat d'associacions, es trien cent distintes. Es realitza l'extracció de dades de diversos indicadors per a cadascuna d'elles, utilitzant diversos programes creats per a este projecte, que automàticament arrepleguen les dades dels dos buscadors analitzats, Google i Bing.

Finalment, es processen i analitzen per a conéixer el tractament que donen els motors de busca a la informació sobre Malalties Rares. Dels 100 portals analitzats, únicament el 25% obté bons resultats. L'impacte en general no és bo, de mitja es situa en 5,25 punts segons el Factor d'Impacte en la Web (WIF). A causa de la correlació que existeix entre els diferents resultats obtinguts, Google és el més indicat per a trobar informació relativa a Malalties Rares. En funció dels resultats obtinguts, es conclou que la presència i visibilitat en la Web de les Malalties Rares hauria de ser millorada.

Paraules clau: motors de busca, malalties rares, visibilitat web.

Índice de contenidos

Índice de tablas

Índice de figuras

1 Introducción

Según la Real Academia de la lengua española *"raro"* significa por definición:

1. Que se comporta de modo inhabitual.
2. Que es extraordinario, poco común o frecuente.
3. Escaso en su clase o especie.
4. Insigne, sobresaliente o excelente en su línea.
5. Extravagante de genio o de comportamiento y propenso a singularizarse.
6. Dicho principalmente de un gas enrarecido: que tiene poca densidad inconsistencia.

Y es precisamente el significado de los tres primeros puntos, tanto unidos como por separado, la clave para entender el porqué del nombre de las enfermedades raras.

Las enfermedades raras son aquellas que afectan a una pequeña parte de la población y cuyo origen y volatilidad imposibilita el adecuado tratamiento de la misma (Orpha.net, 2012). Dependiendo del país existen matices en la definición por prevalencia de enfermedad rara. Por ejemplo, en Europa se considera enfermedad rara cuando afecta a 1 persona de cada 2.000 mientras en los Estados Unidos lo es si existen menos de 200.000 personas afectadas (Unidos, 2002).

EURORDIS, la Organización Europea para las Enfermedades Raras, estima la existencia de entre 6.000 y 8.000 enfermedades raras, que en general son de tipo crónico, invalidantes y cuyo origen, en más de un 80% de los casos, es genético (EURORDIS, 2012).

Pese a tratarse de un número relativamente alto, debido a la pequeña incidencia en la población, generalmente son desconocidas para el grueso de la ciudadanía y la información disponible en la web es limitada y se encuentra poco accesible, dificultando el conocimiento general y el uso de la misma entre pacientes y personal médico-investigador en particular.

Debido a la poca visibilidad de las enfermedades raras en el conjunto de la población es necesario aplicar a la información que pueda ofrecerse de las mismas la máxima ayuda posible para que todo el mundo pueda llegar fácilmente a ella. Dado que internet juega un papel importante en el proceso de búsqueda de información, se pretende utilizar este trabajo para conocer la presencia y visibilidad de las enfermedades raras en la web para poder ayudar a informar sobre las mismas promocionando y potenciando la visibilidad de la información relacionada con el campo.

1.1 Justificación

Según la Organización Mundial de la Salud (Salud, 2012), cerca de un 8% de la población mundial se encuentra afectada por enfermedades raras, en España la cifra supera los tres millones de afectados. Esto implica una serie de problemas:

- Complica el diagnóstico correcto en primera instancia.
- Dificulta el día a día de los pacientes y su entorno.
- Al existir una prevalencia mínima se dificulta el desarrollo de estrategias y productos terapéuticos.
- El acceso a la información es limitado, ya sea relacionada con la propia enfermedad como con lugares de ayuda.

Dos de estos puntos problemáticos hacen referencia a la falta de información en diferentes ámbitos, para el paciente y su entorno o para el colectivo médico. Desde los diferentes organismos internacionales se ha propuesto una base de Redes de Referencia que permita ofrecer información de calidad a los afectados y sus familias, así como ayudar a los profesionales y los centros de referencia a intercambiar conocimientos (Europea, Redes de Referencia - CE, 2016).

Del mismo modo existen asociaciones, federaciones, centros de investigación o institutos, tanto nacionales como internacionales, que ofrecen información dirigida a los dos colectivos, pero que es necesario conocer en primera instancia para poder acceder a la información.

A esto debe añadirse que no todas las enfermedades raras reciben la misma atención. Por ejemplo existen investigaciones que son motivadas por el interés del investigador, investigaciones que se inician debido al mecenazgo de una

asociación o particular que indica la enfermedad a investigar, e investigaciones que son propiciadas gracias a la publicidad que se ofrece desde algún colectivo.

Un ejemplo claro de este ultimo punto es la campaña viral "El cubo helado" [1]qu e se realizó en 2014 para lograr recaudar fondos para la investigación de la Esclerosis Lateral Amiotrófica (ELA). Esta campaña tenia como fin dos objetivos, concienciar a la población acerca de la enfermedad y recaudar fondos para su investigación de un modo llamativo: los participantes debían pasar el testigo a tres personas para que estos realizasen el reto, propiciando una reacción en cadena, y mojarse con un cubo de agua helada o realizar una donación a alguna asociación relacionada con la enfermedad.

Gracias a la participación de personas famosas y lo llamativo de la campaña se lograron recaudar solo en España más de 700.000 euros (World, 2015), haciendo visible la enfermedad y dejando patente la falta de recursos económicos en relación con las enfermedades raras.

Debido a que no todas las enfermedades tienen una oportunidad similar para hacerse visibles, es necesario que se disponga de toda la ayuda posible en cualquier medio para dar la mayor difusión posible a la información relacionada con las enfermedades raras.

Según los datos del Instituto Nacional de Estadística en relación a la evolución del uso de las TIC por personas de entre 16 y 74 años (INE, 2014) la tendencia de uso de las mismas se encuentra al alza. Como se puede observar en la Figura 1 cada día más usuarios se conectan con mayor frecuencia a la red.

Según la encuesta publicada por el Observatorio Nacional de las Telecomunicaciones y la Sociedad de la Información en abril del presente año 2016 (ONTSI, 2016), la cantidad de usuarios que utiliza Internet para buscar información relacionada con la salud aumenta proporcionalmente año tras año, siendo actualmente el 60,5% de la población.

[1] http://adelaweb.org/campana-cubo-de-agua-helada

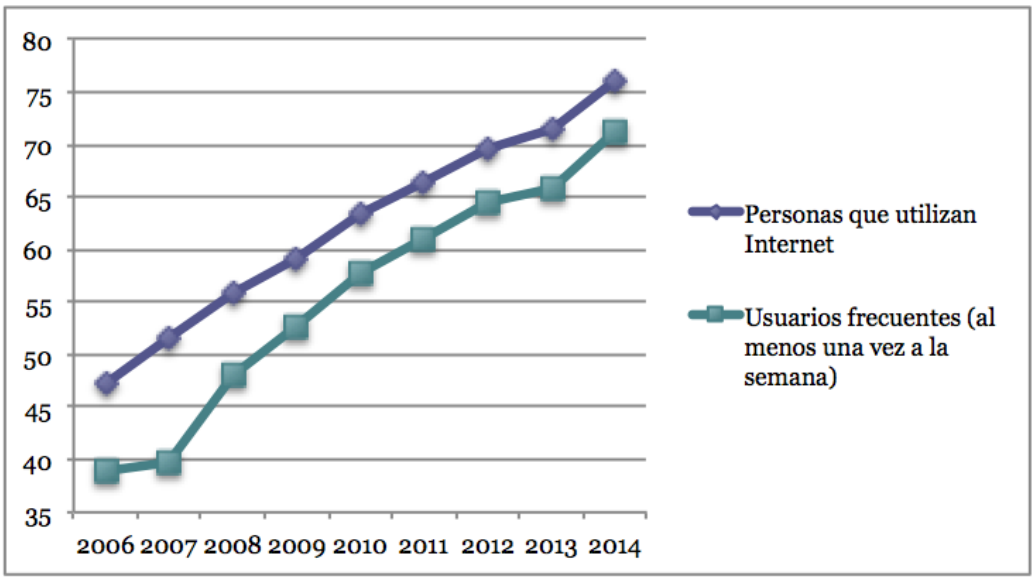


Figura 1: Evolución uso de las TIC por personas de entre 16 y 74 años (INE, 2014)

El uso de las tecnologías de la información y la comunicación se convierte en la tercera fuente de información relacionada con la salud para los usuarios (ONTSI, 2016), siendo las dos primeras los profesionales médicos y farmacéuticos. Las búsquedas más importantes realizadas en el medio se encuentran relacionadas con:

- Nutrición, alimentación y estilo de vida saludable.
- Diagnóstico o enfermedad.
- Síntomas para identificar una dolencia o patología.
- Medicamentos recetados.

De este modo la Web se convierte en la plataforma principal para la difusión de información relacionada con la salud. Por ello, se realizará un análisis de interés por las Enfermedades Raras en la Web, también se buscará la cantidad de información que existe en la web de cada una de ellas.

Debido a la evolución natural a la que se someten todas las cosas en el universo, la Web ha experimentado una serie de cambios desde su invención en 1989 por Tim Berners-Lee. En la versión actual de la web se enfatiza el contenido centrado en el usuario, la usabilidad y la interoperabilidad y cuando esto se aplica al cuidado de la salud se denomina Salud 2.0 (Belt, 2010).

Sus usos son variados, pueden ir desde mantener informado sobre algún campo en particular, pasando por educación médica o colaboración en

investigaciones tanto por profesionales, como por pacientes desde cualquier parte del mundo. Este ultimo punto en el caso de las enfermedades raras supone el fin de una barrera mayor, ya que debido a la baja incidencia en la población, en muchas ocasiones es difícil encontrar la muestra necesaria para llevar a cabo investigaciones o extraer conclusiones acerca de tratamientos.

Dado que los motores de búsqueda son la puerta principal de entrada a un mundo virtual de información, y son utilizados para la extracción de datos en Internet, es imprescindible analizar el tratamiento que dan a la información que existe en la red relacionada con enfermedades raras.

El funcionamiento interno de los buscadores es estudiado y analizado en multitud de campos y durante mucho tiempo con mayor o menor éxito. Así como las posibilidades de posicionamiento en las páginas de resultados de los motores de búsqueda (SERPs) mediante la disciplina de Optimización para Motores de Búsqueda (en inglés SEO).

Pero este tipo de análisis no trata de buscar una solución o explicación a los posibles problemas, como no encontrarse en los primeros resultados de búsqueda, que existan relacionados con el posicionamiento de una temática concreta, únicamente tratan de solventar la dificultad de visibilidad aplicando unas técnicas preexistentes.

Además, los motores de búsqueda no ofrecen toda la información de un modo fácil o sencillo de extraer, actualmente Bing ofrece una API, pero limita su uso mensual gratuito. Google no ofrece ninguna API de resultados de búsqueda, por lo que es más difícil recopilar la información. A esto se debe añadir que los resultados que ofrecen no son completos, por lo que supone un grado añadido de dificultad el poder recopilar correctamente, y de un modo sencillo, la información relacionada con las métricas que permiten conocer la presencia y visibilidad web de las Enfermedades Raras.

Para poder saber cual es la presencia de las páginas relacionadas con las Enfermedades Raras, se centra el foco del análisis en la cobertura que los motores de búsqueda tienen sobre las asociaciones de pacientes y familiares, por ser estos portales los que más información contienen generalmente. De este modo se puede medir cuantitativamente la información que existe sobre el tema propuesto.

Para poder conocer esta presencia, primero se debe localizar las Asociaciones sobre las que se realiza el análisis, para ello será necesario buscar un listado de asociaciones por enfermedad, para después cruzarlo con la información

obtenida anteriormente y localizar las asociaciones relacionadas con las enfermedades que más interesan.

Del mismo modo, se pretende realizar una comparativa entre los buscadores propuestos y, tras analizar los resultados obtenidos, se espera poder encontrar un nuevo sistema que permita ejecutar los procesos de medición de un modo más sencillo y fiable, incorporando diferentes técnicas e indicadores para realizar las mediciones.

En última instancia, se espera poder diseñar un método de investigación que permita cuantificar los diferentes portales dedicados a Enfermedades Raras para lograr obtener la mayor información posible acerca de las mismas.

Con la realización de este Proyecto Final de Máster se espera poder entender mejor el comportamiento de los motores de búsqueda más allá de la literatura existente en la actualidad, y poder ayudar a su vez a que las Enfermedades Raras tengan la visibilidad y presencia en la web ideal para poder llegar, no sólo a pacientes y profesionales médicos, sino a toda la sociedad.

1.2 Objetivos

De acuerdo con la justificación planteada anteriormente, el presente trabajo de fin de máster tiene como objetivo principal analizar la presencia y visibilidad web de los portales relacionados con enfermedades raras mediante la cuantificación de diferencias en el tratamiento de los portales por parte de diferentes motores de búsqueda.

Los objetivos concretos que se plantean son las siguientes etapas:

- Medir y analizar la cantidad de información existente actualmente sobre las enfermedades raras en internet para conocer el interés que suscitan.

- Estudiar y seleccionar los dominios relacionados con el tema necesarios para realizar el estudio.

- Diseñar y programar arañas específicas para la obtención de indicadores métricos de forma automática, precisa y gratuita.

- Medir el tamaño de los dominios, incluyendo los diferentes tipos de documentos e información, para conocer y contextualizar su presencia en la Web en función de diversos buscadores.

- Analizar el posicionamiento y tratamiento de los resultados relacionados con un tema concreto.

- Proponer un modelo de análisis mediante cuantificación para diferentes buscadores web.

- Elaborar un directorio compuesto por los resultados obtenidos en materia de dominios relacionados con el tema.

1.3 Avance metodológico

Utilizando un proceso inductivo, se explorará y analizará la problemática de visibilidad en los diferentes buscadores. Se logra, a su vez, obtener diferentes perspectivas sobre la investigación en función de los resultados obtenidos.

La investigación será de tipo descriptiva, ya que se tratará de especificar las propiedades y características analizadas, para poder indicar tendencias o líneas a seguir dentro del posicionamiento y visibilidad web.

La metodología que se aplica en este trabajo consta de los siguientes puntos:

- Extraer una muestra significativa de los dominios a analizar.

- Analizar métodos, procesos y herramientas para la extracción de información de la manera más rápida y precisa posible.

- Generar un procedimiento de captura de datos, componiendo o configurando los scripts o herramientas necesarias para la obtención de los datos que conformarán el grueso de información del trabajo.

- Procesar los datos obtenidos para su correcto manejo y posterior análisis.

- Presentación e interpretación de los datos, realizando un análisis de los mismos para extraer el conocimiento que puedan aportar.

El enfoque de este trabajo es puramente cuantitativo, las preguntas que se deben responder se encuentran delimitadas en los objetivos y todas ellas se pueden responder mediante respuestas muy concretas.

Se escoge este enfoque ya que mediante la recolección de datos se pretende conocer y medir diferentes puntos en la investigación para poder encontrar soluciones. Algunas de las preguntas que se pretende poder resolver son:

- ¿Cómo de distintos son los resultados en función de los buscadores, el idioma o la localización?
- ¿Qué tipos de documentos se encuentran indexados?
- ¿Cuál es la profundidad de indexación por buscador?
- ¿Existen similitudes o casos únicos destacables entre los resultados?

1.4 Relación asignaturas

Este Trabajo Fin de Máster corresponde al proyecto final del Máster Universitario en Gestión de la Información de la Universidad Politécnica de Valencia. Se encuentra estrechamente relacionado con las asignaturas que se imparten en el titulo siguientes:

- *Analítica web y cibermetría*: sienta las bases principales del camino que sigue el proyecto. Permite conocer las métricas, indicadores, proyectos y tecnologías de búsqueda que explican el entorno web y su funcionamiento.
- *Técnicas de investigación e innovación*: permite conocer las herramientas metodológicas fundamentales que guían el proyecto, las técnicas de medición estudiadas y el análisis de datos visto han sido los pilares fundamentales del mismo.
- *SEO y SEM*: la asignatura permite conocer el funcionamiento interno de los buscadores y el sistema de posicionamiento mediante SEO, se estudian las diferentes técnicas aplicables que permiten lograr los

resultados de posicionamiento esperados y que ayudan a desarrollar este proyecto.

- *Explotación de datos masivos*: debido a la gran cantidad de datos que se manejan a lo largo del proyecto es necesario contar con esta asignatura, la cual permite dar un tratamiento, gestión y análisis de los datos recogidos.

- *Fuentes de datos e información*: la asignatura permite abrir el campo de conocimiento de lugares de los cuales servirse para extraer la información relacionada con el proyecto, así como poder entender el tipo de información que está siendo recolectada.

- *Sociedad de la información*: en la asignatura se presentaron las Tecnologías de la Información y su entorno, permitiendo conocer los servicios y productos que conforman la plataforma que se analizan en el proyecto.

- *Almacenamiento y recuperación de información*: esta asignatura permite conocer las tecnologías de almacenamiento, procesado y recuperación de grandes volúmenes de datos que se utilizan en el trabajo.

- *Gestión de datos: web semántica y open data*: la asignatura permite conocer los diferentes sistemas de apertura de datos e información que serán aplicados en este proyecto.

- *Información multimedia en entornos multidispositivo*: en la asignatura se trabajan métodos de publicación de datos en la web que serán útiles para la finalización del trabajo.

1.5 Estructura de la memoria

El presente trabajo se estructura en seis apartados cuyo contenido se explica a continuación:

El capítulo 2 se dedica al estado de la cuestión, exponiendo el marco teórico en el cual se encuentra el trabajo desarrollado. Se pretende poder ofrecer una

visión del conjunto de aspectos que conforman la base para entender el fin último del trabajo.

En el tercer capítulo se detalla la metodología seguida, describiendo y especificando los detalles completos de la sistemática aplicada durante la realización del estudio. Se propone el modelo de análisis a utilizar y se describen los procesos de recogida de datos.

En el cuarto capítulo se realiza una exposición de los resultados obtenidos, efectuando una discusión de los mismos, presentándolos y explicándolos aportando una descripción detallada de la información recopilada en el capitulo anterior.

En el quinto capítulo se exponen las conclusiones y se describe el trabajo futuro, presentando las conclusiones extraídas del presente trabajo e indicando posibles trabajos futuros que surgen de la problemática encontrada en el transcurso del trabajo. El trabajo finaliza con el sexto capitulo, en el cual se detalla la bibliografía utilizada.

Finalmente, en el Anexo I se puede encontrar el listado completo con la información de las 438 asociaciones listadas, que incluye siglas, nombre completo y URL de cada una de ellas.

2 Estado de la cuestión

Tal y como se indica en la introducción, el presente trabajo busca ayudar a conocer la visibilidad de las enfermedades raras en internet. Para ello es necesario presentar el estado en que se encuentra el marco teórico y conjunto de herramientas que permitirán realizar el análisis final.

2.1 Objeto de análisis: las asociaciones.

No es necesario que una enfermedad sea rara para que ésta genere una serie de problemas, preocupaciones, desconfianza y multitud de sentimientos más tanto al enfermo como a su entorno. Pero en el caso de las enfermedades raras existen una serie de componentes que hacen que el sobrellevar las mismas sea más difícil que en el caso de una enfermedad de tipo más común, y la principal de ellas es la falta de información.

En el momento en que se diagnostica a un paciente con una enfermedad rara, los profesionales médicos son capaces de aportar el grueso de información que éste necesitará en ese momento, realizando las explicaciones necesarias y aportando conocimiento de un nuevo modo de vida. Tras esto, el paciente y sus familiares pueden pasar por diferentes etapas y modos de sobrellevarlo, pero llegará un momento en el que precisen respuestas y una mayor cantidad de información (Salcedo, 2011), por ello recurrirán a la siguiente fuente de información sobre salud: Internet (ONTSI, 2016).

Desde diferentes organismos, tanto nacionales como internacionales, públicos o privados, se trata de promover e incentivar el acceso a la información relacionada con las enfermedades raras a través de la Web y por ello se llevan a cabo diferentes proyectos o iniciativas con diversos grados de apertura de información. Los principales ejemplos son:

- *EURORDIS*[2] : se trata de una Red Europea de Referencia creada por la Comisión Europea y cuyo objetivo principal es permitir a los profesionales y los centros de referencia de distintos países intercambiar conocimientos (Europea, Comisión Europea - Salud Pública, 2016) pero que también ofrece información útil para los enfermos ya

[2] http://www.eurordis.org

que sus dirigentes son pacientes. Se encarga de representar a 716 organizaciones de pacientes de 63 países y sus principales deberes son:

- Aplicar criterios europeos a las enfermedades raras que requieren un tratamiento especializado.
- Servir de centros de investigación y conocimiento para el tratamiento de pacientes de distintos países de la UE.
- Garantizar la disponibilidad de los tratamientos necesarios.

- *Orpha.net*[3]: es un portal de información de referencia en enfermedades raras y medicamentos huérfanos, dirigido a todos los públicos. Su objetivo principal es contribuir a la mejora del diagnostico, cuidado y tratamiento de los pacientes con enfermedades raras (Orpha.net, 2012). Está formado por un consorcio de 40 países, coordinados por el equipo francés del Instituto Nacional Francés de la Salud y de la Investigación Médica, el cual se encarga de financiar el proyecto junto con otros socios.

- *Federación Española de Enfermedades Raras (FEDER)*[4]: se trata de una entidad de utilidad pública cuya misión es dar voz a las personas con enfermedades poco frecuentes y sus familias (FEDER, 2016). Desde la federación trabajan por crear una red sólida de asociaciones en toda España para mejorar la calidad de vida de los afectados, defendiendo sus derechos y promoviendo su visibilidad en el conjunto de la población.

- *Ciberer*[5]: se trata de un área temática creada por el Centro de Investigación Biomédica en Red, para servir de referencia, coordinar y potenciar la investigación sobre las enfermedades raras en España, formada por 62 grupos de investigación ligados a 29 instituciones consorciadas (Ciberer, 2016). Desde esta plataforma se puede acceder a una gran cantidad de información sobre investigación relacionada con las Enfermedades Raras, siendo una de las herramientas más interesantes Maper:

[3] http://www.orpha.net
[4] http://www.enfermedades-raras.org
[5] http://www.ciberer.es/

- *Maper*[6] es un mapa interactivo de proyectos en ER investigadas en España. De este modo se puede disponer de una base de datos actualizada y fiable de las enfermedades que se están investigando activamente.

