TESIS DOCTORAL

Descubrimiento y evaluación de recursos web de calidad mediante Patent Link Analysis

Cristina I. Font Julián

Dirigida por:

Dr. José Antonio Ontalba y Ruipérez
Dr. Enrique Orduña Malea

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Departamento de Comunicación Audiovisual, Documentación e Historia del Arte

Valencia, junio 2021

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

DCADHA

DPTO. DE COMUNICACIÓN AUDIOVISUAL DOCUMENTACIÓN E HISTORIA DEL ARTE

La presente tesis doctoral ha sido financiada por el Gobierno de España mediante el contrato predoctoral para la formación de doctores FPI BES-2017-079741 otorgada por el Ministerio de Ciencia e Innovación.

A mis padres y mi hermana

"El último hombre que sabía cómo funcionaba probablemente había muerto en la sala de torturas hacía muchos años. O tan pronto como fue instalada. Matar al creador era un método tradicional de proteger la patente"

– Terry Pratchett
(Dioses menores)

Mi agradecimiento sincero de corazón a todos los que me han ayudado y apoyado a lo largo de todo el camino para que esta tesis pudiese ver la luz.

Cristina I. Font Julián

Sumario

Índice de figuras

Índice de tablas

Siglas y Abreviaturas

ARIPO: Organización Regional Africana de la Propiedad Intelectual
CIP: Clasificación Internacional de Patentes
EAPO: Organización Euroasiática de Patentes
EPO: European Patent Office
INID: Identificación Numérica Internacionalmente acordada en materia de Datos
OAPI: Organización Africana de la Propiedad Intelectual
OEP: Organización Europea de Patentes
OEPM: Oficina Española de Patentes y Marcas
OMPI: Organización Mundial de la Propiedad Intelectual
OP CCG: Oficina de Patentes del Consejo de Cooperación de los Estados Árabes del Golfo
PCT: Patent Cooperation Treaty (Tratado de Cooperación en materia de Patentes)
WIPO: World Intellectual Property Organization
USPTO: United States Patent and Trademark Office (Oficina de Patentes y Marcas de Estado Unidos)

Resumen

Español

Las patentes son documentos legales que describen el funcionamiento exacto de una invención, otorgando el derecho de explotación económica a sus dueños a cambio de dar a conocer a la sociedad los detalles de funcionamiento de dicha invención. Para que una patente pueda ser concedida debe cumplir tres requisitos: ser novedad (no haber sido expuesto o publicado con anterioridad), cumplir la actividad inventiva y tener aplicación industrial. Es por ello que las patentes son documentos valiosos, ya que contienen una gran cantidad de información técnica no incluida antes en otro tipo de documento (publicado o disponible).

Debido a las características particulares de las patentes, los recursos que éstas mencionan, así como los recursos que mencionan a las patentes, contienen enlaces que pueden ser útiles y dar apoyo a diversas aplicaciones (vigilancia tecnológica, desarrollo e innovación, Triple-Helix, etc.) al disponer de información complementaria, así como de la creación de herramientas y técnicas que permitan extraerlos y analizarlos.

Por este motivo, la presente tesis doctoral plantea la posibilidad de descubrir recursos web de calidad a través de los enlaces mediante el diseño de un método adecuado para su extracción y análisis eficiente aplicando el uso de la técnica cibermétrica de análisis de enlaces.

Para ello, se ha descrito en el estado de la cuestión dos grandes bloques: las patentes y la cibermetría, con el objetivo de ofrecer un marco de trabajo que permita entender correctamente el alcance del análisis. Tras la exposición teórica del campo de estudio, se realiza una explicación del método propuesto para alcanzar los objetivos que definen la tesis. Este método se encuentra divido en dos bloques complementarios: Patent Outlink y Patent Inlink, que juntos conforman la técnica de Patent Link Analysis. Ambos bloques siguen un proceso de trabajo equivalente. En primer lugar, se selecciona la fuente de la cual extraer los datos, tras esto se recopilan aquellos necesarios y se almacenan en bruto con doble copia para evitar pérdidas de información, de este modo es posible recuperar en todo momento la información inicial.

Se analizan los datos brutos para entender el tipo de información y estructura en la que se encuentran y que permita preparar un sistema de extracción específico para

cada bloque. En el caso del Bloque Patent Outlink, se desarrolla un programa ad hoc para la extracción de enlaces contenidos en los documentos en patentes, que permite extraer la mayor cantidad posible de enlaces con el mínimo ruido, y recopilando a su vez información relacionada con la sección de aparición o la categoría de la patente de modo que se pueda realizar un análisis en mayor profundidad.

Para realizar el estudio se selecciona la Oficina de Patentes y Marcas de Estados Unidos (USPTO), recogiendo todas aquellas patentes concedidas entre los años 2008 y 2018 (ambos incluidos). Una vez extraída la información a analizar en cada bloque se cuenta con: 3.133.247 de patentes, 2.745.973 millones de enlaces contenidos en patentes, 2.297.366 millones de páginas web de patentes enlazadas, 17.001 paginas únicas web enlazando a patentes y 990.663 patentes únicas enlazadas desde documentos web.

Los resultados del análisis de Patent Outlink muestran como tanto la cantidad de patentes que contienen enlaces (20%), como el número de enlaces contenido en patentes (mediana 4-5) es todavía bajo, pero ha crecido significativamente durante los últimos años y se puede esperar un mayor uso en el futuro. Existe una diferencia clara en el uso de enlaces entre áreas de conocimiento (42% pertenecen a Física, especialmente Computación y Cálculos), así como por secciones dentro de los documentos, explicando los resultados obtenidos y la proyección de análisis futuros.

Los resultados del análisis de Patent Inlink identifica una cantidad considerable menor de dominios webs que enlazan a patentes (17.001 frente a 256.724), pero existen más enlaces por documento enlazante (el número de enlaces total es similar para ambos bloques de análisis). Así mismo, los datos muestran una elevada dispersión, con unos pocos dominios generando una gran cantidad de enlaces. Ambos bloques muestran la existencia de una alta relación con empresas y servicios tecnológicos, existiendo diferencias relativas a los enlaces a Universidades y Gobiernos (más enlaces en Outlink).

Los resultados muestran que el modelo de análisis propuesto permite y facilita el descubrimiento y evaluación de recursos web de calidad. Así mismo, se concluye que la cibermetría, mediante el análisis de enlaces, aporta información de interés para el análisis de los recursos web de calidad a través de los enlaces contenidos y dirigidos a documentos de patentes.

El método propuesto y validado permite de un modo eficiente, eficaz y replicable la extracción y análisis de los enlaces contenidos y dirigidos a documentos de patentes. Permitiendo, a su vez, definir, modelar y caracterizar el Patent Link Analysis como un subgénero del Link Analysis que puede ser utilizado para la construcción de sistemas de monitorización de link intelligence, de evaluación y/o de calidad entre otros, mediante el uso de los enlaces entrantes y salientes de documentos de patentes aplicable a universidades, centros de investigación, así como empresas públicas y privadas.

Inglés

Patents are legal documents that describe the exact operation of an invention, granting the right of economic exploitation to its owners in exchange for describing the details of the operation of said invention. For a patent to be granted, it must meet three requirements: be novel (not have been previously exhibited or published), comply with the inventive step and have industrial application. That is why patents are valuable documents, since they contain a large amount of technical information not previously included in another type of document (published or available).

Due to the particular characteristics of patents, the resources that they mention, as well as the resources that mention patents, links contained can be useful and give support to various applications (technological surveillance, development and innovation, Triple-Helix, etc.) by having complementary information, as well as the creation of tools and techniques that allow them to be extracted and analyzed.

For this reason, this doctoral thesis raises the possibility of discovering quality web resources through links by designing a suitable method for their extraction and efficient analysis that applies the use of the cybermetric technique of link analysis.

For this, two large blocks have been described in the state of the art: patents and cybermetrics, with the aim of offering a framework that allows a correct understanding of the scope of the analysis.

After the presentation of the field of study, an explanation of the proposed method is made to achieve the objectives that define the thesis. This method is divided into two complementary blocks: Patent Outlink and Patent Inlink, which together make up the Patent Link Analysis technique. An equivalent work process is followed for both blocks. First, the source from which to extract the data is selected, after this the necessary data is collected and stored raw with double copies to avoid loss of information, in this way it is possible to recover the initial information at any time.

The raw data is analysed to understand the type of information and structure to prepare a specific extraction system for each block. In the case of the Patent Outlink Block, an ad hoc program is developed for the extraction of links contained in patent documents, which allows the extraction of as many links as possible with minimal noise, also collecting information related to the section of appearance, and category of the patent in order to perform a more in-depth analysis.

To carry out the study, the United States Patent and Trademark Office (USPTO) is selected, collecting all patents granted between 2008 and 2018 (both included). Once the information to be analyzed has been extracted in each block, there are: 3,133,247 patents, 2,745,973 million links contained in patents, 2,297,366 million linked patent web pages, 17,001 unique web pages linking patents and 990,663 Unique patents linked from web documents.