En el caso de EURORDIS, Orpha.net y FEDER son portales que están centrados en información de nivel alto, es decir son similares a grandes bases de datos que contienen un gran volumen de información sobre enfermedades raras en general, pero no un gran volumen sobre ninguna en particular. Si bien es cierto que apuntan a portales desde los que se puede recabar más información; aunque aplicando un sesgo, ya que debe existir una relación de *asociación* entre las dos fuentes, por lo que si no son colaboradores no aparecerán en los registros.

En estas tres herramientas se puede encontrar información sobre actualidad relacionada con el campo, descripciones de las enfermedades, enlaces a algunas asociaciones, testimonios e información sobre medicamentos o investigaciones.

Así mismo cabe señalar la existencia del Registro Nacional de Enfermedades Raras del Instituto de Salud Carlos III, en el que se dispone de un buscador con un listado de las Enfermedades Raras en España, aunque únicamente lista las enfermedades y en caso de que el usuario requiera mayor información, se le redirige a Orphanet.

Más allá de los organismos mencionados, existen asociaciones, federaciones o fundaciones relacionados con las enfermedades raras. Como se indica anteriormente, algunos de ellos pueden encontrarse a través de las listas directorio que aparecen en algunos de los portales vistos, junto con la enfermedad o enfermedades a los que se asocian. Otros, en cambio, deben ser encontrados a través de búsquedas en internet u otras personas conocedoras de la existencia del organismo.

Para la realización del trabajo se han listado un total de 438 asociaciones relacionadas con una o más enfermedades raras en España (en el Capitulo 3 se verá en profundidad el método utilizado para alcanzar esta cifra). Esta cantidad no es accesible de un modo sencillo y son necesarias horas de investigación y trabajo para llegar a él, este es uno de los motivos que sustentan el presente trabajo. Es sobre este conjunto de asociaciones sobre el

[6] http://www.ciberer-maper.es/

que se realiza el estudio y análisis, centrado en su visibilidad y presencia en la web.

2.2 Técnica de análisis

2.2.1 Cibermetría

La primera vez que el prefijo *ciber-* tuvo un significado relacionado con el espacio electrónico, fue de la mano de William Gibson tras la publicación de varias novelas (Sen, 2004). Con el paso de los años el término evoluciona y actualmente se relaciona con las tecnologías de la información, Internet y la realidad virtual.

El sufijo *-metría* significa "*medida*" o "*medición*", en este caso, del *ciber*espacio. Por lo tanto en el caso de la cibermetría nos indica, que la disciplina se fundamenta en la cuantificación, y posterior análisis, de los objetos de estudio encontrados en el espacio virtual que conforma la red (Orduña-Malea & Aguillo, 2014).

La cibermetría aplica los principios de la *Cienciometría*, *Bibliometría* e *Infometría* al estudio de la Web, su similitud con las áreas es de solapamiento debido a que comparten un mismo enfoque, el estudio cuantitativo sobre la información (Björneborn, 2004).

En la Figura 2 se puede ver la relación entre las disciplinas indicadas anteriormente junto con la *Webmetría*. Para poder comprender mejor la extensión de la Cibermetría, se detallan las definición es del resto de disciplina:

- *Infometría*: estudio cuantitativo de la producción, almacenamiento, recuperación, diseminación y utilización de la información (Wolfram, 2000). Esta disciplina investiga la existencia de regularidades empíricas en estas actividades y los intentos de desarrollar modelos matemáticos y teorías que permitan comprender los procesos de información.

- *Bibliometría*: estudio estadístico sobre publicaciones realizadas, como libros o artículos. Éste análisis utiliza datos relacionados con los autores y las publicaciones científicas, y con los artículos y las citas que reciben, para calcular el impacto y visibilidad tanto de individuos como de

equipos de investigación, instituciones y países, así como sus trabajos. Permite además, identificar redes nacionales e internacionales, y localizar el desarrollo en ciencia y tecnología (Portal, 2001).

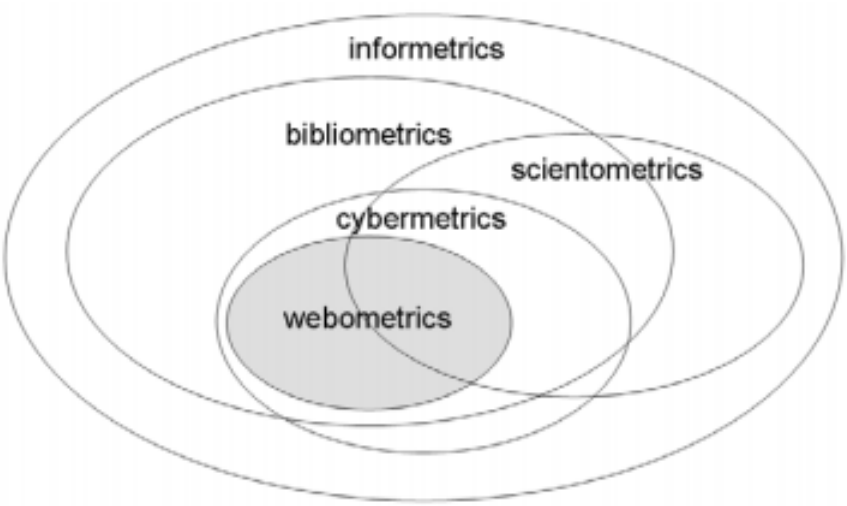


Figura 2: Relación entre disciplinas (Björnesson, 2004)

Tabla 1: Resumen Actividades de investigación en función de Especialidad.

Infometría	Actividad informativa
Bibliometría	Actividad bibliográfica
Cienciometría	Actividad de investigación
Webmetría	Actividad en la Web

- *Cienciometría*: estudio que permite la métrica y análisis de la comunicación en ciencia, tecnología e innovación (Leydesdorff, 2012). Según este autor, la unidad de análisis principal de la técnica son los documentos. Los métodos que se utilizan en este campo son cualitativos, cuantitativos y enfoques computacionales.

- *Webmetría*: estudio cuantitativo del *World Wide Web*, que analiza el número y tipos de enlaces o *hiperlinks*, la estructura de la Web y su uso. Este estudio se realiza utilizando un acercamiento bibliométrico e infométrico para cuantificar los aspectos constructivos y de uso de las fuentes de información en la Web, sus estructuras y tecnologías (Björneborn, 2004).

En la Tabla 2 se detalla, por disciplina y de modo esquemático, los objetos de estudio de cada una de ellas, ejemplos de varibles utilizadas para su análisis, métodos aplicables en función del análisis y los objetivos:

Tabla 2: Adaptación de la tipología para la definición y clasificación de las distintas disciplinas (Macías-Chapula, 2001).

Disciplina	Objeto de estudio	Ejemplos Variables	Métodos	Objetivos
Infometría	Palabras, documentos, bases de datos.	Métricas para recuperación, relevancia, recordatorio.	Modelos vector-espacio, Probabilísticos, lenguajes de procesamiento.	Aumentar eficiencia de la recuperación.
Bibliometría	Libros, documentos, revistas, artículos, autores y usuarios.	Citas, frecuencia de aparición de palabras, longitud de oraciones.	Clasificación, frecuencia, distribución.	Asignar recursos.
Cienciometría	Disciplinas, materias, campos.	Aspectos diferenciadores de las disciplinas.	Análisis de conjunto y de correspondencia.	Identificar nodos de interés, estado de comunicaciones.
Cibermetría	Recursos de información, estructuras y tecnologías en Internet.	Herramientas de búsqueda, revistas, autores, formatos.	Clasificación, frecuencia, distribución, modelos estadísticos.	Cuantificar la información electrónica en la Red.
Webmetría	Aspectos cuantitativos de la construcción y uso de la Información.	Hosts, servidores Web, usuarios, dominios.	Técnicas bibliométricas para relacionar diferentes sitios de la Web.	Analizar los componentes de la Web.

En resumen, estas las disciplinas se centran en el estudio cuantitativo de ciertos componentes, diferenciándose estos en función de la rama y alcance: la información en cualquiera de sus formas y tipos en el caso de la Infometría; la producción, diseminación y uso de la información científica en el caso de la

Bibliometría; la ciencia como disciplina en el caso de la Cienciometría y la web a través de los enlaces en el caso de la Webmetría.

Para poder comprender mejor las aplicaciones de la Cibermetría, a continuación se detallan algunos ejemplos sobre las mismas en su estudio aplicado a la Red (Rodriguez, 2006):

- Número, alcance y temas de las redes de información.
- La distribución existente entre las redes por países.
- El volumen y tipología de las colecciones de información disponibles, según tamaño y tipo.
- La distribución existente entre los diversos tipos de redes.
- La evaluación de las redes.

Pero estos estudios pueden encontrar diferentes tipos de problemas en el momento de su aplicación en el análisis de la Red. Internet se caracteriza por su dinamismo y constante estado cambiante, por lo que los documentos contenidos en el medio se someten al mismo tipo de inconsistencia en sus estados, generando un movimiento fluctuante entre su disponibilidad y persistencia (Arroyo, 2005).

Esto implica que existe una dificultad inherente al medio para poder medir con exactitud su tamaño, si bien es cierto que se puede extraer una fotografía del momento concreto en el que se realiza la medición que permita tener una idea general y, mediante la repetición del estudio, se pueda trazar una evolución para tratar de alcanzar una idea más exacta.

Otra de las limitaciones existente, se encuentra relacionada con la *visibilidad* de los componentes que se encuentran en el mismo. Internet se puede dividir en dos partes, una visible y otra invisible. El Internet visible es aquel cuyo contenido se encuentra, o puede ser, indizado, por lo que podrá ser encontrado por el usuario o un motor de búsqueda de un modo sencillo.

En la Figura 3 se puede ver una representación realizada por Michael Bergman que muestra de forma gráfica las diferencias entre las dos *Internets*

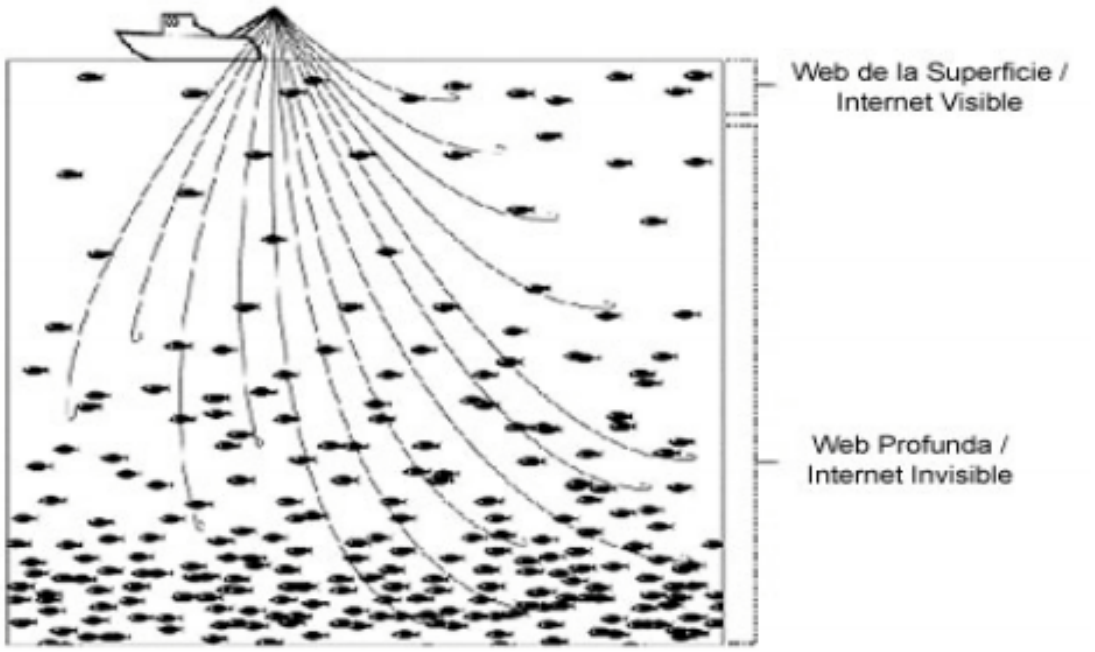


Figura 3: Ilustración representativa de Internet Profunda (Bergman, 2011)

En cambio, Internet invisible o Internet profunda[7], es la parte de Internet no indizada por los motores de búsqueda estándar, convirtiéndose así en una zona de difícil acceso para el usuario común (Lapuente, 2013).

Existe mucha información difícil de indizar debido a las características de la misma: volúmenes de datos extremadamente grandes, información fugaz, bases de datos dinámicas, lugares con pasarelas o contraseñas especiales de acceso, páginas sin conexión, etc (Aguillo, 2000).

Para poder comparar la diferencia entre los dos tipos de Internet, y comprender mejor la cantidad de información que no se encuentra visible para el usuario en general, en la siguiente tabla se puede observar las diferencias de tamaño, según un estudio realizado por la Universidad de California (Lyman, 2003):

[7] Término procedente del inglés: *Deep Web*.

Tabla 3: Tamaño de Internet en Terabites

Medio	Terabites en 2002
Web Visible	167
Web Profunda	91.850
Emails (Originales)	440.606
Mensajería instantánea	274
Total:	**532.897**

Del mismo estudio se extrae una serie de cifras interesantes para el presente trabajo, pese a que la fecha del mismo no es cercana, medir todo Internet no es sencillo y por ello no existe una gran cantidad de bibliografía. En la Tabla 4 se muestra la diferencia de cobertura que existe entre las dos *Internets*:

Tabla 4: Diferencia de cobertura entre Internet Profunda / Visible

Área temática	Diferencia cobertura
Salud	5,5%
Educación	4,3%
Estilo de vida	4,0%
Ciencias	4,0%
Noticias, Media	12,2%

La información que contiene la Tabla 4, indica por tanto, que una gran cantidad de información que podría llegar a ser muy útil tanto para pacientes como para profesionales (médicos o investigadores) no se encuentra fácilmente accesible.

Por otro lado, las herramientas utilizadas en general para indexar Internet son los motores de búsqueda y al tratarse de plataformas privadas de uso público, sus métodos de indexación quedan ligados a la capacidad económica de la empresa, su tamaño y cobertura. Debido a esto, serán analizados con mayor detenimiento posteriormente.

2.2.1.1 Líneas de investigación en Cibermetría:

Las líneas de estudio principales que pasan a desarrollarse en el Capitulo se centran en las partes:

- Descriptiva: centrada en el estudio de la propia disciplina, concentrándose en los aspectos relacionados con la cobertura, definición, unidades de medida y naturaleza de indicadores.

- Aplicada: centrada en el análisis del objeto de estudio.

- Instrumental: centrada en el estudio de las fuentes de información y los métodos de extracción de la información.

2.2.2 Cibermetría descriptiva

Como se ha descrito anteriormente, Internet se puede dividir en dos subconjuntos, uno más grande que el otro, siendo el espacio a analizar el más pequeño, que corresponde al Internet visible y su estructura.

Primero se debe recordar que Internet no es una red centralizada, y se puede dividir en una parte relacionada con los contenidos, y otra con los protocolos y recursos físicos que permiten su funcionamiento.

Tabla 5: Elementos físicos y virtuales de Internet.

Elementos Físicos	Elementos virtuales
Infraestructura física (ordenadores, servidores, etc)	Aplicaciones
Infraestructura de Comunicación (routers, hubs, etc)	Servicios (web, correo, chat, etc)
	Contenidos (objetos digitales)

Debido al enfoque del presente trabajo, se pasa a analizar los elementos que conforma el presente estudio. En este caso se trata del contenido asociado a Internet visible.

Es difícil indicar una cifra exacta sobre la cantidad de contenido existente en Internet, más allá de trabajos estimativos como el visto anteriormente. Para tratar de cuantificar la cantidad de contenidos, los principales estudios se

centran en la métrica relacionada con las páginas webs, es decir, tratar medir la cantidad de documentos disponibles en el *World Wide Web*.

Debido a que cada segundo se vuelvan en Internet millones de datos y, como hemos visto antes, sólo se encuentra indizado un porcentaje mínimo de Internet, únicamente pueden existir estimaciones a este respecto. Pese a ello, Internet Live Stats[8], como se puede ver en la Figura 4, calcula que en 2015 habían cerca de 900 millones de páginas, tras un descenso de 100 millones aproximados con respecto del año anterior.

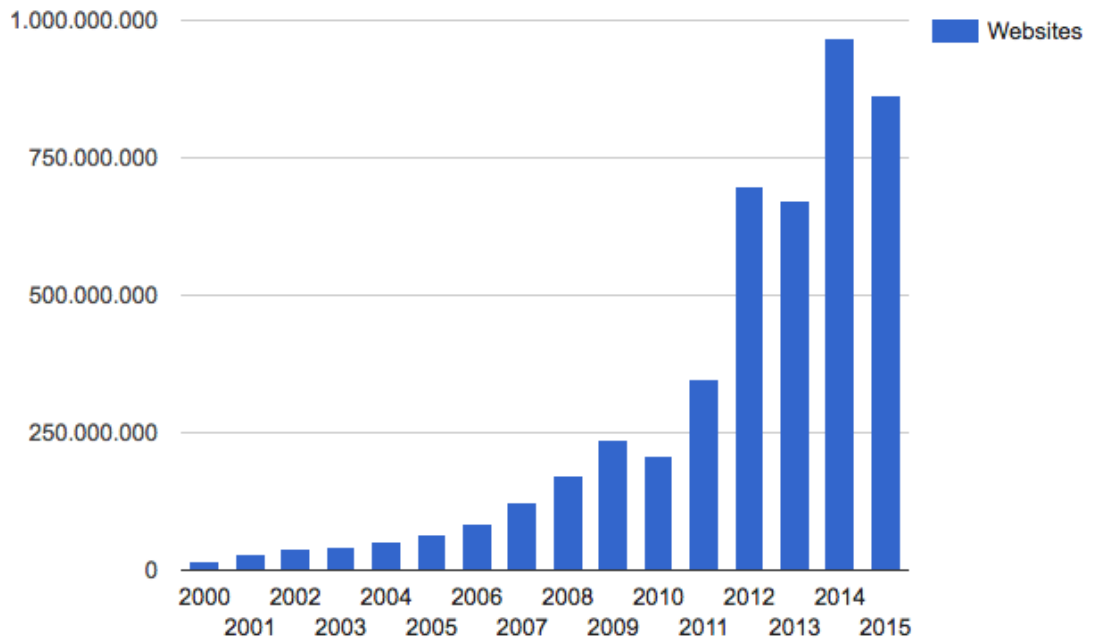


Figura 4: Número total de páginas en Internet (Fuente: Internetlivestats)

Tal y como se ha comentado anteriormente, las páginas indizadas son un porcentaje mínimo en comparación con la cantidad de contenido existente. Según el portal World Wide Web Size[9], el número actual de contenido indizado es de al menos 50 billones de páginas.

[8] http://www.internetlivestats.com/
[9] http://www.worldwidewebsize.com/

En la Figura 5 se puede observar la cantidad de contenido estimado que posee el motor de búsqueda Google indizado en billones de páginas, en el periodo comprendido entre Junio, Julio y Agosto de 2016:

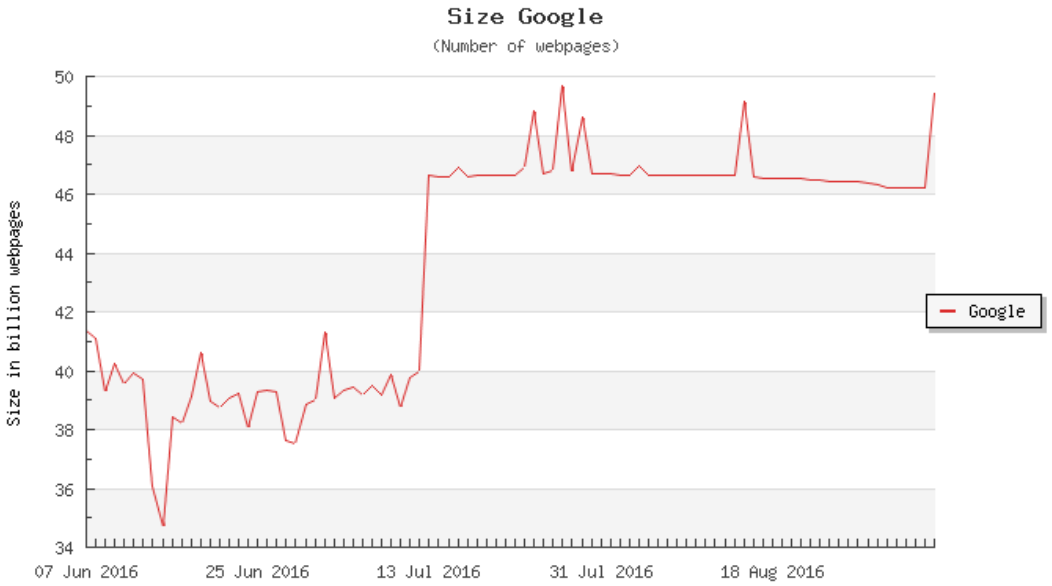


Figura 5: Cantidad de páginas indizadas por el motor de búsqueda Google (Fuente: WorldWideWebSize para el periodo Junio-Agosto de 2016).

Del mismo modo, se puede observar en la Figura 6, el número de páginas indizadas por Bing para el periodo equivalente:

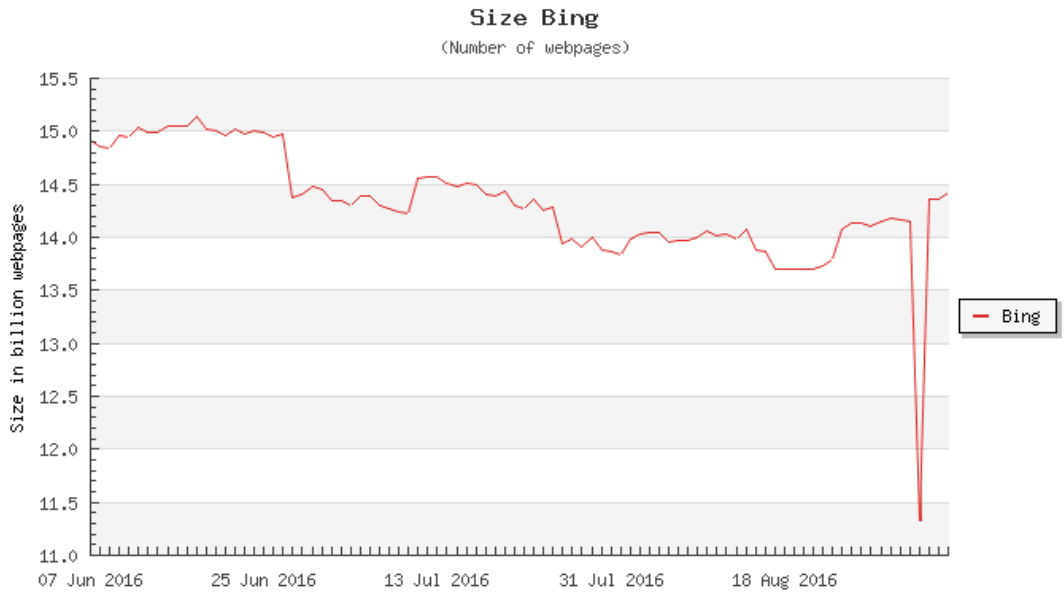


Figura 6: Cantidad de páginas indizadas por el motor de búsqueda Bing, (Fuente: WorldWideWebSize para el periodo Junio-Agosto 2016).

Una vez vista la cantidad de información existente y encontrada en la web es necesario conocer los elementos digitales que conforman el espacio, y que deben ser definidos para poder conocer las métricas aplicables, según el Consorcio de Internet[10] son:

- *Ficheros HTML*: *HyperText Markup Language* es un lenguaje de marcado estándar utilizado para la creación de páginas web. Actualmente se utiliza junto con otros lenguajes como la Hojas de Estilo en Cascada (CSS) y Javascript, para ofrece al usuario una mejor experiencia y calidad (W3C, 1997).

  El lenguaje HTML se caracteriza por el uso de etiquetas, que permiten la utilización otros elementos digitales embebidos, como imágenes o vídeos o textos. Una página web puede encontrarse formada por uno o más ficheros HTML.

- *Elementos gráficos fijos o animados*: estos elementos pueden ser imágenes fijas, en formatos como PNG o JPG. Los elementos animados pueden ser objetos de tipo GIF o vídeos. Estos elementos pueden mostrarse embebidos en una página o mostrados de forma individual. En el caso de los elementos animados de tipo vídeo, para que puedan ser mostrados en el navegador, es preciso que sean acompañados por reproductores o elementos que permitan su visualización.

- *Otros componentes multimedia*: como pueden ser sonidos o programas. Del mismo modo que los elementos gráficos animados, éstos deben ser acompañados de elementos que permitan su visualización en un navegador en caso de que deban ser mostrados de ese modo.

  Estos elementos están dispuestos para ofrecer un mayor número de funciones e interactividad. Algunos ejemplos de estos elementos son de tipo Flash y Java.

- *Variantes de lenguaje hipertextual*: algunos de estos lenguajes son por ejemplo XML, SGML o RDF. Son lenguajes orientados a definir la

[10] https://www.w3c.org/

estructura y semántica de un documento pero que no forman parte de la visualización final. En general son lenguajes jerárquicos y cuya gramática puede ser extendida, facilitando su programación. Una ventaja de los mismos es la capacidad de ser exportados por aplicaciones para poder ser *leídos* y utilizados de otros modos al inicialmente propuesto.

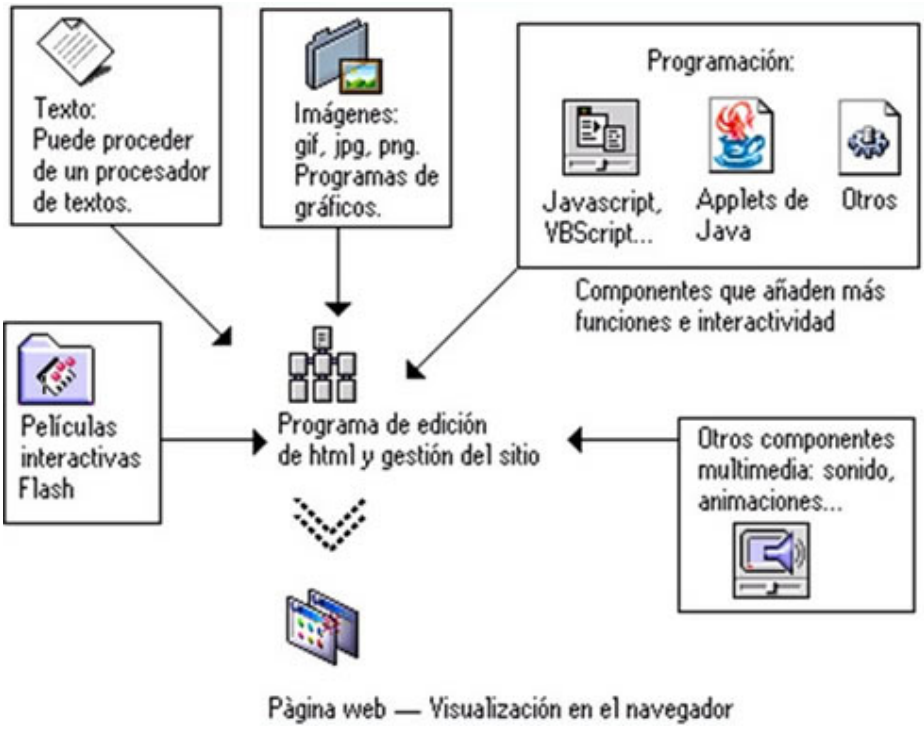


Figura 7: Elementos de contenido web [Fuente: disenowebakus.net].

- *Ficheros ricos*: los ficheros ricos son aquellos documentos o elementos web que se encuentran en un formato, generalmente propietario, como pueden ser PDFs o documentos Word (PPTs, DOCs, XLSs, etc). Generalmente aparecerán como enlaces para ser descargados ya que es preciso un sistema de lectura concreto. Si bien es cierto que el uso de los documentos PDfs como elementos web se encuentra generalizado y existen diversas soluciones para su visualización embebida.

- *Enlaces hipertextuales*: los enlaces, vínculos o links, son los encargados de vertebrar la web. Son el nexo de unión entre la información existente en Internet y uno de los motivos por el que se puede llamar Red (Lapuente, 2013).

  La definición que el W3Consortium tiene sobre los enlaces hipertextuales es:

  "Una relación entre dos anclas, guardadas en la misma o diferente base de datos"

  De modo que la función de los enlaces es la de conectar los diferentes nodos o elementos que conforman la web, por lo que los enlaces generan un camino desde un punta A de información a otro punto, que no tiene porque ser consecutivo (B), ofreciendo la posibilidad de realizar un tipo de navegación no secuencial.

  Según la finalidad de los mismos, como se puede observar en la Tabla 6, se pueden definir de la siguiente forma:

Tabla 6: Resumen tipología de los Enlaces (Orduña-Malea, 2011)

Tipo	Finalidad	Ejemplo
Organizativos	Organizan la información en un espacio web.	"Página siguiente"
Por Contenido	Relacionados con el significado del enlace	"Más información"
Intrínsecos	Remiten a otros ficheros para formar la misma web.	Uso de anclas: **Ancla**
Internos	Conectan mediante URL con otras páginas o ficheros del mismo sitio web.	** Inicio** Desde cualquier página dentro del dominio es.wikipedia.org
Externos	Remiten a URLs externas.	** Inicio** Desde cualquier página fuera del dominio wikipedia.org
Entrante	Inlink – puede ser tanto interno como externo	A <- B
Saliente	Outlink – puede ser tanto interno como externo	A -> B
Autoenlace	La fuente y el destino coinciden	A - A

Como se muestra en la Figura 7, la unión de todo o partes de lo anteriormente citado, visualizado a través de un navegador es lo que se conoce como página web. Las páginas puede ser estáticas o dinámicas. Las páginas web estáticas son aquellas básicamente informativas, no permiten la interacción del usuario con las mismas y ofrecen siempre el mismo contenido.

En cambio, las páginas web dinámicas son generadas mediante aplicaciones web, por lo que su contenido varía en función de una serie de parámetros o peticiones a la misma. Esto permite que tanto el usuario como los encargados del portal controlen, en mayor o menor medida, el contenido de la misma, implicando a su vez que puede no volver a repetirse la visualización de la página, y dificultando su correcta medida o análisis.

Teniendo en cuenta las complejidades de medición de todos los elementos indicados anteriormente, es posible aplicar métricas que permitan tener constancia de una serie de indicadores.

Existe una gran cantidad de tipos de indicadores que se pueden aplicar, pero dadas las características del presente proyecto se mencionan los que conciernen al mismo mediante un listado no exhaustivo (CSIC, 2003):

- Tamaño: se relaciona con el número de ficheros alojados en dominio, pueden realizarse dos aproximaciones, una en la que se incluyen únicamente los documentos HTML, y otra en la que se incluyen todos los elementos web.

Tabla 7: Indicadores Cibermétricos de Tamaño

Indicadores Cibermétricos de Tamaño	Unidad
Número de páginas	Grupo de países, país, región, dominio, subdominio, sede, documento, fichero
Nº total de objetos	
Nº de ficheros ricos	
Nº de ficheros media	
Nº de ficheros dinámicos/ejecutables/otros	
Nº aplicaciones web	
Tamaño de los ficheros	
Número de palabras	

- Visibilidad: relacionado con el número total de enlaces externos recibidos por un portal.

Tabla 8: Indicadores Cibermétricos de Visibilidad

Indicadores Cibermétricos de Visibilidad		Unidad
Total	Nº enlaces recibidos de diferentes sedes	Dominio, subdominio, sede
Institucional	Nº enlaces página principal	
Nacional/Internacional	Nº enlaces recibidos del mismo/diferentes dominios	

- Mención: se relaciona con el número de veces que un término aparece como resultado de una búsqueda.

- Popularidad: se relaciona con las visitas recibidas por un sitio web.

Tabla 9: Indicadores Cibermétricos de Popularidad

Indicadores Cibermétricos de Popularidad		Unidad
Consumo	Distribución de visitas página/directorio	Dominio, subdominio, sede
	Ficheros volcados	
	Distribución por institución	
	Distribución por país	
Posición	Posición relativa (Alexa)	

- Impacto: se trata de un indicador combinado, se calcula mediante la división de la Visibilidad por el Tamaño. Tiene tres variantes: una total, con todos los enlaces; una externa, dónde se utilizan únicamente los enlaces entrantes; y otra interna, dónde se utilizan únicamente los enlaces salientes (Smih, 1999).

Tabla 10: Indicadores Cibermétricos de Impacto

Indicadores Cibermétricos de Impacto		Unidad
Posición	Posición en buscador/término	Dominio, subdominio, sede
Impacto Web	WIF = Visibilidad/Tamaño[11]	
Importancia	Tamaño + Importancia + Popularidad	

[11] Se detalla únicamente el WIF total, entrarían en la misma categoría los indicados con anterioridad.

2.2.3 Cibermetría aplicada: Salud 2.0

Internet se definía en sus inicios como la "red de redes" debido a la interconexión existente entre las plataformas y estructuras que la conformaban (Salcedo, 2011). Actualmente la definición tiene más sentido que nunca ya que la capacidad de comunicar entre usuarios, permitiendo generar nuevas redes virtuales centradas en diferentes tipos de temáticas, y con la facilidad que aportan las nuevas herramientas sociales, es mucho más sencilla.

Esto es debido al concepto indicado en la introducción como Web 2.0, esta versión se centra en el usuario, poniendo a su disposición una serie de herramientas que aportan riqueza a su experiencia de navegación. No se trata tanto de un cambio tecnológico, como de enfoque y objetivos de la red (Definicion.de, 2016).

Las características principales de esta Web 2.0 radican en la contribución activa del usuario (O'Reilly, 2005), esto queda reflejado en aspectos como:

- La accesibilidad de los recursos
- El aumento de blogs, wikis y webs creadas por los usuarios.
- El incremento en el uso de las redes sociales.
- La importancia del *long tail*.
- El contenido agregado por los usuarios.
- El etiquetado colectivo.

Este cambio en el modelo comunicativo de la red puede ser aplicado a diferentes campos permitiendo su especialización como los Negocios 2.0, la Enseñanza 2.0, el Gobierno 2.0, la Ciencia 2.0 o la Salud 2.0, siendo este último el que da una base de conceptualización al presente trabajo.

La Salud 2.0 defiende el modelo indicado anteriormente, aplicado a la mejora de los servicios en la relación medico-paciente-enfermedad. Esto permite la existencia de un nuevo tipo de paciente, más informado y con conocimiento de diversas experiencias, lo que se puede traducir una mejora en su calidad de vida debido a las implicaciones que esto supone (Salcedo, 2011).

Si bien es cierto que el gran volumen de información que existe en Internet puede debilitar el mismo sistema al que pertenece, debido a las propiedades inherentes a la misma como son su autoría, control y propiedad (Giustini, 2008), y el peligro que puede suponer que sean los pacientes o personas sin

conocimientos especializados, los encargados de la transmisión de cierto tipo de información.

Es importante destacar que la información sanitaria que se comparte sea de calidad, libre de falacias y actualizada. Para los usuarios en España, la fiabilidad de la información encontrada sobre cuestiones relacionadas con la salud se evalúa con una nota de 7'2 sobre 10 (BBVA, 2008). Es por ello que se debe acentuar lo máximo posible la visibilidad de las asociaciones y organismos oficiales para que la información de confianza prevalezca sobre aquella que puede generar ruido o entorpecer la relación generada por la Salud 2.0.

Tabla 11: Encuesta Pew Internet y American Life Project Agosto 2006. Total (N=2928) Enfermos crónicos (N=269) Sin enfermedades (N=1711)

Temas de Salud buscados en Internet			
Tema	Porcentaje de usuarios que han realizado la búsqueda sobre el tema		
	Usuarios crónicos	Usuarios sin enfermedades	Todos
Enfermedad concreta o problema médico	73%	62%	64%
Tratamiento o procedimiento específico	64	49	51
Dieta, nutrición, vitaminas o suplementos nutricionales	53	48	49
Ejercicios o fitness	46	44	44
Prospecto fármaco	51	35	37
Doctor u Hospital en particular	33	28	29
Seguro médico	30	28	28
Tratamientos o medicina alternativos	42	25	27
Problemas por depresión, ansiedad, estrés o salud mental	30	21	22
Peligros ambientales para la salud	24	21	22
Tratamientos o medicina experimental	30	17	18
Inmunización y vacunas	13	16	16
Salud dental	16	14	15
Ayuda médica	24	11	13
Salud sexual	11	11	11
Dejar de fumar	18	8	9
Problemas con drogas u alcohol	8	8	8

En cuanto al consumo de esta información, según un estudio realizado por Pew Internet[12], el 86% de los usuarios de internet en América que viven con una enfermedad crónica o discapacidad, buscan información relacionada con al menos uno de los 17 temas sobre salud mostrados en la Tabla 1 (Internet, 2007), en comparación con el 79% de usuarios de internet que no tienen esas condiciones de salud.

Según el mismo estudio, el 75% de las personas con problemas de salud severos indicaron que basan la toma de decisiones respecto a sus tratamientos, su interacción con el doctor o la habilidad de tratar con su enfermedad basándose en lo leído en internet.

Un dato remarcable del estudio, y que respalda completamente la necesidad el presente trabajo, es que el 31% de los pacientes con enfermedades crónicas se sintieron frustrados por la falta de información o la incapacidad de encontrar lo que trataban de buscar en línea, en comparación con el 20% de los pacientes sin problemas de salud severos que se sintieron del mismo modo ante el mismo problema.

En cuanto a estudios previos realizados sobre Salud 2.0 y Cibermetría, la bibliografía existente cuenta con estudios realizados sobre todo fuera de España.

Aunque si se aplica para realizar rankings de Hospitales, del mismo modo que existen Rankings sobre Universidades a nivel mundial. Del Ranking en España se encarga el Laboratorio de Cibermetría[13] que pertenece al Centro de Información y Documentación Científica (CINDOC) parte del Centro Nacional de Investigación de España (CSIC), y realizan análisis cuantitativos de Internet y sus contenidos.

En estos momentos[14] el Ranking de Hospitales se encuentra en fase beta, pero cuando se encuentre en fase de producción su objetivo será el de "convencer a las comunidades académicas y políticas de la importancia de la publicación web, no solo para la diseminación del conocimiento académico, sino también como una forma de medir la actividad científica, el rendimiento y el impacto"[15].

En cuanto a análisis realizados fuera de nuestras fronteras, se han realizado análisis relacionados con la información sobre salud en línea para conocer el

[12] http://www.pewinternet.org/
[13] http://www.webometrics.info/
[14] Agosto de 2016.
[15] http://hospitals.webometrics.info/es/metodologia

nivel de patrocinio, tipo de plataforma y tipo de estructuración de enlaces (Groselj, 2013), pese a tratarse de un acercamiento de estilo más webmétrico, permite conocer el grado de visibilidad en función tipo de contenido ofrecido y el tamaño y características del dominio.

Otro estudio relacionado con Salud 2.0, llevado a cabo por Leanne Bowler, Wan-Yin y Daquin He (Leanne Bowler, 2011), en el que se estudia la visibilidad de seis portales relacionados con salud juvenil, concluye que los portales estudiados tienen una visibilidad limitada comparada con el resto de portales relacionados con salud.

Relacionado con la visibilidad de las Enfermedades Raras, se encuentra el estudio realizado por A. Castillo, P. López y M.C. Carretón (A Castillo, 2015), en el que se analiza el tipo de herramientas utilizadas para comunicación, en 143 portales de organizaciones españolas relacionadas con Enfermedades Raras. Algunos de los resultados extraídos del informe, indican que el 82% de las organizaciones tiene un portal funcional, y un 58% cuenta con un tipo de web 2.0. Los portales analizados se dirigen principalmente a los propios pacientes y allegados, en menor porcentaje a los medios de comunicación y profesionales.

2.2.4 Cibermetría instrumental

Es importante conocer las Fuentes de información que permiten desarrollar el trabajo en el campo Cibermétrico, por ello se describen en el presente apartado, aquellas fuentes utilizadas para el desarrollo del presente trabajo.

2.2.4.1 Motores de búsqueda

Los motores de búsqueda son la puerta de entrada a Internet, las herramientas que nos permiten buscar archivos almacenados en los servidores que conforman la Web. El primer buscador aparece en 1990 desarrollado por Alan Emtage, estudiante de la universidad McGill de Montreal (SearchEngineHistory, 2016), pero no es hasta la aparición de Google, y su expansión como gigante, cuando se convierten en herramientas indispensables.

Funcionan como una gran base de datos, mediante diferentes tecnologías, recopilan y almacenan la información encontrada, clasificándola del modo más

conveniente, para después poder servirla a los usuarios en sus paginas de resultados (SERPs)[16].

El funcionamiento general se divide en tres pasos: rastrear, indexar y buscar:

- *Rastrear*: en este paso los sistemas buscan información en la red que no tengan previamente guardada o sea conocida para los mismo. Para lograr esto se utilizan unas herramientas llamadas *"arañas"*, una araña es un programa informático que, autónomamente, lee un fichero estándar robots.txt y el cual le dirige a los archivos que conforman un determinado dominio para indexarlos.

- *Indexar*: el siguiente paso, una vez se conoce la existencia de los archivos, es almacenar la información encontrada, añadiendo asociaciones con términos o *palabras clave* que permitan definirla, para poder servirla al usuario cuando éste realice una búsqueda con alguno de los términos relacionados, creando así un *índice web*. Las palabras descriptivas son extraídas por la araña de los documentos encontrados, puede ser el texto del cuerpo del documento, metadatos, descripciones, etc.

- *Buscar*: el paso en el que entra en juego el usuario, desde el portal de búsqueda se realiza una solicitud de información que puede contener cualquier tipo de formula. Normalmente se trata de palabras clave que son recibidas por parte del buscador, analizadas y mediante las cuales se extrae de la base de datos la información previamente indizada.

Los algoritmos de búsqueda son el secreto mejor guardado de las compañías y, con el fin de poder servir a los usuarios la información más relevante, utilizan diversos sistemas de medición para otorgar una puntuación a los resultados, y proporcionar al usuario una respuesta acorde a su búsqueda.

Para esto, cada uno de los archivos indexados son medidos en función de unos parámetros, que varían dependiendo del buscador, y se utilizan para generar un ranking de importancia de los resultados en el que se basan para ofrecer el resultado de búsqueda al usuario. El sistema de medición se retroalimenta en función de la interacción del usuario con los resultados, para sumar o restar puntos y mejorar futuras búsquedas.

[16] *Search Engine Results Pages*

Existen tres tipos de buscadores (Wikipedia, Wikipedia, 2016):

- *Jerárquicos*: son los más parecidos a la descripción realizada anteriormente, ofrecen una interfaz desde la que realizar la búsqueda contra una base de datos y ofrecer los resultados. Ejemplos: *Google*[17] o *Bing*[18].

- *Directorios*: son directorios de enlaces a páginas que ofrecen motores de búsqueda interna. Por ejemplo: *Open Directory Project*[19].

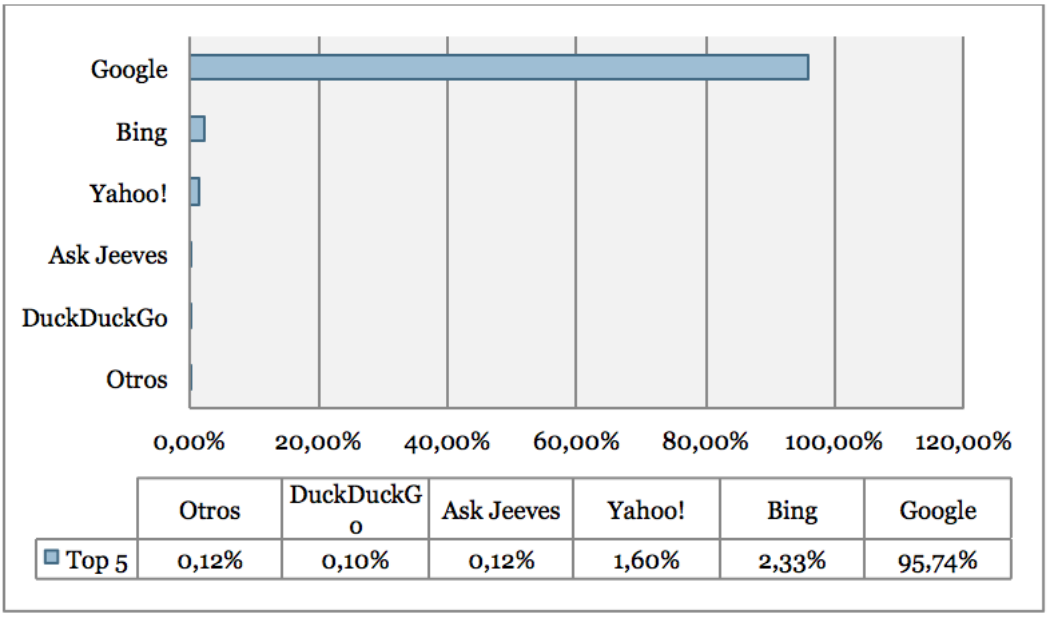
- *Metabuscadores*: se trata de interfaces desde las que se reenvía la búsqueda del usuario múltiples buscadores. Como ejemplo se encuentran *DogPile*[20] o *Metacrawler*[21].