The results of the Patent Outlink analysis show that both the number of patents that contain links (20%) and the number of links contained in patents (median 4-5) is still low but has grown significantly in recent years, and it can be expected more use in the future. There is a clear difference in the use of links between areas of knowledge (42% belong to Physics, especially Computing and Calculus), as well as by sections within the documents, explaining the results obtained and the projection of future analyzes.

The Patent Inkinks analysis results identify considerably fewer web domains linking to patents (17,001 vs. 256,724), but there are more links per linking document (the total number of links is similar for both analysis blocks). Likewise, the data shows a high dispersion, with a few domains generating a large number of links. Both blocks show the existence of a high relationship with companies and technological services, with differences relating to Universities and Governments (more links in Outlink).

The results show that the proposed analysis model allows and facilitates the discovery and evaluation of quality web resources. Similarly, it is concluded that cybermetrics, through the analysis of links, provides information of interest for the analysis of quality web resources through the links contained and directed to patent documents

The method proposed and validated in this thesis allows in an efficient, effective and replicable way the extraction and analysis of the links contained and directed to patent documents. Allowing, in turn, to define, model and characterize the Patent Link Analysis as a subgenre of the Link Analysis that can be used for the construction of link intelligence monitoring systems, evaluation and/or quality among others, through the use of the incoming and outgoing links of patent documents applicable to universities, research centers, as well as public and private companies.

Valenciano

Les patents són documents legals que descriuen el funcionament exacte d'una invenció, atorgant el dret d'explotació econòmica als seus amos a canvi de donar a conéixer a la societat els detalls de funcionament d'aquesta invenció. Perquè una patent puga ser concedida ha de complir tres requisits: ser novetat (no haver sigut exposat o publicat amb anterioritat), complir l'activitat inventiva i tindre aplicació industrial. És per això que les patents són documents valuosos, ja que contenen una gran quantitat d'informació tècnica no inclosa abans en un altre tipus de document (publicat o disponible).

A causa de les característiques particulars de les patents, els recursos que aquestes esmenten, així com els recursos que esmenten les patents, contenen enllaços que poden ser útils i donar suport a diverses aplicacions (vigilància tecnològica, desenvolupament i innovació, Triple-Helix, etc.) en disposar d'informació complementària, així com de la creació d'eines i tècniques que permeten extraure'ls i analitzar-los.

Per aquest motiu, la present tesi doctoral planteja la possibilitat de descobrir recursos web de qualitat a través dels enllaços mitjançant el disseny d'un mètode adequat per a la seua extracció i anàlisi eficient, aplicant l'ús de la tècnica cibermétrica d'anàlisi d'enllaços.

Per a això, s'ha descrit en l'estat de la qüestió dos grans blocs: les patents i la cibermetria, amb l'objectiu d'oferir un marc de treball que permeta entendre correctament l'abast de l'anàlisi.

Després de l'exposició teòrica del camp d'estudi, es realitza una explicació del mètode proposat per a aconseguir els objectius que defineixen la tesi. Aquest mètode es troba dividisc en dos blocs complementaris: Patent Outlink i Patent Inlink, que junts conformen la tècnica de Patent Link Analysis. Per a tots dos blocs es segueix un procés de treball equivalent. En primer lloc, es selecciona la font de la qual extreure les dades, després d'això es recopilen aquells necessaris i s'emmagatzemen en brut amb doble còpia per evitar pèrdues d'informació, d'aquesta manera és possible recuperar en tot moment la informació inicial.

S'analitzen les dades brutes per entendre el tipus d'informació i estructura en la cual es troben les dades i que permeta preparar un sistema d'extracció específic per a cada bloc. En el cas de el Bloc Patent Outlink, es desenvolupa un programa ad hoc per a l'extracció d'enllaços continguts en els documents en patents, que permet extreure la major quantitat possible d'enllaços amb el mínim soroll, i recopilant al seu torn informació relacionada amb la secció d'aparició o la categoria de la patent, de manera que es puga realitzar una anàlisi en major profunditat.

Per a realitzar l'estudi es selecciona l'Oficina de Patents i Marques dels Estats Units (USPTO), recollint totes aquelles patents concedides entre els anys 2008 i 2018 (tots dos inclosos). Una vegada extreta la informació a analitzar en cada bloc es compta amb: 3.133.247 de patents, 2.745.973 milions d'enllaços continguts en patents, 2.297.366 milions de pàgines web de patents enllaçades, 17.001 pàgines úniques web enllaçant a patents i 990.663 patents úniques enllaçades des de documents web.