Según el estudio mencionado con anterioridad realizado por la ONTSI sobre los ciudadanos ante la e-Sanidad (ONTSI, 2016), el 85% de los usuarios realiza las consultas sobre salud en internet a través de los buscadores. En cuanto a las búsquedas directas a portales relacionados con enfermedades específicas, según el estudio sobre e-Pacientes realizado por Pew Internet (Internet, 2007), el 37% de los enfermos crónicos comienza la búsqueda de información en un portal concreto, frente al 27% de los enfermos comunes en la misma situación.

StatCounter[22] es un servicio de analítica web que permite conocer el uso y las tendencias de navegación, así como otros datos, de los internautas. Además, ofrece gratuitamente los datos recopilados en su portal para que puedan ser utilizados. Según los datos de que dispone del uso en España el navegador más utilizado para el último año (Julio 2015-Julio 2016) es Google por un 95,74%.

La Figura 8 muestra el uso de todos los navegadores y sus porcentajes en España para el mismo periodo:

[17] http://www.google.com
[18] http://www.bing.com
[19] http://www.dmoz.org
[20] http://www.dogpile.com
[21] http://www.metacrawler.com
[22] http://gs.statcounter.com



	Otros	DuckDuckGo	Ask Jeeves	Yahoo!	Bing	Google
Top 5	0,12%	0,10%	0,12%	1,60%	2,33%	95,74%

Figura 8: Top 5 Buscadores España Julio 15 - Julio 16 (Fuente: StatCounter)

Para la realización del proyecto se utilizan tres buscadores, de los que se realizará un estudio previo, para poder ejecutar correctamente un análisis posterior de sus características que ayude a la obtención de resultados.

Como se ha indicado en el apartado 2.2.2, los motores de búsqueda presentan una serie de limitaciones, ya se ha presentado el problema de indización y alcance sobre Internet.

Además, se debe añadir los sesgos que existen en cuanto a cobertura. Al tratarse de sistemas automatizados se asume que los resultados son neutrales y sin sesgo que los limite, pero debido a que se trata de herramientas y servicios ofrecidos por compañías privadas, esto no es una realidad acertada (Goldman, 2006).

Por las propias políticas de las empresas, paginas que se encuentren indizadas pueden ser eliminadas de la base de datos o no mostradas a los usuarios al realizar una búsqueda.

Los algoritmos que se utilizan para realizar la búsqueda, indización y posterior creación del ranking de resultados de la web, se basan en factores indicados por los equipos humanos de trabajo, por lo que los resultados se encontrarán limitados por éstos a su vez.

Que exista un sesgo en los buscadores no es malo per se, ya que permite eliminar gran parte del contenido basura y trata de ofrecer los mejores

resultados, evitando a su vez la anarquía y preservando su credibilidad (Goldman, 2006).

El sesgo también puede ser debido al idioma o la localización, la mayoría de los buscadores ofrecen información en función de los gustos del usuario, historial o el idioma, por lo que se genera de este modo un sesgo al tratar de ofrecer información personalizada (Vaughan & Thelwall, 2003).

Los buscadores que se analizan en los siguientes apartados comparten ciertas características. Una de las más importantes es la posibilidad de realizar consultas aplicando operadores booleanos o mediante operadores clave.

Los operadores booleanos que pueden ser utilizados son (Google, Operadores de búsqueda, 2016):

- OR o | : para buscar páginas que contengan los dos términos indicados.
- + : para incluir palabras especificas. Esto resulta interesante en el caso de las consideradas "palabras ruido", como los artículos, ya que los buscadores tienen a eliminarlos de la consulta.
- - : para excluir el término indicado de la búsqueda.

Algunos de los operadores o palabras clave son:

- Filetype: para buscar todos los elementos indizados de un formato concreto, PDFs por ejemplo.
- Site: para limitar los resultados al dominio concreto indicado.
- Link: muestra las paginas que apuntan a dicha URL.
- Related: muestra contenido web similar al URL indicado en la consulta.
- El uso de comillas: al delimitar una palabra con comillas, se indica que los resultados deben incluir únicamente aquellas paginas que contengan exactamente las palabras incluidas.

Uno de los análisis cuantitativos que puede realizarse utilizando los motores de búsqueda mediante consulta directa es la *Estimación de Hits (Hit Count Estimates)*, es decir, la cantidad de resultados que devuelve la consulta, que sería el equivalente a cantidad de documentos indizados para el termino o términos indicado.

Esta técnica, no es perfecta. El hit count puede ser una estimación de resultados para volúmenes muy grandes, esto es debido generalmente a la necesidad de servir los resultados con un límite de tiempo para mejorar la

experiencia de usuario, pero supone que a grandes cantidades de resultados, la aproximación sea mayor.

Otra de las limitaciones a tener en cuenta es la eliminación de duplicados, los navegadores generalmente indican un número de hits, pero mientras se realiza un análisis de los resultados, el contador disminuye. Esto es debido a que los navegadores eliminan resultados que consideran duplicados, estos duplicados pueden ser debidos a páginas web dinámicas por ejemplo.

a. Google

El buscador de Google es el primer servicio que ofrece la compañía desde 1998, fue desarrollado un año antes por Larry Page y Sergey Brin y actualmente es el buscador más utilizado, no solo en España, si no en todo el mundo (Wikipedia, Wikipedia, 2016).

El buscador se centra de un modo muy intenso en la atención al usuario, por ello, además de buscar resultados relacionados con la palabra o palabras introducidas, ofrece una serie de características para mejorar la experiencia de navegación. Algunas de estas características incluyen datos meteorológicos, resultados de operaciones matemáticas, conversión de unidades, seguimiento de paquetes, información de aeropuerto, etc.

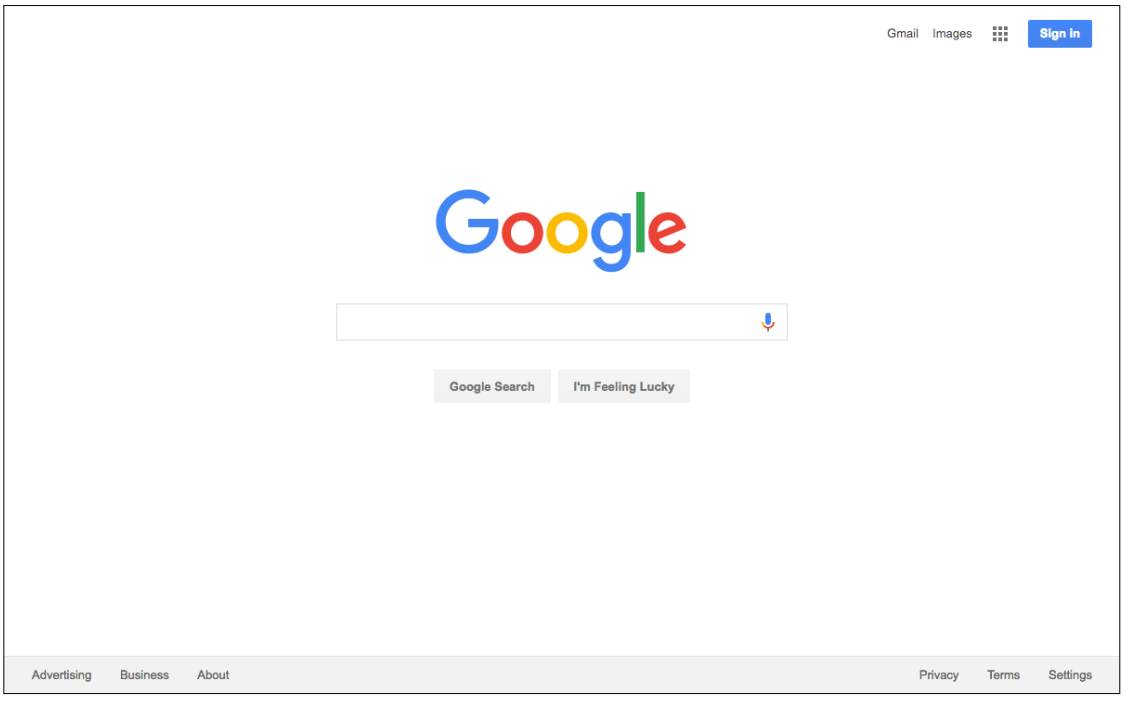


Figura 9: Página principal Buscador - Google

Es considerada una de las bases de datos informática más grandes del mundo, debido a que como hemos visto antes, es el motor de búsqueda con más elementos indizados de la web.

Su robot o araña más famoso se llama *Googlebot,* es el encargado de recoger los enlaces y mostrarlos como resultados tras una búsqueda. Tienen además otros robots encargados de extraer noticias u otros elementos web.

Al realizar una búsqueda ofrece dos tipos de ayuda, una relativa a la escritura de la consulta, mientras el usuario escribe ofrece sugerencias de queries, y la otra, llamada "Voy a tener suerte", ofrece un resultado concreto a una consulta determinada, sin pasar por la lista de resultados. Es decir, carga directamente una web determinada, normalmente coincide con el primer resultado a mostrar en el SERP.

En el caso de google el Hit Count aparece en la parte superior izquierda, bajo el menú de resultados de búsqueda como se puede observar en la Figura 10:

Figura 10: Contador de hits en Google

i. Dr. Google

En la descripción sobre el motor se han indicado algunas de las características que posee para ayudar a los usuarios. Todas las características tienen su base de partida en las búsquedas de los usuarios. Por ejemplo, Google Trends utiliza todas las búsquedas realizadas a sus bases de datos, para permitir la visualización mediante gráficas de la evolución de búsquedas sobre un término concreto de interés.

Otra de las características muy interesantes se encuentra relacionada con las búsquedas sobre salud (Google, Búsquedas médicas en Google, 2016).

El motor de búsqueda añadió el pasado año 2015 una funcionalidad basada en tarjetas de información, que permite al usuario que busca información relacionada con una enfermedad, y recibir un resumen con datos añadidos acerca de los síntomas y tratamientos.

Al buscar sobre una enfermedad, aparece la tarjeta en la parte derecha de la pantalla del navegador, con dos o tres pestañas:

- *Acerca de*: aparece una breve descripción de la enfermedad, incluida la duración, tratamientos existentes y facilidad de propagación. Una imagen asociada, la frecuencia de la enfermedad en el país de búsqueda. Formas de contagio en caso de ser contagiosa. Edades y géneros afectados.

- *Síntomas*: se indican los síntomas más comunes relacionados con esta enfermedad.

- *Tratamientos*: esta pestaña no se encuentra disponible en todos los países. En los que si aparece, indica medicamentos comunes utilizados para tratar la enfermedad. Otros tipos de tratamientos, como terapias, cirugías y dispositivos. Y tipos de especialistas médicos a quienes consultar sobre la enfermedad.

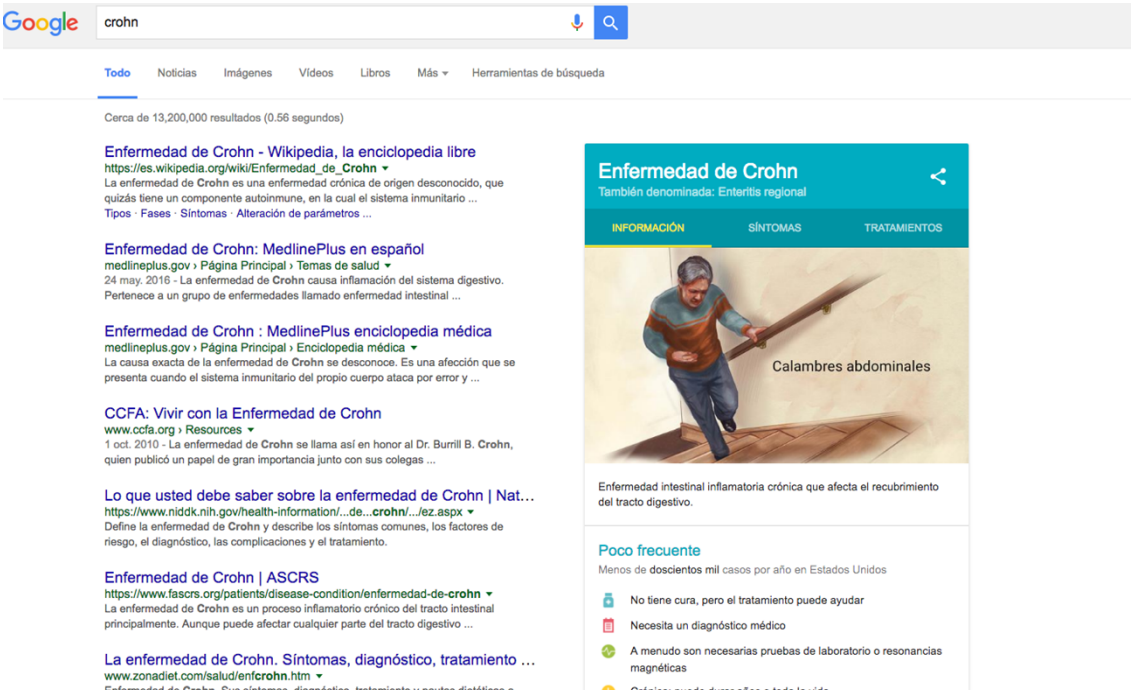


Figura 11: Ejemplo de visualización de tarjeta sobre enfermedad en Google

Las fuentes de información que utiliza son variadas, se trata de una combinación de algoritmos y profesionales de la salud. Mediante los algoritmos encuentran y extraen la información relativa a la enfermedad de base de datos de buena calidad en toda la Web. Después un equipo de profesionales revisa cuidadosamente la información y la perfecciona para mostrarla al publico. Por ultimo los ilustradores gráficos añaden una imagen.

Algunas fuentes de las que obtienen la información médica son:

- *Agencias gubernamentales*:
  - Institutos Nacionales de la Salud (National Institutes of Health, NIH)
  - Biblioteca Nacional de Medicina (National Library of Medicine, NLM).
  - Centros para el Control y la Prevención de Enfermedades (Centers for Disease Control and Prevention, CDC).
  - Instituto Nacional del Cáncer (National Cancer Institute, NCI).
  - Administración de Alimentos y Medicamentos (Food and Drug Administration, FDA).
  - ClinicalTrials.gov
  - Organización Mundial de la Salud (OMS)
  - Ministerio de Salud de Brasil

- *Ilustradores médicos:*
  - Suman Kasturia
  - Molly Borman-Pullen
  - Catherine Delphia
  - Cassio Lynm
  - Michele Graham
  - Christy Krames
  - Todd Buck
  - Jennifer E. Fairman

- *Asociaciones:*
  - Mayo Clinic, Estados Unidos
  - Apollo Hospitals, India
  - Columbia Asia Hospitals, India
  - Albert Einstein Hospital, Brasil
  - Lumiata
  - VoxHealth
  - Partnership for Drug-Free Kids

Un detalle importante es la existencia de un apartado en la tarjeta, en el que se indica si la enfermedad requiere de un diagnostico medico, y ofrece un listado de profesionales o especialistas a los que el usuario puede dirigirse. Esto se encuentra implementado para evitar que la gente se autodiagnostique o automedique.

Actualmente[23] el sistema cuenta aproximadamente con unas 1.000 tarjetas, el número exacto no es conocido, y tampoco existe un listado en el que poder consultarlas, por lo que en el Capitulo 3 se realizará un pequeño análisis del medio.

b. Bing

Bing es el buscador de la empresa Microsoft, en funcionamiento desde 2009. Se basa en los motores anteriores de la compañía: MSN Search y Windows Live Search.

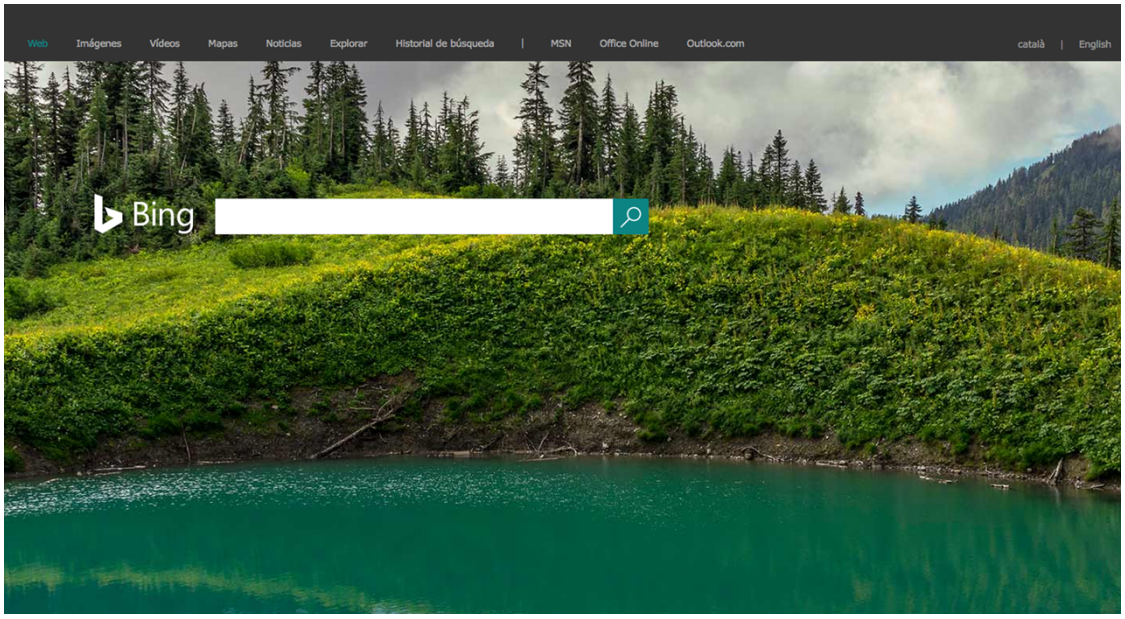


Figura 12: Página principal del motor de búsqueda Bing

[23] Agosto 2016

Algunas características señaladas que posee son el cambio diario en su imagen de fondo, a diferencia de Google, en la pagina de resultados cuenta con un listado añadido en el que indican *Búsquedas Relacionadas*, y otro con *Búsquedas Anteriores*.

Igual que Google, tiene características especiales relacionadas con las búsquedas, como pueden ser resultados de partidos en el momento, cálculos matemáticos, diccionario, etc. Así como imágenes o vídeos enlazados directamente en la primera página de resultados:

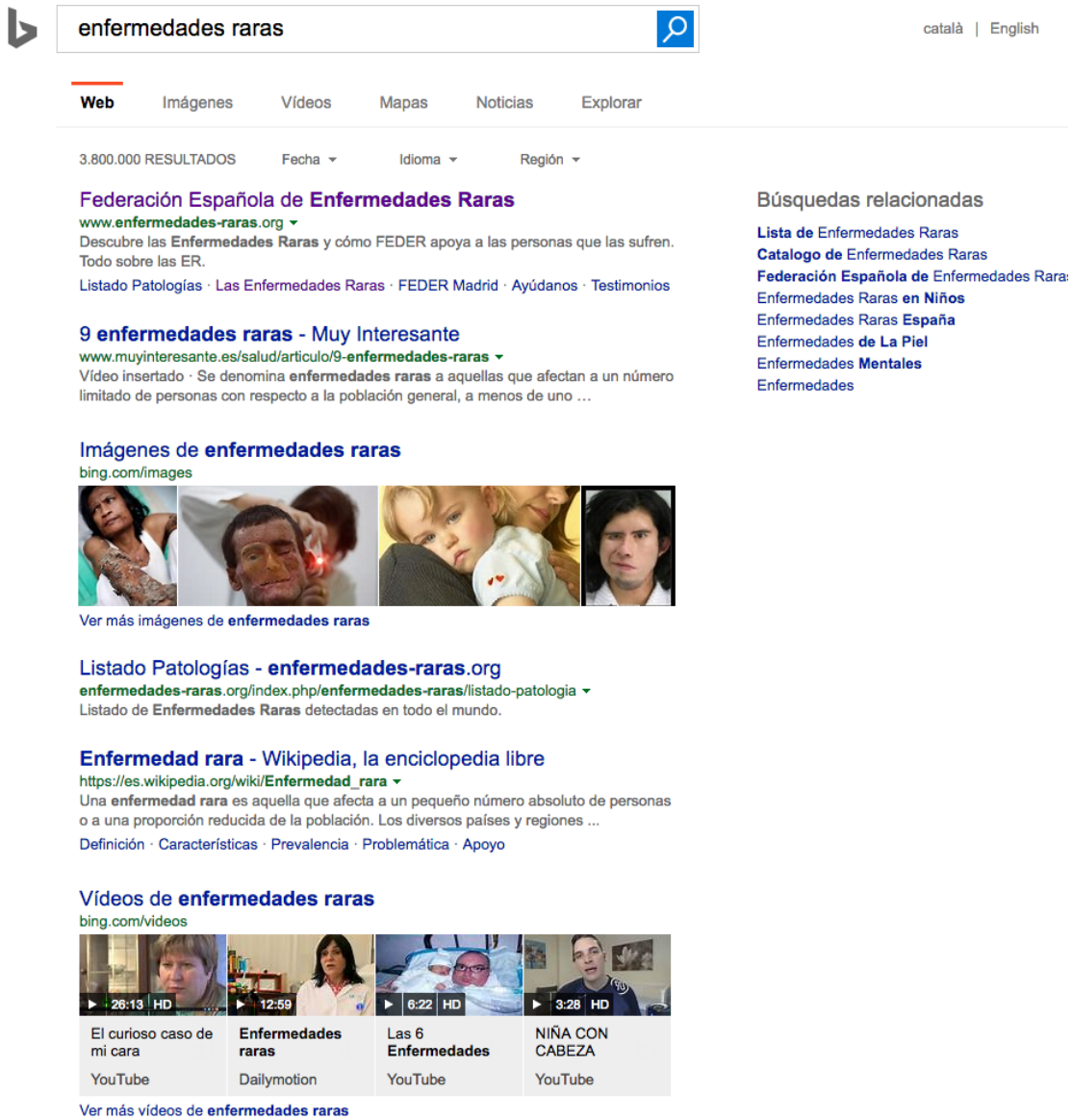


Figura 13: Página de resultados del motor de búsqueda Bing

ii. Consulta a la API

Bing, a diferencia de Google, ofrece a los usuarios una API (*Application Programming Interface*) para consulta a su motor de búsqueda.

Esta API permite descargar en formato CSV, XML o JSON todos los resultados que se obtendrían a través del navegador, de un modo sencillo y rápido.

Para solicitar resultados se pueden seleccionar entre diferentes tipos de fuentes (imágenes, video, Web, noticias, búsquedas similares), añadiendo si se quiere otros parámetros:

- Filtro madurez.
- Filtro de imágenes: tamaño y aspecto.
- Latitud y longitud
- Mercado: país y lenguaje.
- Categoría de noticias.
- Orden de noticias.
- Opciones: como resaltar en los resultados los términos de la búsqueda.
- Filtros de vídeo.
- Orden de vídeos.
- Tipo de archivo Web.

El uso de la API se limita a 5.000 transacciones gratuitas mensuales por cuenta de correo electrónico registrado. Las transacciones son el resultado de realizar la búsqueda y extraer cada una de las paginas de resultados, en cada página se muestran 50 resultados, por lo que para obtener los resultados de una consulta con 100 resultados, la API lo interpreta como dos transacciones pese a que únicamente se realiza una consulta.

c. Majestic

Debido a la desaparición de buscadores alternativos como AltaVista o Yahoo! y la limitada cobertura de otros como DuckDuckGo, para la realización del presente análisis de opta por utilizar Majestic como buscador alternativo.

Majetic-12 es un buscador distribuido a nivel mundial que nace en 2004, su fin ultimo es el mismo que el de los motores de búsqueda vistos anteriormente,

pero en este caso no existe una gran infraestructura perteneciente a una compañía, si no que se apoya en la comunidad.

Funciona mediante la instalación del cliente MJ12Node en el ordenador personal de cualquier usuario, de modo que todos los nodos recopilan información y la envían a la base de datos, generando así un motor de búsqueda distribuido.

Actualmente[24] se han descargado aproximadamente 5.000 millones de URLs, la Figura 14 muestra la evolución de las páginas rastreadas por día durante los últimos 10 años:

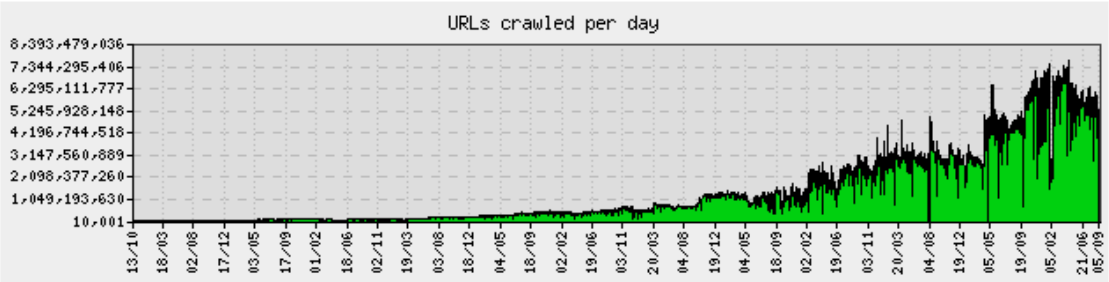


Figura 14: URLs rastreadas al día desde el inicio de Majestic-12

Actualmente el uso del buscador como usuario normal, no es accesible del mismo modo que lo son Google y Bing. Si bien es cierto que si devuelve las 1.000 primeras consultas en caso de que exista esa cifra o más como resultados, ni la interfaz, ni la visualización de resultados es amigable.

Majestic, a su vez, ofrece una API desde la cual poder descargar todos los datos almacenados relativos a un dominio o web. Estos datos se relacionan con el SEO y contienen información como *backlinks*, dominios de enlace, posición en el ranking de Alexa, etc.

Esta API será utilizada como método para descargar datos relativos a los dominios a analizar en el presente trabajo. Para conocer los detalles de la misma, se puede acceder al Glosario[25] disponible en su web.

[24] Agosto 2016
[25] https://es.majestic.com/support/glossary

2.2.4.2 Arañas

En el Apartado 2.2.4.1 Motores de búsqueda se ha visto como los motores de búsqueda utilizan Rastradores para encontrar la información en la Web y posteriormente indexarla.

Estos rastreadores, son también conocidos como arañas, robots de búsqueda, *crawlers*, *spiders* o *bots*. Su función es doble, son capaces de "rastrear" la web en busca de nueva información mediante los enlaces que existen, y también de extraer la información encontrada.

Reciben el nombre de *arañas* debido a que recorren la Web en busca de la información. En inglés el término Web significa *"tela de araña"* por lo que los programas son pequeñas *arañas* que trabajan *tirando del hilo* de la red para encontrar toda la información posible. Así mismo, en castellano si adquieren el segundo sentido de la palabra, ya que "arañan" las webs que recorren, recogiendo la información clave que se desee para ser almacenada.

La información que pueden recoger es muy variada, algunos ejemplos son (Ochando, 2012):

- Datos de imágenes, documentos, metadatos, etc.
- Colecciones de textos.
- Número total de enlaces analizados.
- Número de dominios, sitios y páginas web analizadas en cada nivel de profundidad.
- Distribución de dominios de tipo genérico y geográfico, según sitios y páginas web.
- Distribución de tipos de documentos según su extensión y formato.
- Análisis de macroestructuras web.
- Sitios web más enlazados.
- Trazado de hipervínculos entre sitios y páginas web que permite la elaboración de gráficas topográficas de los lugares analizados.

Las arañas pueden ser programadas de diferentes maneras, por lo que sus formas de trabajo varían en función de los resultados que se desean obtener. Existen diferentes tipos de herramientas que permiten la extracción y análisis

de los resultados de los motores de búsqueda, pero ninguna se adaptaba a las necesidades del presente proyecto, por lo que se opta por programarlas.

a. Scrapy

Se trata de un framework de código abierto y colaborativo, que permite extraer los datos deseados de diferentes tipos de webs. *Scrapy*[26] utiliza el lenguaje de programación *Python*[27], un lenguaje de alto nivel, interpretado y dinámico.

El funcionamiento es relativamente sencillo, una vez programada la araña cuanta con tres ficheros principales:

- *Configuración*: en este archivo se detalla el comportamiento de todos los componentes que conforman la araña, como por ejemplo: qué araña se debe utilizar, la velocidad a la que debe recorrer los sitios, los módulos que deben ser importados, el tiempo de descarga máximo que se debe utilizar, la cantidad de dominios en paralelo que se pueden recorrer y la identificación para evitar problemas con los motores de búsqueda o webs.

- *Items*: en este archivo se indican los datos que se desean recoger. Scrapy puede devolver los datos utilizando los diccionarios de Python, pero estos carecen de estructura, por lo que es necesario dársela y que la respuesta sea clara y sin errores.

  Para ello se utiliza la clase Items, que conforma un contenedor que recoge los datos obtenidos con la araña y los ofrece como resultados estructurados según interese al usuario.

  Se puede añadir, además, los metadatos necesarios para poder entender mejor los resultados obtenidos, de este modo utilizando los objetos Field, se puede recabar la información de modo estructurado.

- *Araña*: en este archivo se define el modo mediante el cual se debe extraer la información de una pagina, incluyendo el modo de realizar el rastreo y cómo extraer datos estructurados de las paginas.

  Su funcionamiento aproximado es:
  1. Las primeras solicitudes se obtienen mediante el uso de *start_requests()*, el que por defecto genera las solicitudes en

[26] http://scrapy.org
[27] https://www.python.org

función del as URLs especificadas en start_urls y el método de parseo para las llamadas.

2. En la función de callback, se parsea la respuesta, la página, y se devuelve a los diccionarios para extraer los objetos indicados en *ítem* y *Request*. Estas solicitudes contendrán también una devolución de llamada (posiblemente la misma), y serán descargadas por Scrapy.

3. En las funciones de callback, se parsean los contenidos de la pagina, normalmente mediante Selectores y se generan los *ítems* con los datos parseados.

4. Por último, los datos se devolverán al usuario para su guardado a base de datos, mediante tuberías, o escritos en un archivo mediante exportación.

Las arañas como herramientas son realmente útiles, ya que automatizan el trabajo manual a una velocidad mucho mayor. Durante el presente trabajo se han utilizado 4 arañas diferentes para extraer toda la información necesaria, que serán indicadas en el Capitulo 3.

Este tipo de programa presenta varios tipos de limitaciones y problemas:

- En el caso de Google, cuando el sistema detecta una gran cantidad de solicitudes, para evitar que se utilicen sistemas automatizados, aparece un captcha a ser resuelto. Para solucionar este problema, se opta por combinar dos soluciones, primero disminuir la velocidad a la que se realizan las consultas y que el tiempo entre ellas sea mayor que en otras arañas. La segunda solución, pasa por indicarle a la araña que abra una ventana en el navegador con la respuesta recbida si no coincide con lo que debería encontrar en ella mediante la orden *open_in_browser()*, para que pueda ser completado el captcha manualmente.
  En general, no se llega a necesitar la segunda opción, se realizan todas las consultas sin que aparezcan las preguntas, a excepción de unas pocas ocasiones mientras se realiza la extracción de hits por enfermedad como se verá en el siguiente capitulo.

- Las arañas se ven sometidas a la voluntad de los programadores de páginas web, en caso de que cambien una mínima parte del código dejarán de funcionar correctamente y deberán ser reprogramadas.

Deben ser utilizadas con mesura ya que, pese a que en este proyecto se encuentren siendo utilizadas para extraer datos de grandes empresas, en caso de que se utilicen con portales más pequeños podrían generar una serie de problemas como los detallados a continuación (Thelwall & Stuart, Web crawling Ethics revisited: cost, privacy, and denial of service, 2006):

- *Coste*: muchos hostings aumentan el precio del ancho de banda una vez este es sobrepasado, por lo que los propietarios de los portales se podrían ver afectados por elevado tráfico en el portal.

- *Denegación de servicio*: Del mismo modo, en caso de realizar un trabajo excesivo en un espacio limitado de tiempo, podría verse afectada la capacidad del servidor de ofrecer resultados a las peticiones que se dirijan, por lo que bien la araña, bien el resto de usuarios, se verían afectados por errores o la no visualización de la información.

- *Privacidad*: los robots son programas automáticos, por lo que no son capaces de distinguir el tipo de información que recogen, por lo que es el programador quien debe asegurar la anonimidad de los datos, para evitar que contengan información sensible.

- *Copyright*: pese a que la información se encuentra pública en Internet, ésta puede encontrarse bajo licencias de copyright que limiten su uso, por lo que se deberá tener en cuenta los posibles problemas legales que suscite el uso de las arañas.

En el estudio realizado por Thelwall & Stuart, se indican una serie de políticas que los usuarios de rastreadores o arañas deberían tener en cuenta:

- Tratar de buscar fuentes alternativas de información.
- Considerar la importancia de la información que va a ser recogida.
- Ser consciente de las implicaciones de costes que puede generar.
- Preparar una posible ayuda económica para los propietarios de los recursos en caso de causar sobre costes o denegación de servicio.

- Siempre seguir las indicaciones de Robots.txt[28]

2.2.5 Estudios previos

Existen diversos estudios que realizan un análisis, tanto de los resultados obtenidos por diferentes motores búsqueda, como de los diferentes métodos que se puede utilizar para la recolección de la información en la Web.

Mike Thelwall y Pardeep Sud (Thelwall & Sud, 2011), proponen utilizar las citas a URL y menciones a instituciones, comparándolos con el conteo de links, para calcular conocer la cobertura de los motores de búsqueda, la variación en la recuperación de la información, las posibles anomalías que puedan tener los motores de búsqueda y la polisemia de las consultas.

En este estudio se realiza una comparativa entre Yahoo! y Bing, utilizando diferentes métodos mencionados anteriormente. Utilizando las URLs de las instituciones, se realizan las consultas a las APIs de los motores de búsqueda. Tras aplicar correlación de Spearman a los resultados obtenidos, para evaluar el grado de concordancia existente entre las respuestas obtenidas.

Los resultados indican que existe poca diferencia entre el conteo de URL y el de dominios, los autores atribuyen estos resultados a que los motores de búsqueda filtran los resultados, eliminando duplicados como se ha comentado anteriormente. Pese a que los resultados se encuentran relacionados, los autores sugieren que se debe realizar las tres medidas diferentes ya que los resultados no son intercambiables.

En el estudio realizado por Judit Bar-Ilan (Bar-Ilan, 2005), presenta unas métricas basadas en las introducidas por Fagin et al., en las que se centra el análisis de sets de resultados no idénticos obtenidos tras realizar las búsquedas. Las nuevas métricas propuestas se utilizan únicamente sobre los documentos que se repiten en los resultados, utilizando el coeficiente de correlación de Spearman.

Para realizar el estudio, se toman en consideración únicamente las consultas que devuelven menos de 1.000 resultados. De los resultados descargados, se filtran las URLs y se generan las listas de resultados por motor de búsqueda,

[28] Robots.txt es un fichero de protocolo que se encuentra almacenado en el mismo servidor dónde están almacenados los recursos web, encargado de indicar si la información puede ser indexada.

de cada URL se indica su rango en la lista de cada motor de búsqueda y se rehace el ranking de nuevo.

Para obtener los resultados, aplican el coeficiente de Spearman a cada par de motores de búsqueda por cada una de las consultas realizadas. Además, calculan la correlación media entre pares de motores de búsqueda, utilizando el solapamiento de URLs descargadas.

Los resultados obtenidos indican que para consultas con porcentajes pequeños de resultados, Google es el buscador que mejor cobertura tiene. En este caso comparan Google con Altavista, siendo su correlación muy alta a excepción de alguna consulta aislada. Actualmente esta comparativa no es posible realizarla debido a la desaparición del motor.

Por último, el análisis realizado por Judit Bar-Ilan, Mazlita Mat-Hassan y Mark Levene (Bar-Ilan, Mat-Hassan, & Levene, 2005), en el que se exploran las diferentes metricas aplicables a la comparativa de rankings de resultados de los motores de búsqueda.

En este caso, se calcula primero el tamaño del solapamiento entre las listas de resultados por buscador. Después se aplica a los resultados la regla *Footrule* de Spearman. Esta regla calcula la distancia entre dos permutaciones.

Esta métrica puede extenderse al caso en el que dos listas no sean idénticas, pero debe incluirse un emplazamiento aleatorio para los resultados que se encuentran fuera de la misma.

Debido a los problemas que existen al aplicar estas métricas, los autores proponen el uso de una propia a la que denominan M. Esta formula trata de capturar la suposición de que los rankings idénticos o casi idénticos entre los primeros resultados de los buscadores es más interesante para el usuario que los documentos que aparecen en las partes finales de los rankings.

El estudio analiza por lo tanto los resultandos mediante el solapamiento de resultados, footrule de Spearman, la medida de Fagin y la métrica propuesta por los autores.

Algunos de los resultados principales obtenidos mediante el estudio secundan el anterior estudio realizado por Bar-Ilan, no existe un gran solapamiento entre buscadores, en el caso de imágenes el solapamiento es incluso menor.

3 Metodología

Dado que ya se conocen tanto las fuentes de información, como las herramientas a utilizar para poder recopilar y analizar la información. Se puede pasar a presentar la parte práctica del proyecto.

3.1 Diseño de la investigación

En el presente punto se indica los diferentes pasos que se realizan a lo largo de la investigación para la recogida de la información necesaria. No existe una línea temporal continua entre etapas, ya que algunas de las mismas pueden llevarse a cabo en paralelo.

- Fase 1: Análisis del interés por enfermedad.
  - Búsqueda de enfermedades.
  - Recopilación de datos relativos al interés.

- Fase 2: Búsqueda y selección de asociaciones.
  - Búsqueda de asociaciones relacionadas con Enfermedades Raras.
  - Selección de las asociaciones sobre las que se realizará el posterior análisis.

- Fase 3: Recopilación de datos.
  - Recopilación de datos en Google.
    - Recopilación de datos de Dr. Google.
  - Recopilación de datos en Bing.
  - Recopilación de datos en Majestic.

- Fase 4: Fase de preparación de los resultados.
  - Preparación de los datos finales.
  - Introducción de las métricas utilizadas.

- Fase 5: Fase de análisis de resultados
  - Esta fase será presentada en el Capitulo 4.

3.2 Fase 1: Análisis de interés por enfermedad

3.2.1 Búsqueda de enfermedades

En esta fase se realiza la búsqueda, recopilación y análisis de los datos relativos al interés por enfermedad. Como se indica anteriormente, existen aproximadamente 7000 enfermedades raras, debido a la cantidad de enfermedades, y la problemática relativa a la información que pueda existir, se decide conocer el interés que suscitan las mismas.

Para ello se descarga el listado disponible de enfermedades desde Orphadata[29], un portal que recopila la información que se ofrece en Orphanet.

El archivo se descarga en formato XML, en castellano, una fracción de su contenido se puede observar en la Figura 15:

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<JDBOR date="2016-05-01 04:14:53" version="1.2.4 / 4.1.6  (orientdb version)" copyright="Orphanet (c) 2016">
  <DisorderList count="9369">
    <Disorder id="17601">
      <OrphaNumber>166024</OrphaNumber>
      <ExpertLink lang="es">http://www.orpha.net/consor/cgi-bin/OC_Exp.php?lng=es&amp;Expert=166024</ExpertLink>
      <Name lang="es">Displasia epifisaria múltiple tipo Al-Gazali</Name>
      <DisorderFlagList count="1">
        <DisorderFlag id="476">
          <Label>on-line</Label>
        </DisorderFlag>
      </DisorderFlagList>
      <SynonymList count="2">
        <Synonym lang="es">Displasia epifisaria múltiple - macrocefalia - facies distintivas</Synonym>
        <Synonym lang="es">Displasia múltiple epifisaria con macrocefalia - rostro distintivo</Synonym>
      </SynonymList>
      <DisorderType id="21394">
        <OrphaNumber>377788</OrphaNumber>
        <Name lang="es">Enfermedad</Name>
      </DisorderType>
      <ExternalReferenceList count="2">
        <ExternalReference id="78407">
          <Source>ICD-10</Source>
          <Reference>Q77.3</Reference>
          <DisorderMappingRelation id="21534">
            <OrphaNumber>377808</OrphaNumber>
            <Name lang="es">el término específico se aplica a un término más amplio</Name>
          </DisorderMappingRelation>
          <DisorderMappingICDRelation id="21604">
            <OrphaNumber>377818</OrphaNumber>
            <Name lang="es">Atribuido</Name>
          </DisorderMappingICDRelation>
          <DisorderMappingValidationStatus id="21611">
            <OrphaNumber>377819</OrphaNumber>
            <Name lang="es">Validado</Name>
          </DisorderMappingValidationStatus>
        </ExternalReference>
        <ExternalReference id="38949">
          <Source>OMIM</Source>
          <Reference>607131</Reference>
          <DisorderMappingRelation id="21527">
            <OrphaNumber>377807</OrphaNumber>
            <Name lang="es">correspondencia exacta (los términos y los conceptos son equivalentes)</Name>
          </DisorderMappingRelation>
          <DisorderMappingICDRelation/>
          <DisorderMappingValidationStatus id="21611">
            <OrphaNumber>377819</OrphaNumber>
            <Name lang="es">Validado</Name>
          </DisorderMappingValidationStatus>
        </ExternalReference>
      </ExternalReferenceList>
    </Disorder>
```

Figura 15: Ejemplo de datos relativos a una enfermedad en el archivo XML sobre Enfermedades Raras

[29] http://www.orphadata.org/

Como se puede observar en la Figura 15, dentro de la etiqueta <Disorder> se encuentra la información relativa a una única enfermedad.

Para poder realizar el primera análisis, se precisan únicamente los nombres de las enfermedades raras. Por lo que es necesario extraer la información relativa a la etiqueta <Name>.

Debido a que el archivo cuenta con 473.595 líneas, se utiliza una librería Python llamada lxml[30] para el procesado del archivo, de modo que no se realice de forma manual. Como se puede ver en la Figura 17, utilizan los siguientes comandos para extraer los datos necesario:

```
from lxml import etree
doc = etree.parse("/../es_listado_enfRaras.xml")
raíz = doc.getroot()
jdb = raíz[0]
enfermedades = jdb.findall("Disorder/Name")
output = open('output.csv', 'w')
for element in enfermedades:
    print >> output, element.text
```

Figura 16: Líneas de comandos introducidos en Terminal.

En la línea 3 de código se selecciona la raíz (primera etiqueta) para poder trabajar con la información que existe dentro. Después, con el contenido en el array jdb, se extraen todas las etiquetas que se encuentran dentro de la etiqueta <Disorder> y se llaman <Name>.

Por ultimo, se vuelcan todos los datos obtenidos en un archivo .csv para poder trabajar con los mismos. Por lo tanto se genera un archivo que contiene todos los nombres de las enfermedades extraídas de la base de datos de Orpha.net.

3.2.2 Recopilación de datos.

Cómo se ha indicado anteriormente, la popularidad de un tema puede ser medida en función de la cantidad de hits que tiene una consulta en un motor de búsqueda. Por lo tanto, y debido al alto porcentaje de uso en la población

[30] http://lxml.de

española del buscador, se utiliza Google para conocer la cantidad de resultados que suscitan cada una de las enfermedades.

Para lograr esto se crea una araña cuyo funcionamiento básico será:

- Ir a Google
- Realizar la búsqueda con el nombre de la enfermedad.
- Recopilar la cantidad de hits, en caso de que existan, encontrados para la consulta.
- *Arañar* los datos.
- Guardar los datos en un fichero JSON.

La programación de la araña cuenta con tres ficheros, como se indica con anterioridad, estos tienen el siguiente contenido:

- Configuración (Figura 18):

```
# -*- coding: utf-8 -*-

# Scrapy settings for google_scraper project
BOT_NAME = 'google_scraper'

SPIDER_MODULES = ['google_scraper.spiders']
NEWSPIDER_MODULE = 'google_scraper.spiders'

USER_AGENT = 'Mozilla/5.0 (Windows NT 6.1; WOW64) AppleWebKit/537.36
(KHTML, like Gecko) Chrome/46.0.2490.86 Safari/537.36'

CONCURRENT_REQUESTS = 50
CONCURRENT_REQUESTS_PER_DOMAIN = 50
AUTOTHROTTLE_ENABLED = False
DOWNLOAD_TIMEOUT = 1200
#HTTPCACHE_ENABLED=True
```

Figura 17: Código Configuración Araña Hits Google

- Items (Figura 19):

```
# -*- coding: utf-8 -*-

# Define here the models for your scraped items
import scrapy


class GoogleScraperItem(scrapy.Item):
    hit_number = scrapy.Field()
    url = scrapy.Field()
```

Figura 18: Código Items Araña Hits Google

- Araña (Figura 20):

```
# -*- coding: utf-8 -*-
from scrapy import Spider
from scrapy.selector import Selector
from google_scraper.items import GoogleScraperItem
from scrapy.utils.response import open_in_browser

class GoogleSpiderSpider(Spider):
    name = "google_spider"
    allowed_domains = ["google.com", "google.es"]

    def __init__(self, filename=None):
        if filename:
            with open(filename, 'r') as f:
                queries = []
                for title in f.readlines():
                    queries.append(title.strip())
                self.start_urls = queries

    def parse(self, response):
        item = GoogleScraperItem()
        html = Selector(response).xpath('//html').extract()[0]
        did_not_match = html.find(' - did not match any documents.')

        try:
            no_results =
Selector(response).xpath('//*[@id="topstuff"]/div[2]/div/div[1]/text()').extract()[0]
            no_results = True
```

```
        except:
            no_results = False
        if no_results == True or did_not_match > 0:
            item['hit_number'] = 0
        else:
            hit_number =
Selector(response).xpath('//*[@id="resultStats"]/text()').extract()[0]
hit_number = filter(lambda x: x.isdigit(), hit_number)
            item['hit_number'] = hit_number
            item['url'] = response.url
        yield item
```

Figura 19: Código Araña Hits Google

El siguiente paso es generar las consultas que utiliza la araña para descargar los datos. Para introducir las consultas se trabaja con un fichero de tipo texto, al que la araña accede y del que lee línea a línea la consulta que debe descargar.

Las consultas cuentan con el siguiente formato:

https://www.google.es/search?&q="nombre de la enfermedad"

Una vez generado el archivo de texto con todas las consusltas a realizar, desde un terminal se ejecuta la orden para que la araña comience a trabajar:

```
> scrapy crawl google_spider –o hitsEnf.json –t json filename=hitsEnf.txt
```

- Donde "google_spider" es el nombre de la Araña que se ejecuta.
- hitsEnf.json el fichero que se genera.
- hitsEnf.txt el fichero del que se leen las consultas.

Una vez el programa finaliza, se obtiene un fichero de tipo JSON en el que se encuentra la URL de la que se extrae la consulta y el número de hits que aparece en Google.

De este modo se encuentran las Enfermedades Raras que generan mayor interés. Para la realización del proyecto se trabaja con el Top 50 de las asociaciones con más hits. Por lo que se trabaja con la lista ordenada de mayor a menor por hits y se seleccionan las primeras 50 enfermedades.

Debido a que todas las enfermedades que se listan en Orphanet no son consideradas enfermedades raras en España, dado el porcentaje de prevalencia en la población, se realiza una búsqueda en la base de datos del Registro Nacional de Enfermedades Raras del Instituto de Salud Carlos III, y aquellas enfermedades que se encuentren entre los primeros resultas y no lo hagan en la base de datos, se descartan.

En la Tabla 12 se puede ver el resultado obtenido tras aplicar los criterios indicados anteriormente:

Tabla 12: Listado Top 50 Enfermedades Raras con Hit de resultados en Google

Enfermedad Rara	hit_number
MODY	52.200.000
PANDAS	21.900.000
Alopecia	9.460.000
MELAS	5.230.000
Lepra	4.290.000
Neuroblastoma	3.630.000
Vasculitis	3.620.000
Glioblastoma	3.150.000
Sarcoidosis	2.870.000
Uveitis	2.600.000
Osteosarcoma	2.410.000
Linfoma	2.250.000
Retinoblastoma	1.950.000
Hemofilia	1.720.000
Osteonecrosis	998.000
Porfiria	875.000
Poliomielitis	856.000
Abetalipoproteinemia	843.000
Fibrosarcoma	842.000
Sialidosis	810.000
Encefalitis	748.000
Aniridia	738.000
Epispadias	573.000
Osteogenesis imperfecta	506.000
Ataxia - telangiectasia	499.000
Ictiosis	491.000
Insulinoma	483.000
Neurofibroma	458.000

Linfedema	451.000
Distrofia muscular	439.000
Galactosemia	436.000
Enfermedad inflamatoria intestinal	398.000
Oligodendroglioma	397.000
Inmunodeficiencia adquirida	375.000
Acromegalia	373.000
Esclerodermia	330.000
Fenilcetonuria	297.000
Microtia	279.000
Policitemia	211.000
CADASIL	203.000
Dextrocardia	200.000
Acondroplasia	197.000
Fiebre de Lyme	165.000
Pelagra	162.000
Miocardiopatia	152.000
Hipertension pulmonar	136.000
Distrofia muscular de Duchenne	135.000
Fibrosis quistica	128.000
Retinosis pigmentaria	123.000

3.3 Fase 2: Búsqueda y selección de asociaciones.

3.3.1 Búsqueda de asociaciones relacionadas con Enfermedades Raras.

Esta fase se plantea la prueba de diferentes métodos para escoger el más indicado y que mejor se adapte al proyecto.

3.3.1.1 Primer método

En este primer método, para la búsqueda y selección de asociaciones se plantea utilizar los datos de las enfermedades más interesantes de la fase anterior, tomando el Top 50 y, utilizando Google, se decide buscar en las 5 primeras páginas (10 resultados por página) de resultados, las asociaciones que aparecen para generar un listado de las mismas, utilizando alguna de las siguientes combinaciones de palabras según el orden indicado a continuación:

1. Asociación Española “nombre de la enfermedad”
2. Fundación Española “nombre de la enfermedad”
3. Asociación enfermos “nombre de la enfermedad”
4. Fundación enfermos “nombre de la enfermedad”

Este planteamiento llevado a la práctica no logra generar los resultados esperados, ya que, en general, no se encuentran resultados de asociaciones al realizar la búsqueda.

Debido a la gran cantidad de tiempo que se utiliza en las primeras consultas, y los resultados obtenidos, se decide probar un segundo método y descartar este, al menos hasta agotar las posibilidades de extracción de asociaciones por otros métodos.

3.3.1.2 Segundo método

Para el segundo método se realiza una búsqueda en Orphanet, FEDER y EURORDIS, las tres cuentan con un listado de asociaciones del que se pueden extraer los datos.

Por lo tanto se accede a las mismas y se descargan todos los datos relativos a las asociaciones de pacientes en España que posee cada una.

En el Anexo I se puede encontrar el resultado obtenido de cruzar las tres listas, habiendo eliminado duplicados y errores.

En la Tabla 13 se puede observar el número de asociaciones extraídos de cada una de las organizaciones y el número total de asociaciones únicas:

Tabla 13: Número de asociaciones por organización y total asociaciones.

Organización	Total
FEDER	312
Orphanet	262
EURORDIS	68
No duplicadas	*438*

De cada asociación se obtienen: las siglas en caso de que las tenga, el nombre de la asociación y la URL en caso de tener sitio web.

Este método resulta ser más rápido y eficaz que el anterior, por lo que el método 1 se descarta como solución.

3.3.2 Selección de las asociaciones sobre las que se realizará el posterior análisis.

Una vez se tiene el registro de las asociaciones completo, se seleccionan aquellas asociaciones que se encuentran relacionadas con las enfermedades listadas en el Top 50. Escogiéndose dos asociaciones por enfermedad.

Para seleccionar las asociaciones, se buscan las listadas por enfermedad y se seleccionan las asociaciones encontradas primero en FEDER, después en Orphanet y por último en EURORDIS. En caso de alcanzar el número de asociaciones antes de llegar al tercer listado, se pasa a la siguiente búsqueda.

Para realizar este paso, cabe señalar que existen asociaciones que tratan sobre más de una enfermedad, por lo que otro criterio aplicado es la búsqueda en el listado de la siguiente asociación en caso de que la encontrada primero ya se encontrase asociada a una enfermedad.

En Tabla 14 se encuentran los nombres de las asociaciones que se utilizarán para la realización del presente análisis, pudiendo consultar las respectivas URLs en el Anexo I:

Tabla 14: Listado de las 100 asociaciones seleccionadas.

Nombre Asociación	Nombre Asociación
Asociación de niños con enfermedades raras de Baleares	Asociación Cadasil España
Asociación Nacional de Addison y Otras Enfermedades Endocrinas	Federación Española de Padres de Niños con Cáncer
Asociación Española OsteoCondromas Múltiples Congénitos	Asociación española Hallervorden Spatz
Asociación Familiares y Enfermos Neuromusculares de Valencia	Asociación de Esclerodermia de Castellón
Asociación Lyme Crónico España	Federació Catalana de Malalties Poc Freqüents
Asociación Molinense de Enfermedades Raras	Federación de Ataxias de España
Asociación Madrileña de Osteogonesis Imperfecta	Federación Española de Hemofilia
Asociación de PKU y OTM de Madrid	Fundación Piel Sana
Asociación de Familiares y Amigos de Pacientes con Neuroblastoma	Asociación Española para la Galactosemia
Asociación Discapacitados Otros Ciegos de	Afectados por Linfedema Primario y Secundario de

España	Sevilla
Asociación Niños con cáncer	Asociación Española de Ictiosis
Associació Catalana d'Atàxies Hererditàries	Asociacion Microtia España
Asociación Cadasil España	Federación Española de Padres de Niños con Cáncer
Luchando contra el Lyme	Centro Nacional de Investigaciones Oncológicas
Asociacion Galega De Lupus	Asociación Corazón y Vida de Sevilla
Federación Española de Enfermedades Metabólicas Hereditarias	Asociación Dedines
Associació de cardiopaties congènites	Asociación Dgenes
Associació catalana de dèficits immunitaris primaris	Asociación de Afectados por Displasia Ectodérmica
ACCU Cataluña	Asociación Duchenne Parent Project España
ACCU España	Enfermedades autoinmunes Vasculitis CyL
Asociación Catalana de Neurofibromatosis	Federación Española de Enfermedades Raras
Asociación Española de Afectados por Linfoma, Mieloma y Leucemia	Asociación de personas y familias afectadas de extrofia vesical, cloacal, epispadias y patologías afines
Asociación Española de Afectados por Sarcomas	Fundación Contra la Hipertensión Pulmonar
Asociación Española Contra el Cáncer	Federación Gallega de Enfermedades Raras y Crónicas
Asociación Española de Déficits Inmunitarios Primarios	Federación Española de Lupus
Asociación Española Familiar Ataxia Telangiectasia	Federación de Personas Sordas de la Comunidad de Madrid
Asociación de Enfermos de Patologías Mitocondriales	Fundación Gaspar Casal
Asociación de Enfermedades Raras	Confederación Española de Familias de Personas Sordas
Asociación Humanitaria de Enfermedades Degenerativas y Síndromes de la Infancia y Adolescencia	Federación Española de Fibrosis Quística
Asociación de Huesos de Cristal de España (SID)	Asociación de enfermos de Fiebre Mediterránea Familiar de España
Asociación de Autoinmunes y Lupus de Almería	Asociación Madrileña de Fibrosis Quística
Asociación Madrileña de Personas con Artritis Reumatoide	Fundación ALPE Acondroplasia
Asociación de Niños con Discapacidad de Almería	Fundación Rebecca de Alba
Asociación Española de Aniridia	Grupo Español de Investigación en Sarcomas
Asociación ourensana de esclerose múltiple,	Asociación Nacional de Hipertensión Pulmonar

ELA, parkinson e outras enfermidades neurodexenerativas	
Asociación de Pacientes Coronarios	Associació catalana d'afectats de limfedema
Asociación de Pacientes con Tumores Raros de España	Asociación Valenciana de Afectados de Lupus
Asociación Andaluza de Hemofilia	Asociación Miastenia de España
Asociación de Afectados por Tumores Cerebrales en España	Asociación Española de Mucopolisacaridosis y Síndromes Relacionados
Federación Española de Enfermedades Neuromusculares	Asociación de Afectados de Neurofibromatosis
Asociación de Enfermedades Neuromusculares de Córdoba	Asociación Síndrome de Noonan Asturias
Asociación de enfermedades Neuromusculares de Andalucía	Asociación Nacional de Porfiria
Asociación Española de Extrofia Vesical	Asociación Afectados de Polio y Síndrome Post-Polio de España
Asociación de familias con perthes	Asociación Postpolio Madrid
Asociación de Hemofilia de la Comunidad de Madrid	Asociación Amigos de los Leprosos-Raoul Follereau
Asociación Síndrome Hemolítico Urémico Atípico	Fundación Retina España
Asociación Oncológica de Madrid	Asociación de afectados por retinosis pigmentaria-Retinitis pigmentosa de la Comunidad Valenciana
Asociación de Pacientes de Uveítis	Federación de Asociaciones de Retinosis Pigmentaria de España
Asociación de Enfermos Neuromusculares de Bizkaia	Asociación Nacional de enfermos de Sarcoidosis
Asociación Nacional para Problemas de Crecimiento	Sociedad Española de Diabetes
Federación Andaluza de Asociaciones de Ataxias	Associació Syndrom

3.4 Fase 3: Recopilación de datos relativos a las asociaciones.

Una vez se tiene el listado de las asociaciones y sus URLs, se pasa a recoger los datos para cada uno de los buscadores. Las arañas de Google y Bing son similares a la mostrada anteriormente en el apartado 3.2.2 simplemente se cambia en el fichero Item -> **hit_number** por **url_serp**.

Los cambios específicos que se realizan en el fichero Araña se indican en los apartado correspondientes a los buscadores.

Los ficheros de consultas que alimentan estas arañas están formados por el mismo tipo de consultas que en el caso anterior, pero cambiando el nombre de la enfermedad por la URL de la asociación, y añadiendo el operador de búsqueda *site*.

3.4.1 Recopilación de datos en Google.

En este caso se cambian las líneas que hacen referencia a **hit_number** por las siguientes:

```
url_serp_list = sel.xpath('//*[@id="rso"]/div/div[3]/div/h3/a/text()').extract()
url_serp_list = [re.search('q=(.*)&sa',n).group(1) for n in url_serp_list]
item['url_serp'] = url_serp_list
```

Figura 20: Código extracción resultados Google

3.4.1.1 Recopilación de datos en Dr. Google

En el Apartado i, se indica que de las tarjetas sobre enfermedades que aparecen al realizar una búsqueda en Google.com no existe un listado o directorio para poder conocer de cuales existen.

Dado que las enfermedades con tarjeta pueden o no ser raras, se prepara una Araña para conocer la cantidad de enfermedades raras con tarjeta.

En esta araña, el código necesario para implementarla es:

```
try:
        no_results =
Selector(response).xpath('//*[@id="rhs_block"]/div/div[1]/div/div[1]/div[2]/div/div/di
v/div[1]/div/div[1]/div/div[1]/span/text()').extract()[0]
        no_results = True
    except:
    no_results = False
    if no_results == True or did_not_match > 0:
        item[results] = no_results
    else:
        item['results'] = 0
```

Figura 21: Araña para conocer si existe tarjeta sobre Enfermedad Rara en Dr. Google

Las consultas que alimentan el fichero son las mismas que las utilizadas para la Araña de conteo de hits.

3.4.2 Recopilación de datos en Bing.

En el caso de Bing para la recogida de hits debe cambiarse el XPath que indica dónde se encuentra el dato a cambiar por el siguiente:

```
//*[@id="b_tween"]/span[1]/text()
```

Y para la recogida de datos relativos a las URLs de resultados, la araña debe cambiar los XPath por:

```
//*[@id="b_results"]/li[2]/div[1]/h2/a/text()
```

3.4.3 Recopilación de datos en Majestic.

En el caso de Majestic, los datos recogidos forman parte de un tipo de búsqueda totalmente distinta, por lo que la araña cambia completamente. Una fracción de la programación de la misma se muestra en la siguiente porción de código:

```
def __init__(self, filename=None):
        if filename:
            with open(filename, 'r') as f:
                queries = []
                for title in f.readlines():
                    queries.append(title.strip())
                self.start_urls = queries

    def parse(self, response):
        sel = response
        item = MajesticScraperItem()

        item['trustMetric'] =
sel.xpath('//table[@class=\'resultsTable\']/tr[43]/td[2]/text()').extract_first()
        item['trustFlow'] =
sel.xpath('//table[@class=\'resultsTable\']/tr[42]/td[2]/text()').extract_first()
        item['topicalTrustFlow_value_0'] =
sel.xpath('//table[@class=\'resultsTable\']/tr[39]/td[2]/text()').extract_first()
        item['topicalTrustFlow_topic_2'] =
sel.xpath('//table[@class=\'resultsTable\']/tr[38]/td[2]/text()').extract_first()
item['refSubNets'] =
sel.xpath('//table[@class=\'resultsTable\']/tr[32]/td[2]/text()').extract_first()
        item['refIPS'] =
sel.xpath('//table[@class=\'resultsTable\']/tr[31]/td[2]/text()').extract_first()
```

Figura 22: Código Araña para Majestic

3.5 Fase 4: Preparación de los resultados.

3.5.1 Preparación de los datos finales.

Una vez recopilados todos los datos, estos se encuentran en diversos archivos y formatos, por lo que es necesario tratarlos para poder trabajar con ellos en el análisis.

Para extraer los datos de los ficheros que se han descargado de todas las arañas, se trabaja con JSON, CSV y un programa Python que lee los datos y los exporta para no tener que trabajar manualmente la extracción de datos de los ficheros.

```
original = ["pdf", ".xls", ".doc", ".jpg", ".png", ".cgi", "aspx", “html”]
matches = []
frequency = []
for word in the original
    if word not in matches
        matches.append(word)
        set frequency = 1
    else
        increase the frequency
sort matches based on frequency
```

Figura 23: Código Phyton para extracción de datos

Todos los datos son exportados a un documento de Excel para poder trabajar con las tablas.

3.5.2 Introducción de las métricas utilizadas.

- *SERP*: los SERPs son los resultados que se obtienen al realizar una búsqueda en un motor. Existen dos tipos, $SERP_1$ y $SERP_n$. En el primer caso, son los hits que aparecen al realizar la búsqueda en la primera página. Los segundos, son aquellos hits que aparecen cuando se llega a la última página.

  Normalmente el número es menor en el segundo caso, ya que los buscadores eliminan resultados duplicados o contenido que no consideran relevante.

- *URLS* únicas: se trata del cálculo diferencia entre los resultados obtenidos por buscador y URL buscada, se trata de un cálculo que permite conocer el porcentaje de solapamiento entre motores de búsqueda.

- *Impacto*: medido en este trabajo con la fórmula tradicional del Web Impact Factor. Para calcular el impacto de las Asociaciones se pasa a utilizar el indicador Web Impact Factor.
  La formula que utiliza este indicador es:

$$WIF = \frac{A}{D}$$

  Dónde A son todos los enlaces que recibe la página y D el número de páginas que se encuentran indexadas por un motor de búsqueda.

  Pese a ser un indicador que ha recibido muchas críticas debido a la existencia de artefactos matemáticos, para conocer el impacto dentro y fuera de un país si es interesante conocer los resultados que pueda ofrecer (Noruzi, 2006).

- *Correlación de Spearman*: medir la fuerza y la dirección de la relación entre dos variables, donde el valor r=1 significa que existe una correlación positiva perfecta y -1 significa una correlación perfecta negativa.

  La formula utilizada es la siguiente:

$$R_s\, r_s = \rho_{rgx,rgy} = \frac{cov(rgx, rgy)}{\sigma rgx\, \sigma rgy}$$

  Donde:

  - X e Y son los valores de los dos rangos a comparar.
  - P es el coeficiente de correlación de Pearson.
  - σ de x y σ de y son las variables relativas a la desviación típica.
  - La fórmula se calcula con un umbral de disimilitud del 0,1.

- CitationFlow: indica el numero de menciones realizado de un dominio determinado. Es decir, la cantidad de veces que alguien menciona mediante un enlace a la web A desde la web B.

- Enlaces de referencia Externos: tras analizar los resultados, únicamente 25 portales se encuentran enlazados por lugares de referencia. Lo que implica una menor visibilidad. Esto debería ser tenido en cuenta por las autoridades competentes, del mismo modo que se crea una base de datos relacionada con las enfermedades, ayudaría a la visibilidad de su información, generar una base de datos centralizada de las asociaciones.

- *Alexa Linking In:* índice de Alexa en relación con los enlaces entrantes a las webs. Este índice no indica el número de enlaces, pero si la cantidad de portales que se encuentran apuntando en la dirección del portal.

4 Resultados y discusión

Una vez recopilados todos los datos, se procede al análisis de los mismos.

4.1 Tamaños por buscadores:

En la Tabla 15 se puede observar una muestra de los resultados obtenidos por asociación y buscador, siendo el número de hits, el dato de estimación de indización indicado por cada buscador; el Número de URLs, el número total de URLs extraídas para la búsqueda y las URLs únicas, las URLs únicas extraídas de la comparación entre resultados:

Tabla 15: Muestra de los resultados obtenidos por asociación y motor de búsqueda

	Google		Bing		URLs
	$SERP_1$	$SERP_n$	$SERP_1$	$SERP_n$	únicas
aecc.es	79800	567	17800	1029	1039
enfermedades-raras.org	42700	569	6610	487	786
asociaciondoce.com	4550	604	4280	274	299
cnio.es	4270	576	6740	975	786
asem-esp.org	3960	329	4230	960	665
mpsesp.org	3700	346	260	277	300
ataxiasandalucia.org	3150	187	138	154	256
fqmadrid.org	2690	241	209	211	305
fegerec.es	2660	202	60	60	220
fgcasal.org	2370	562	3470	1032	945
fesorcam.org	2330	359	4700	1006	640
accuesp.com	1890	245	41	65	288
duchenne-spain.org	1840	545	4450	1040	708
fedaes.org	1800	205	192	205	282
corazonyvida.org	1680	563	5190	135	560
felupus.org	1550	142	403	187	155
sediabetes.org	1510	556	3200	690	570
aacic.org	1350	291	1150	675	387
fiapas.es	1310	230	2270	885	518
fundacionrebeccadealba.org	1270	413	341	340	500

En la Tabla 16 se muestran los datos comparativos de máximos y mínimos de páginas indizadas por Google y Bing. Estos resultados son los obtenidos al recoger los hits que aparecen una vez introducida la consulta.

Tabla 16: Resultados comparativos entre Google y Bing mediante hits

	Google	Bing
Mínimo	8 páginas	8 páginas
Máximo	688 páginas	17800 páginas
1º Cuartil	29% tiene 110 páginas o menos	27% tiene 36 páginas o menos
2º Cuartil	24% tiene entre 330 y 111 páginas	24% tiene entre 144 y 37 páginas
3º Cuartil	25% tiene entre 1020 y 331 páginas	26% tiene entre 453 y 145 páginas
4º Cuartil	22% tiene más de 1020 páginas	23% tiene más de 453 páginas

Del mismo modo se calculan y comparan los datos entre el $SERP_1$ y $SERP_n$, los resultados son mostrados en la Tabla 17:

Tabla 17: Resultados comparativos entre Google y Bing, mediante enlaces reales.

	Google	Bing
Mínimo	7 páginas	8 páginas
Máximo	79800 páginas	17800 páginas
1º Cuartil	26% tiene 55 páginas o menos	27% tiene 36 páginas o menos
2º Cuartil	25% tiene entre 145 y 56 páginas	27% tiene entre 135 y 37 páginas
3º Cuartil	25% tiene entre 232 y 146 páginas	23% tiene entre 305 y 136 páginas
4º Cuartil	24% tiene más de 232 páginas	23% tiene más de 305 páginas

Dado que se tienen los datos del numero de hits iniciales y los datos de enlaces que ofrece el buscador al final, se puede extraer un porcentaje, y su distribución por cuartiles, sobre la cantidad de enlaces finales que muestra el

buscador en función de los hits iniciales, se pueden ver los resultados en la Tabla 18:

Tabla 18: Resultados comparativos entre Google y Bing mediante % $SERP_1$ y $SERP_2$.

	Google	Bing
Mínimo	0,7% extracción/hits	0 extracción/hits
Máximo	109% extracción/hits	466% extracción/hits
1º Cuartil	27% tiene menos de un 26% de extracción/hits	28% tiene menos de un 63% de extracción/hits
2º Cuartil	25% tiene entre un 52% y un 27% de extracción/hits	21% tiene entre un 98% y un 64% de extracción/hits
3º Cuartil	25% tiene entre un 89% y un 53% de extracción/hits	31% tiene entre un 100% y un 99% de extracción/hits
4º Cuartil	23% tiene más de un 90% de extracción/hits	20% tiene más de un 100% de extracción/hits

Destaca en este análisis que existen resultados para los que se han recogido más enlaces reales, es decir, el motor de búsqueda ofrecía más resultados que los indicados en un primer momento mediante hits.

Esto es algo que sucede en los dos motores de búsqueda, pero se da con mayor frecuencia en Bing que en Google. Google ofrece en un resultado un 109% de resultados. En cambio Bing supera el 100% en 20 resultados.

Una de las causas posibles puede que sea la propia infraestructura de los motores. Pese a que no se sigue esa línea de investigación en el presente trabajo, es necesario apuntar, que los motores de búsqueda son físicamente servidores en los que se alojan bases de datos. Repartidos por toda la geografía existen multitud de nodos, que replican la información de unos a otros. Cuando un usuario realiza una petición de búsqueda, normalmente se busca la información en el nodo más cercano, pero puede suceder que por motivos de limite de conectividad, fallo de servidor, etc, la solicitud de información se realice a otro nodo.

Esto explicaría el motivo por el cual existen resultados por encima del 100% de enlaces en el $SERP_n$ frente al $SERP_1$.

En las comparaciones anteriores entre motores utilizando la cantidad de páginas y su distribución mediante cuartiles a modo de comparación (Tabla 16 y Tabla 17) no resulta tan pronunciada la diferencia de distribución entre cuartiles como sucede en el caso de los porcentajes de resultados. En este

sentido Google se mantiene mucho más estable y guarda unas proporciones en cuanto a las diferencias de resultados que Bing no posee.

En la Figura 23 se muestra un gráfico de dispersión que permite conocer la relación por dominio en los buscadores, el eje X representa a Google y el eje Y a Bing, es decir $SERP_1$ de Google contra $SERP_1$ de Bing:

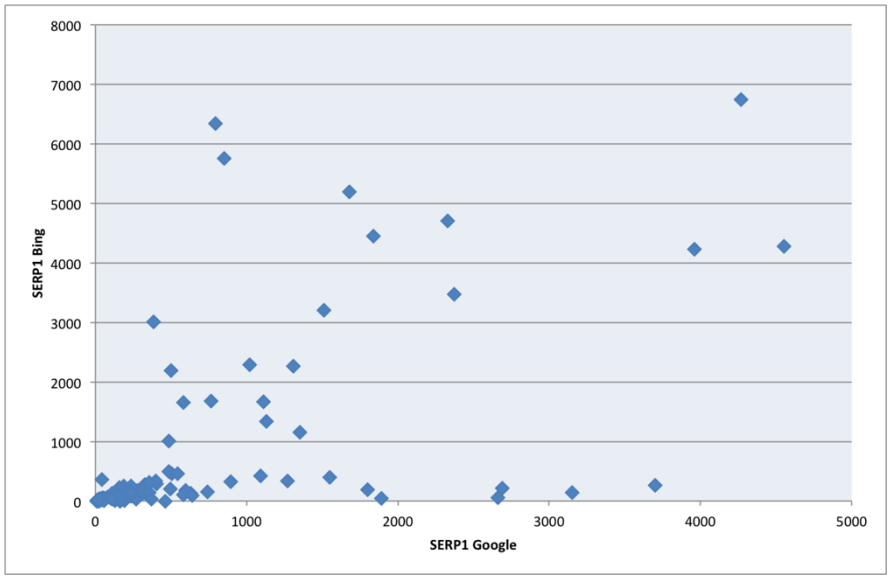


Figura 24: Gráfico de dispersión de Hits en Google y Bing.

De este modo se puede observar como la gran mayoría de asociaciones, se encuentran en el rango de entre 1000 y 500 hits en los buscadores, lo que implica de nuevo que su tamaño no es el adecuado. Así mismo, al ser los valores muy altos, esto implica que la dispersión es alta para valores muy altos, indicando que sus valores son muy diferentes, por ello es necesario medir en los dos buscadores.

En la siguiente figura 24 se puede observar otro gráfico de dispersión, en esta ocasión comparando las menciones de los portales, en los dos motores. En esta ocasión se puede observar como existe una menor dispersión, agrupándose la mayoría de los resultados en los 2.000.

En este caso los ejes son representados de la misma forma. Debe indicarse que en ambos casos se han eliminado los valores extremo que impedían la correcta visualización del conjunto de dispersión.

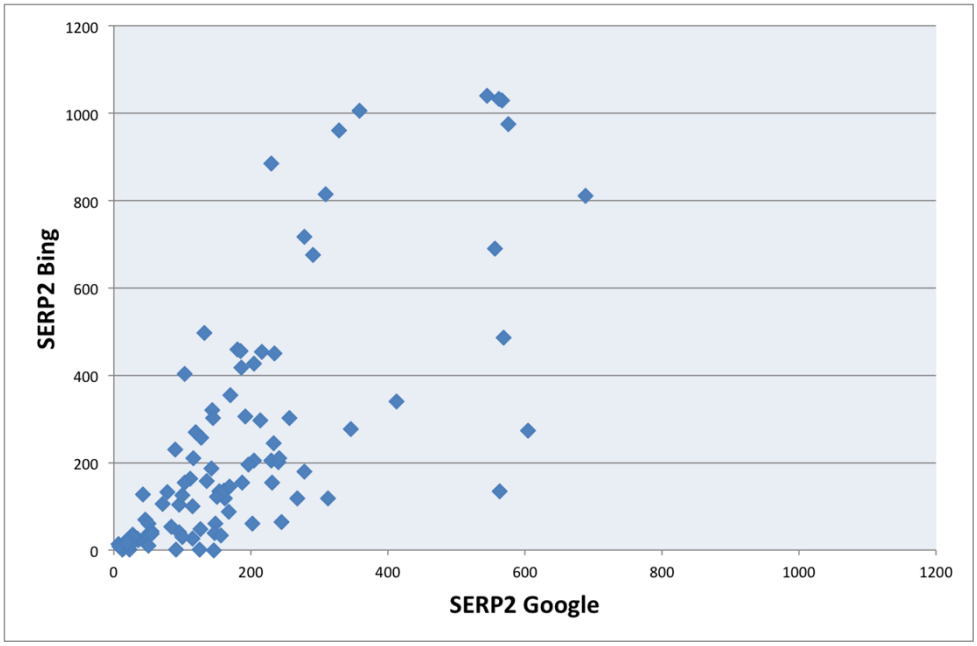


Figura 25: Gráfico de dispersión menciones en Google y Bing.

4.2 Factor de impacto.

El Factor de Impacto obtenido para los dos motores de búsqueda es el siguiente:

Tabla 19: Resultados Factor de Impacto

Nº	URL	WIF Google	WIF Bing
1	abaimar.es	3,40	5,76
2	adisen.es	1,34	6,88
3	aeomc.blogspot.com.es	3,05	0,01
4	afenmva.org	14,33	24,74
5	alcelyme.org	3,87	356000,00
6	amerenfermedadesraras.blogspot.com.es	7,59	10,68
7	amoimadrid.org	2,21	15,26
8	asfema.org	1,09	1,36
9	asociacion-nen.org	0,45	0,01
10	asociaciondoce.com	0,30	0,00
11	aspanion.es	22,04	6,44
12	associaciocatalanaataxies.blogspot.com	0,91	58,67
13	cadasil.org	0,08	1285,00
14	cancerinfantil.org	9,41	67,79

15	enachasociacion.blogspot.com	12,00	24,58
16	esclerodermia-adec.net	1,38	0,02
17	fcmpf.entitatsbcn.net	25,00	10900,00
18	fedaes.org	1,51	55,73
19	fedhemo.com	1,88	71,77
20	fundacionpielsana.es	2,00	1,25
21	galactosemiaespana.wordpress.com	1,09	24,56
22	https://sites.google.com/site/adelprisesevilla	0,23	0,08
23	ictiosis.org	2,23	7,96
24	infoame.org	2,05	8,83
25	luchando-contra-el-lyme.webnode.es	8,04	7,82
26	lupusgalicia.org	1,73	73,28
27	metabolicos.es	2,47	50,24
28	aacic.org	2,60	0,00
29	acadip.org	0,80	23,51
30	accucatalunya.cat	5,01	3,38
31	accuesp.com	3,24	59,27
32	acnefi.org	2,70	0,85
33	aeal.es	2,99	35,62
34	aeasarcomas.org	1,34	14,13
35	aecc.es	0,63	106,74
36	aedip.com	9,46	3000,00
37	aefat.es	5,04	15030,30
38	aepmi.org	13,22	109,60
39	aeracmeim.org	1,63	0,13
40	ahedysia.org	1,64	6,27
41	ahuce.org	2,74	0,00
42	alal.es	4,72	75609,76
43	amapar.org	4,12	105,84
44	anda.com.es	4,04	29900,00
45	aniridia.es	3,13	2201,55
46	aodem.com	1,31	0,00
47	apacor.org	11,44	1555,17
48	apture.org	7,95	1,69
49	asanhemo.org	4,72	0,00
50	asate.es	2,04	1653,70
51	asem-esp.org	1,17	1,03
52	asencordoba.org	25,13	21250,00
53	asense-a.org	0,92	1,14
54	asexve.es	9,09	3209302,33
55	asfape.org	25,43	12,31
56	ashemadrid.org	23,01	0,02
57	ashua.es	15,64	6,23

58	asion.org	5,38	0,03
59	asociacionauvea.es	5,16	0,32
60	asociacionbene.com	0,99	23,96
61	asociacioncrecer.org	2,29	0,98
62	ataxiasandalucia.org	0,03	11,59
63	cnio.es	2,14	0,00
64	corazonyvida.org	1,42	2,49
65	dedines.es	36,63	218,24
66	dgenes.es	4,06	4,35
67	displasiaectodermica.org	47,04	46666,67
68	duchenne-spain.org	1,69	17,08
69	eavacyl.com	0,43	0,56
70	enfermedades-raras.org	0,40	1,53
71	extrofia.info/asafex/	1,10	0,08
72	fchp.es	2,85	24,36
73	fegerec.es	1,13	4,95
74	felupus.org	1,54	43,18
75	fesorcam.org	1,20	1,22
76	fgcasal.org	1,78	1,59
77	fiapas.es	2,91	2,66
78	fibrosisquistica.org	3,50	0,00
79	fmfspain.com	1,03	644,12
80	fqmadrid.org	0,99	4,88
81	fundacionalpe.org	2,10	8,07
82	fundacionrebeccadealba.org	4,83	4,55
83	grupogeis.org	2,08	56,40
84	hipertensionpulmonar.es	3,11	40,08
85	limfacat.org	10,10	55,45
86	lupusvalencia.org	6,36	35,00
87	miasteniagravis.es	1,77	3,56
88	mpsesp.org	0,44	33,73
89	neurofibromatosis.es	1,82	45,77
90	noonanasturias.com	1,40	11,37
91	porfiria.org	5,02	0,03
92	postpolioinfor.org	4,15	22,86
93	postpoliomadrid.org	6,26	3,45
94	raoulfol.com	2,10	0,07
95	retina.es	1,20	23943,66
96	retinacv.es	1,13	1,97
97	retinosisfarpe.org	1,79	9,61
98	sarcoidosis.es	1,76	94,17
99	sediabetes.org	3,90	3,04
100	syndrom.org	9,50	1,89

Como puede observarse existen Factores de Impacto muy dispares, en la siguiente tabla se puede observar el mínimo, máximo y la media por motor de búsqueda:

Tabla 20: Resumen WIF mínimo, máximo y promedio.

	Google	Bing
WIF Mínimo	0,03	0,001*
WIF Máximo	47,04	3209302,33
WIF Promedio	5,25	38391,40

*Existe un valor nulo que en la tabla general se muestra como 0, por lo que el valor mínimo es el indicado en la Tabla 20.

En este caso, en Google si se obtienen unos resultados que permiten visualizar el impacto de los portales asociados con Enfermedades Raras. El 90% de los valores se encuentra por debajo del 10 de impacto, por lo que indica que su impacto no es muy bueno y debería de mejorar de cara a poder ayudar con la difusión de los contenidos.

En el caso de Bing, los resultados son más dispares, existe una gran cantidad de datos que no corresponden exactamente con un número normal de impacto web. Los elevados resultados se deben a que los hits recogidos como número total de links son muy elevados.

Esto puede ser debido a que Bing no sólo devuelve los resultados que realmente enlazan al contenido sobre el que se está buscando, si no que también muestran información relacionada o que el motor considera que puede ser de utilidad al usuario, por lo que realmente este indicador no parece ser realmente válido para la mayoría de asociaciones analizadas con este motor de búsqueda.

4.3 Correlación de Spearman

En la Tabla 21 se pueden observar los datos recogidos de SERPs, enlaces únicos y consultas a los motores con las URLs de las asociaciones, tras aplicar el coeficiente de correlación de Spearman.

Tabla 21: Coeficiente de Spearman sobre datos recogidos

	$SERP_1$ G	$SERP_1$ B	$SERP_n$ G	$SERP_n$ B	Únicos
$SERP_1$ G	1	0,785	0,955	0,783	0,939
$SERP_1$B		1	0,794	0,756	0,848
$SERP_n$ G			1	0,790	0,907
$SERP_n$ B				1	0,857
Únicos					1

Se puede observar como los resultados de $SERP_1$ entre los dos buscadores tienen una valor de 0,78, lo que implica que su correlación es alta.

Lo mismo sucede con el valor de $SERP_n$, su valor es de 0,79. Muy similar al anterior, por lo que los dos resultados varían de la misma manera.

La correlación que existe entre las búsquedas de Google, se encuentra fuertemente correlacionada, por lo que sus resultados serán los mejores ya que casi se acerca a 1 el valor.

Los resultados son todos superiores a 0,5, esto supone que la fuerza de la correlación también es alta. Esto podría significar que se pueden utilizar indistintamente ambos buscadores, pero lo cierto, es que tanto en valores únicos como en SERP1 la correlación de Google es mejor.

4.4 Enlaces externos:

De los datos recogidos en Majestic, existen varios índices que permiten conocer la visibilidad de un dominio.

- *CitationFlow*: En este caso el valor mínimo es de 0 y el máximo de 58, pese a que no se corresponde, los valores son muy similares a los calculados anteriormente con el WIF.

- *Enlaces de referencia Externos*: tras analizar los resultados, únicamente 25 portales se encuentran enlazados por lugares de referencia. Lo que implica una menor visibilidad.

- *Alexa Linking In:*
  - El 33% de los resultados tiene menos de 10 portales enlazando.
  - El 43 % de los resultados tiene entre 50 y 11 portales enlazando.
  - El 13% de los resultados tiene entre 51 y 99 portales enlazando.

- El 11% de los resultados tiene más de 100 portales enlazando, siendo el máximo 1026.

Gracias a esto se puede observar como la visibilidad de los portales web de las Asociaciones no es bueno.

4.4.1 Dr. Google

Un resultado obtenido transversalmente en la investigación, es el relativo a la cantidad de tarjeta que se encuentran en Google, con información sobre diferentes enfermedades.

La Araña programada extrae la información relativa a si la tarjeta de datos existe o no como resultado de la búsqueda. Los resultados obtenidos indican que el 1% del total del listado de las Enfermedades Raras tienen una tarjeta creada en Google.

Al cruzar los datos con el Top 50 de enfermedades seleccionadas para realizar el estudio, 40 de ellas tienen tarjeta. Debido a que las tarjetas de Google se van generando en función de las búsquedas que recibe el motor de búsqueda, se concluye que el Ranking de interés en Enfermedades generado en la Fase 1 del presente proyecto se encuentra acorde con el interés que demuestran los usuarios.

5 Conclusiones

Las Enfermedades Raras, aunque no todas en la misma medida, suscitan interés dentro de la Web. De las enfermedades analizadas, el número de documentos online en los que se menciona la enfermedad se encuentra en 52 Millones de entradas en Google para la enfermedad de MODY, y las 15 enfermedades con Hits más altos cuentan con más de un millón de resultado cada una. El resto de las enfermedades analizadas, poseen todas más de 100.000 resultados. Por lo que su presencia en Internet es significativa teniendo en cuenta el número de menciones.

En cambio, como se comenta en el apartado 3.3.1, la búsqueda de asociaciones relacionadas con las mismas no resulta sencilla. Los resultados ofrecidos por Google en este aspecto no ofrecían toda la información que se necesitaba para el estudio, por lo que es necesario buscar otras fuentes de información. Estas fuentes, a su vez, no ofrecen toda la información completa, es obligatorio buscar en diferentes portales para recabar toda la que se encuentre disponible y después unificarla, con el consecuente tiempo y esfuerzo que esto conlleva.

Por lo tanto existe una asimetría clara entre la atención que reciben las Enfermedades Raras en la Web, medidas mediante el número de menciones anteriormente indicado, y la visibilidad de la asociaciones que tratan de las mismas.

Como resultado de ese esfuerzo, este trabajo ofrece un listado completo, posiblemente el más exhaustivo en la actualidad, de 438 asociaciones, con sus nombres, siglas (en caso de que la asociación las tenga) y URLs, agrupando todas las que se encuentran disponibles y que, como trabajo futuro, se podría publicar como un directorio o portal web interactivo, de modo que sea mucho más fácil y sencillo acceder a dicha información. Por lo que el presente trabajo aporta una solución a un problema que existe actualmente y ayuda a mejorar la visibilidad de las asociaciones y, de forma indirecta, de la información sobre las enfermedades raras.

En relación con los portales analizados, los que más presencia tienen se encuentran relacionados con diferentes tipos de enfermedades, desde el cáncer, las enfermedades raras en niños a problemas de piel u ojos. Existiendo 25 resultados con más de 1.000 hits, lo que implica que sus resultados son reseñables.

En cambio, las otras 75 asociaciones restantes cuentan con menos de 1.000 hits, siendo un 25% del total las que cuentan con menos de 100. Esto implica que su presencia es mínima. Suponiendo que el contenido indizado por los buscadores sea como mínimo superior al 90% del contenido total, una conclusión que puede extraerse es que la cantidad de contenidos creados por dichas asociaciones es muy baja, lo que podría explicar la carencia de visibilidad detectada anteriormente.

Si tomamos el ejemplo de la Enfermedad Rara "Uveitis", una asociación relacionada con la misma es la Asociación Retina Comunidad Valenciana, que se encuentra en el puesto 50 dentro del ranking creado por las 100 analizadas en cuanto a visibilidad. La enfermedad tiene 1.950.000 resultados en Google, en cambio la asociación tiene 256 hits en Google, y únicamente se encuentra enlazada por 22 sitios según Alexa. Esto supone que tanto su presencia, como su visibilidad son bajas.

Teniendo en cuenta el *Web Impact Factor* como métrica de impacto, y utilizando los datos de Google por los problemas encontrados con Bing que se detallan en el anterior capitulo, el impacto máximo es de 47 puntos y el valor medio de 5,25, esto indica que el impacto de las webs es mínimo, llegando algunos casos a ser prácticamente nulo (menos el 20% tienen menos de 1 punto).

Por otro lado, se detecta una dependencia del instrumento de medida, los motores de búsqueda, en la obtención de resultados. No obstante la correlación obtenida entre los resultados de Google y Bing es relativamente alta.

Para llegar a esta conclusión se utilizan los resultados obtenidos en los dos buscadores, tanto el indicador de resultados inicial ($SERP_1$) como el final ($SERP_n$). Al comparar los resultados, primero por buscador y $SERP_1$, y después con $SERP_n$, los resultados superan el 0,7 de correlación, lo que indica que la correlación es fuerte.

En el caso de la correlación existente entre los $SERP_1$ y $SERP_n$ de un mismo buscador, queda patente que los mejores resultados son los obtenidos por Google, con una correlación de r=0,95 entre los resultados que indica que tiene y los que muestra al final.

Es interesante indicar, que en relación con los documentos únicos, es decir URLs extraídas en ambos motores de búsqueda por cada una de las asociaciones, la correlación de Google es mayor que la de Bing, esto indica que

Google es más estable generando los dos tipos de resultados, ofreciendo mejores resultados relacionados con las asociaciones de Enfermedades Raras.

Por lo tanto, si unimos que la mejor correlación entre $SERP_1$ y $SERP_n$ y entre $SERP_1$ y resultados únicos la obtiene Google, éste es el mejor motor de búsqueda para tratar de extraer la mayor cantidad de información.

Pese a que la correlación es alta entre los dos buscadores, cabe recordar que existen valores atípicos importantes, por lo que los datos deben ser tomados con cautela ya que es importante conocer el por qué de la existencia de estas anomalías para poder llegar a conclusiones absolutas.

Como conclusión final, queda patente que del mismo modo que las Enfermedades Raras son raras, su visibilidad y presencia en la Web también lo es. Esto debería ser tenido en cuenta para mejorarlo y que cualquier persona cuando necesite información acerca de las mismas, pueda encontrarla de la forma más sencilla posible. Aunque para ello se precisa de varios elementos para resolver el problema:

- Por una parte, formar y concienciar a las asociaciones para optimizar la información que ofrecen, de modo que al desarrollarla se posicionen mejor en la Web.
- Y por otra parte, comprender mejor el instrumento de medida para poder mejorar el sistema de medición, haciendo más precisa, rápida, económica y fiable la obtención del impacto web.

Tal y como se indica en el apartado 1.2 , el presente trabajo de fin de máster trata de analizar la presencia y visibilidad web de los portales relacionados con enfermedades raras mediante la cuantificación de diferencias en el tratamiento de los portales por parte de diferentes motores de búsqueda como objetivo principal. Una vez recogida y analizada toda la información necesaria para este proyecto, puede concluirse que el objetivo principal se ha alcanzado satisfactoriamente.

Del mismo modo, en cuanto a los objetivos secundarios se concluye que:

- Se ha analizado y medido la cantidad de información existente actualmente sobre las enfermedades raras en internet, de modo que se ha logrado obtener un listado con las Enfermedades Raras que generan mayor interés en la red.

- Se han estudiado y seleccionado los dominios relacionados con las asociaciones sobre Enfermedades Raras disponibles, logrando obtener un listado con todas las asociaciones disponibles e información sobre las mismas.

- Se ha diseñado y programado diferentes arañas específicas para la obtención de los indicadores métricos utilizados de forma automática, precisa y gratuita.

- Se ha medido el tamaño de los dominios seleccionados para el análisis, incluyendo los diferentes tipos de documentos e información, habiendo podido conocer y contextualizar la presencia de los mismos en la Web en función de diversos buscadores.

- Se ha realizado un análisis relativo al posicionamiento y tratamiento de los resultados relacionados con las Enfermedades Raras, cuyos resultados se encuentran en el Capitulo 5.

- Se ha propuesto un modelo de análisis mediante cuantificación para diferentes buscadores web.

- Se ha elaborado un directorio compuesto por los 438 resultados obtenidos tras el estudio y selección de asociaciones sobre Enfermedades Raras.

Debido a que el análisis Cibermétrico llega a una profundidad mucho mayor de la alcanzada en este trabajo, como trabajo futuro se plantea como reto principal, lograr mediante métodos mejorados una mayor, más rápida y automatizada recogida de datos, que permita realizar el análisis de un modo más sencillo que el presentado.

Otras ideas que pueden ser tenidas en cuenta son:

- Buscar y aplicar nuevos indicadores relacionados con la presencia y la visibilidad web.

- Expandir el trabajo, realizando un estudio temporal para seguir el comportamiento y comprobar si el problema de indización que da valores de $SERP_n$ mayores que $SERP_1$ se estabiliza o si persiste en el tiempo, para entender mejor el funcionamiento y diferencias significativas encontradas en los motores de búsqueda.

- Ampliar el trabajo incorporando medidas sobre infraestructuras, para comprobar la implicación que tienen sobre los resultados obtenidos.

- Generar una versión web con los resultados del presente proyecto, para que puedan ser alcanzados por todas aquellas personas interesadas en los dos temas principales del trabajo: la Cibermetría y las Enfermedades Raras.

- Repetir el presente estudio, ampliando la muestra a nivel Europeo / mundial, para poder conocer el estado del mismo en otros países, de modo que se puedan intercambiar ideas y mejorar el proyecto.

- Estudiar las conexiones existentes entre las propias webs y redes de las asociaciones, de modo que se puedan entender la comunicación y relación que tienen para tratar de mejorar su visibilidad.

6 Bibliografía

A Castillo, P. L. (2015). La comunicación en la red de pacientes con enfermedades raras en España. *Revista Latina de Comunicación Social* (70), 673 a 688.

Aguillo, I. (de de 2000). Internet invisible o infranet: definición, clasificación y evaluación. (J. E. Documentación, Interviewer)

Arroyo, O. P. (2005). Cibermetría. Estado de la cuestión. *9as Jornadas Españolas de Documentación, FESABID* (p. 14). Madrid: FESABID.

Bar-Ilan, J. (03 de 2005). . *Comparing rankings of search engines resultos on the Web* . Jerusalem, Israel: Elsevier.

Bar-Ilan, J., Mat-Hassan, M., & Levene, M. (de 12 de 2005). Methods for comparing rankings of search engine results. *Methods for comparing rankings of search engine results* . Londres, Londres, UK: Elsevier.

Basagoiti, I. (2011). Compartir información sanitaria. ePacientes: comunicación e interacción. *El ePaciente y las Redes Sociales* .

BBVA, F. (2008). *Internet en España*. Fundación BBVA.

Belt, T. H. (2010). *Definition of Health 2.0 and Medicine 2.0: A Systematic Review*. Nijmegen: JMIR Publications.

Bergman, M. K. (2011). The Deep Web: Surfacing hidden value. *Web-based Information Retrieval* , .

Björneborn, L. (2004). *Small-World Link Structures across an Academic Web Space: A Library and Information Science Approach*. Denmark: Royal School of Library and Information Science.

Ciberer. (07 de 07 de 2016). *Ciberer*. Retrieved 07 de 07 de 2016 from Ciberer: http://www.ciberer.es/quienes-somos

CSIC. (02 de 2003). *Cibermetría: Introducción teórico-práctica*. Retrieved 12 de 08 de 2016 from CSIC: http://docplayer.es/3213026-Cibermetria-introduccion-teorico-practica-a-una-disciplina-emergente.html

Definicion.de. (01 de 06 de 2016). *definicion.de*. From definicion.de: http://definicion.de/web-2-0/

Europea, C. (25 de 07 de 2016). *Comisión Europea - Salud Pública*. Retrieved 25 de '07 de 2016 from Comisión Europea - Salud Pública: http://ec.europa.eu/health/rare_diseases/european_reference_networks/index_es.htm

Europea, C. (01 de 01 de 2016). *Redes de Referencia - CE*. Retrieved 29 de 04 de 2016 from Comisión Europea - Salud Pública: http://ec.europa.eu/health/rare_diseases/european_reference_networks/index_es.htm

EURORDIS. (18 de 02 de 2012). *EURORDIS - ¿Qué es una enfermedad rara?* Retrieved 14 de 04 de 2016 from EURORDIS: http://www.eurordis.org/es/content/%C2%BFque-es-una-enfermedad-rara

FEDER. (01 de 04 de 2016). *FEDER*. Retrieved 01 de 04 de 2016 from FEDER: http://www.enfermedades-raras.org/

Giustini, D. (2008). How Web 2.0 is changing medicine. *BMJ* , 1279-1284.

Goldman, E. (01 de 01 de 2006). *Search Engine Bias and the Demise of Search Engine Utopianism*. Retrieved de de from digitalcommons.law.scu.edu: digitalcommons.law.scu.edu/facpubs/76

Google. (de 08 de 2016). *Búsquedas médicas en Google*. Retrieved de 08 de 2016 from Google: https://support.google.com/websearch/answer/2364942?p=medical_conditions&rd=1&hl=es

Google. (de de 2016). *Operadores de búsqueda*. Retrieved de 08 de 2016 from Operadores de búsqueda: https://support.google.com/websearch/answer/2466433?hl=es

Groselj, D. (2013). A webometric analysis of online health information: sponsorship, platform type and link structures. *Emerald Insight* , 209.

INE. (2014). *Encuesta sobre Equipamiento y Uso de las TIC en los hogares 2014*. Instituto Nacional de Estadística.

Internet, P. (07 de 10 de 2007). *Pew Internet*. Retrieved 26 de 07 de 2016 from Pew Internet: http://www.pewinternet.org/2007/10/08/e-patients-with-a-disability-or-chronic-disease/

Lapuente, M. J. (2013). Hipertexto, el nuevo concepto de documento en la cultura de la imagen. *Universidad Complutense de Madrid* , 184.

Leanne Bowler, W.-Y. H. (2011). The visibility of health web portals for teens: a hyperlink analysis". *Online information Review , 35* (3), 443.

Leydesdorff, L. (2012). Scientometrics. *arxiv.org* , 1.

Lyman, P. (2003). How much Information? *Universidad de Carolina* , .

Macías-Chapula, C. A. (2001). Papel de la Cibermetría y la cienciomatría y su perspectiva nacional e internacional. *ACIMED* , 4.

Noruzi, A. (de de 2006). The Web Impact Factor: a critical review. *The Web Impact Factor: a critical review* . Tehran, , Iran: The electronic library.

Ochando, M. B. (26 de 11 de 2012). *Sistemas de recuperación e Internet*. Retrieved de 08 de 2016 from Webmetría y análisis de páginas web: http://ccdoc-sistemasrecuperacioninternet.blogspot.com.es/2012/11/webmetria-y-analisis-de-paginas-web.html

ONTSI. (2016). *Los ciudadanos ante la e-Sanidad*. España: red.es.

Orduña-Malea, E. (11 de 2011). Propuesta de un modelo de análisis redinformétrico multinivel para el estudio sistémico de las universidades españolas (2010). *Propuesta de un modelo de análisis redinformétrico multinivel para el estudio sistémico de las universidades españolas (2010)* . Valencia, Valencia, España: Universidad Politécnica de Valencia.

Orduña-Malea, E., & Aguillo, I. (2014). *Cibermetría: midiendo el espacio red*. Barcelona: Editorial UOC.

O'Reilly. (30 de 09 de 2005). *What is web 2.0*. Retrieved 20 de 07 de 2016 from oreilly: http://www.oreilly.com/pub/a/web2/archive/what-is-web-20.html

O'Reilly, T. (2007). What is web 2.0: Design patterns and business models for the next generation of software. *Communications & Strategies* (65), 17.

Orpha.net. (24 de 10 de 2012). *Orpha.net*. Retrieved 14 de 04 de 2016 from Portal de información de enfermedades raras y medicamentos huérfanos: http://www.orpha.net/consor/cgi-bin/Education_AboutRareDiseases.php?lng=ES

Portal, S. (25 de 09 de 2001). *Statistic Portal*. Retrieved 08 de 08 de 2016 from Statistis Portal: http://oxforddictionaries.com/definition/english/bibliometrics

Python. (de de 2016). *Python*. Retrieved de de from Python: https://www.python.org/doc/essays/blurb/

Rodriguez, A. M. (12 de 09 de 2006). Indicadores cibermétricos: ¿Nuevas propuestas para medir la información en el entorno digital? La Habana, , : .

Salcedo, V. T. (2011). *El ePaciente y las redes sociales*. España: ITACA.

Salud, O. M. (2012). Boletín de la Organización Mundial de la Salud. In O. M. Salud, *Boletín de la Organización Mundial de la Salud* (Vol. 90, pp. 401-476). OMS: OMS.

SearchEngineHistory. (de de 2016). *SearchEngineHistory*. Retrieved 2016 from : http://www.searchenginehistory.com/

Sen, B. K. (2004). Cybermetrics - meaning, definition, scope and constituents. *Annals of Library and Information Studies* , 116.

Smih, A. (de de 1999). A Tale of Two Web Spaces: Comparing Sites Using Web Impact Factors. *A Tale of Two Web Spaces: Comparing Sites Using Web Impact Factors , 55 (5)* . , , : Journal of documeentation.

Thelwall, M., & Stuart, D. (de 11 de 2006). Web crawling Ethics revisited: cost, privacy, and denial of service. *Web crawling Ethics revisited: cost, privacy, and denial of service* . Wolverhampton, , UK: Journal of the American Society for Information Science and Technology.

Thelwall, M., & Sud, P. (15 de 03 de 2011). A comparison of methods for collecting web citation data for academic organizations. *A comparison of methods for collecting web citation data for academic organizations* . Wolverhampton, Wolverhampton, UK: Journal fo the American Society for Information Science and Technology.

Unidos, 1. C. (06 de 11 de 2002). Rare Diseases Act of 2002. *Rare Diseases Act of 2002* . Washington D.C., Columbia, Estados Unidos: Office of Rare Diseases.

Vaughan, L., & Thelwall, M. (09 de 08 de 2003). Search engine coverage bias: evidence and possible causes. *Information and processing and management* .

W3C. (de de 1997). *W3C Html*. Retrieved 03 de 08 de 2016 from W3C: https://www.w3.org/TR/REC-html40-971218/cover.html#toc

Wikipedia. (01 de 01 de 2016). *Wikipedia*. Retrieved 26 de 07 de 2016 from Wikipedia: https://es.wikipedia.org/wiki/Motor_de_b%C3%BAsqueda

Wikipedia. (01 de 01 de 2016). *Wikipedia*. Retrieved 12 de 06 de 2016 from Wikipedia: https://en.wikipedia.org/wiki/Google_Search

Wolfram, D. (2000). Applications of Informetrics to Information Retrieval Research. *Information Science Research* , 77.

World, B. (19 de 08 de 2015). *Fundela Noticias*. Retrieved 02 de 05 de 2016 from Fundela: http://www.fundela.es/noticias/2015/segun-los-cientificos-el-reto-del-cubo-de-agua-ayuda-a-acelerar-un-descubrimiento-sobre-la-ela/

Anexo I: Listado Asociaciones

Siglas	Nombre	URL
AACIC	associació de cardiopaties congènites	aacic.org
ABAIMAR	Asociación de niños con enfermedades raras de Baleares	abaimar.es
ACADIP	ASSOCIACIÓ CATALANA DE DÈFICITS IMMUNITARIS PRIMARIS	acadip.org
ACCUcat	ACCU Cataluña	accucatalunya.cat
ACCU	ACCU España	accuesp.com
ACNEFI	Asociación Catalana de Neurofibromatosis	acnefi.org
ADISEN	Asociación Nacional de Addison y Otras Enfermedades Endocrinas	adisen.es
AEAL	Asociación Española de Afectados por Linfoma, Mieloma y Leucemia	aeal.es
AEASARCOMAS	Asociación Española de Afectados por Sarcomas	aeasarcomas.org
AECC	Asociación Española Contra el Cáncer	aecc.es
AEDIP	Asociación Española de Déficits Inmunitarios Primarios	aedip.com
AEFAT	Asociación Española Familiar Ataxia Telangiectasia	aefat.es
AEOMC	Asociación Española OsteoCondromas Múltiples Congénitos)	aeomc.blogspot.com.es
AEPMI	Asociación de Enfermos de Patologías Mitocondriales	aepmi.org
ACMEIM	Asociación de Enfermedades Raras	aeracmeim.org
AFENMVA	Asociación Familiares y Enfermos Neuromusculares de Valencia	afenmva.org
AHEDYSIA	Asociación Humanitaria de Enfermedades Degenerativas y Síndromes de la Infancia y Adolescencia	ahedysia.org
AHUCE	Asociación de Huesos de Cristal de España (SID)	ahuce.org
ALAL	Asociación de Autoinmunes y Lupus de Almería	alal.es
ALCE	Asociación Lyme Crónico España	alcelyme.org
AMAPAR	Asociación Madrileña de Personas con Artritis Reumatoide	amapar.org
AMER	AsociaciÃ³n Molinense de Enfermedades Raras	amerenfermedadesraras.blogspot.com.es
AMOI	Asociación Madrileña de Osteogonesis Imperfecta	amoimadrid.org
ANDA	Asociación de Niños con Discapacidad de Almería	anda.com.es

ANIRIDIA	Asociación Española de Aniridia	aniridia.es
AODEM	Asociación ourensana de esclerose múltiple, ELA, parkinson e outras enfermidades neurodexenerativas	aodem.com
APACOR	Asociación de Pacientes Coronarios	apacor.org
APTURE	Asociación de Pacientes con Tumores Raros de España	apture.org
ASANHEMO	Asociación Andaluza de Hemofilia	asanhemo.org
ASATE	Asociación de Afectados por Tumores Cerebrales en España	asate.es
FASEM	Federación Española de Enfermedades Neuromusculares	asem-esp.org
ASENCO	Asociación de Enfermedades Neuromusculares de Córdoba	asencordoba.org
ASENSE	Asociación de enfermedades Neuromusculares de Andalucía	asense-a.org
ASEXVE	Asociación Española de Extrofia Vesical	asexve.es
ASFEPE	ASOCIACION DE FAMILIAS CON PERTHES	asfape.org
ASFEMA	Asociación de PKU y OTM de Madrid	asfema.org
ASHEMADRID	Asociación de Hemofilia de la Comunidad de Madrid	ashemadrid.org
ASHUA	Asociación Síndrome Hemolítico Urémico Atípico	ashua.es
ASION	Asociación Oncológica de Madrid	asion.org
NEN	Asociación de Familiares y Amigos de Pacientes con Neuroblastoma	asociacion-nen.org
AUVEA	Asociación de Pacientes de Uveítis	asociacionauvea.es
BENE	Asociación de Enfermos Neuromusculares de Bizkaia	asociacionbene.com
CRECER	Asociación Nacional para Problemas de Crecimiento	asociacioncrecer.org
DOCE	Asociación Discapacitados Otros Ciegos de España	asociaciondoce.com
ASPANION	Asociación Niños con cáncer	aspanion.es
ACAH	Associació Catalana d'Atàxies Hererditàries	associaciocatalanaataxies.blogspot.com
	Federación Andaluza de Asociaciones de Ataxias	ataxiasandalucia.org
CADASIL	Asociación Cadasil España	cadasil.org
	Federación Española de Padres de Niños con	cancerinfantil.org

	Cáncer	
CNIO	Centro Nacional de Investigaciones Oncológicas	cnio.es
	Asociación Corazón y Vida de Sevilla	corazonyvida.org
DEDINES	Asociación Dedines	dedines.es
DGENES	Asociación Dgenes	dgenes.es
AADE	Asociación de Afectados por Displasia Ectodérmica	displasiaectodermica.org
DPPE	Asociación Duchenne Parent Project España	duchenne-spain.org
EAVACYL	Enfermedades autoinmunes Vasculitis CyL	eavacyl.com
ENACH	Asociación española Hallervorden Spatz	enachasociacion.blogspot.com
FEDER	Federación Española de Enfermedades Raras	enfermedades-raras.org
ADEC	Asociación de Esclerodermia de Castellón	esclerodermia-adec.net
ASAFEX	Asociación de personas y familias afectadas de extrofia vesical, cloacal, epispadias y patologías afines	extrofia.info/asafex/
FCHP	Fundación Contra la Hipertensión Pulmonar	fchp.es
FECAMM	Federació Catalana de Malalties Poc Freqüents	fcmpf.entitatsbcn.net
FEDAES	Federación de Ataxias de España	fedaes.org
FEDHEMO	Federación Española de Hemofilia	fedhemo.com
FEGEREC	Federación Gallega de Enfermedades Raras y Crónicas	fegerec.es
FELUPUS	Federación Española de Lupus	felupus.org
FESORCAM	Federación de Personas Sordas de la Comunidad de Madrid	fesorcam.org
FGCASAL	Fundación Gaspar Casal	fgcasal.org
FIAPAS	Confederación Española de Familias de Personas Sordas	fiapas.es
FEFQ	Federación Española de Fibrosis Quistica	fibrosisquistica.org
FMF	Asociación de enfermos de Fiebre Mediterránea Familiar de España	fmfspain.com
	Asociación Madrileña de Fibrosis Quística	fqmadrid.org
ALPE	Fundación ALPE Acondroplasia	fundacionalpe.org
	Fundación Piel Sana	fundacionpielsana.es

	Fundación Rebecca de Alba	fundacionrebeccadealba.org
	Asociación Española para la Galactosemia	galactosemiaespana.wordpress.com
GEIS	Grupo Español de Investigación en Sarcomas	grupogeis.org
	Asociación Nacional de Hipertensión Pulmonar	hipertensionpulmonar.es
ADELPRISE	Afectados por Linfedema Primario y Secundario de Sevilla	https://sites.google.com/site/adelprisesevilla
ASIC	Asociación Española de Ictiosis	ictiosis.org
AME	Asociacion Microtia España	infoame.org
LIMFACA	Associació catalana d'afectats de limfedema	limfacat.org
	Luchando contra el Lyme	luchando-contra-el-lyme.webnode.es
AGAL	Asociacion Galega De Lupus	lupusgalicia.org
AVALUS	Asociación Valenciana de Afectados de Lupus	lupusvalencia.org
	Federación Española de Enfermedades Metabólicas Hereditarias	metabolicos.es
	Asociación Miastenia de España	miasteniagravis.es
MPS	Asociación Española de Mucopolisacaridosis y Síndromes Relacionados	mpsesp.org
	Asociación de Afectados de Neurofibromatosis	neurofibromatosis.es
	Asociación Sindrome de Noonan Asturias	noonanasturias.com
	Asociación Nacional de Porfiria	porfiria.org
	Asociacion Afectados de Polio y Sindrome Post-Polio de España	postpolioinfor.org
	Asociación Postpolio Madrid	postpoliomadrid.org
	Asociación Amigos de los Leprosos-Raoul Follereau	raoulfol.com
	Fundación Retina España	retina.es
	Asociación de afectados por retinosis pigmentaria-Retinitis pigmentosa de la Comunidad Valenciana	retinacv.es
FARPE	Federación de Asociaciones de Retinosis Pigmentaria de España	retinosisfarpe.org
ANES	Asociación Nacional de enfermos de Sarcoidosis	sarcoidosis.es
	Sociedad Española de Diabetes	sediabetes.org
	Associació Syndrom	syndrom.org