Els resultats de l'anàlisi de Patent Outlink mostra com tant la quantitat de patents que contenen enllaços (20%), com el nombre d'enllaços contingut en patents (mitjana 4-5) és encara baix, però ha crescut significativament durant els últims anys i es pot esperar un major ús en el futur. Hi ha una diferència clara en l'ús d'enllaços entre àrees de coneixement (42% pertanyen a Física, especialment Computació i Càlculs), així com per seccions dins dels documents, explicant els resultats obtinguts i la projecció d'anàlisi futurs.

Els resultats de l'anàlisi de Patent Inlinks identifica una quantitat considerable menor de dominis webs que enllacen a patents (17.001 enfront de 256.724), però hi ha més enllaços per document enllaçant (el nombre d'enllaços total és similar per a tots dos blocs d'anàlisi). Així mateix, les dades mostren una elevada dispersió, amb uns pocs dominis generant una gran quantitat d'enllaços. Tots dos blocs mostren l'existència

d'una alta relació amb empreses i serveis tecnològics, existint diferències relatives als enllaços a Universitats i Governs (més enllaços en Outlink).

Els resultats mostren que el model d'anàlisi proposat permet i facilita el descobriment i avaluació de recursos web de qualitat. Així mateix, es conclou que la cibermetria, mitjançant l'anàlisi d'enllaços, aporta informació d'interès per a l'anàlisi dels recursos web de qualitat a través dels enllaços continguts i dirigits a documents de patents.

El mètode proposat i validat en la present tesi permet d'una manera eficient, eficaç i replicable l'extracció i anàlisi dels enllaços continguts i dirigits a documents de patents. Permetent, al seu torn, definir, modelar i caracteritzar el Patent Link Analysis com un subgènere del Link Analysis que pot ser utilitzat per a la construcció de sistemes de monitoratge de link intelligence, d'avaluació i/o qualitat, entre altres, mitjançant l'ús dels enllaços entrants i sortints de documents de patents aplicable a universitats, centres d'investigació així com empreses públiques i privades.

Capítulo 1
Introducción

En este capítulo de introducción, se realiza un recorrido por los motivos que explican la necesidad del desarrollo del estudio planteado, buscando dar un primer paso en el universo de las patentes y la importancia de su estudio. De este modo, es posible situar la necesidad y beneficios de la tesis.

Entre los apartados que lo conforman, se realiza la exposición de los objetivos a contestar con la presente tesis, así como una breve introducción de la metodología propuesta junto con los retos, a todos los niveles, a los que se enfrenta la consecución del análisis.

1.1. La importancia del estudio de las patentes

Según la definición de la Organización Mundial de Propiedad Intelectual (OMPI/WIPO) una patente es un derecho exclusivo que se concede sobre una invención que permite a su titular decidir sobre el uso de ésta por terceros. Estos documentos generan un contrato bidireccional entre el titular y la entidad que los concede, un contrato de tipo *quid pro quo* (o *do ut des*) mediante el que se protegen los intereses del inventor o inventora a cambio de conocer el funcionamiento de la invención.

De esta forma, el dueño de la patente mantiene la propiedad de aquello recogido y descrito en el documento, pudiendo decidir las formas de explotación que más le interesen, incluso en lo relativo a la creación, uso, venta e importación por otros agentes durante un periodo máximo de 20 años. Mientras que la entidad protectora se beneficia gracias a la exposición pública obligatoria del contenido de la patente. Esta exposición del conocimiento recogido en las patentes permite impulsar el desarrollo y la investigación, agilizando y fomentando la innovación de terceros.

La importancia de las patentes queda reflejada en la longevidad de los sistemas de protección, en funcionamiento desde la antigua Grecia hasta nuestros días, que han sabido adaptarse a través de la historia para dejar de ser privilegios otorgados por repúblicas y monarcas hasta sentar las bases fundamentales tras la implantación del primer sistema moderno en Venecia (1474) y llegar ser la salvaguarda concedida por los gobiernos actuales.

Las invenciones más famosas y significativas de la historia moderna forman parte de los diferentes sistemas de protección en todo el mundo, desde la mejora de la maquina de vapor de Watt (1769 – Patente inglesa nº 913), pasando por la dinamita de Nobel (1867 – Patente inglesa nº 1.345, Patente sueca nº 102, Patente francesa nº 72.007 y Patente americana nº 78.317 [año 1868]), el teléfono de Graham Bell[1] (1876 – Patente americana nº 174.465), la bombilla de filamento incandescente de Thomas Edison (1880 – Patente americana nº 223.898), el telégrafo de Tesla (1900 – Patente americana nº649.621), el autogiro de Juan de la Cierva (1920 – Patente española nº 74.322), hasta la más reciente patente para la edición del genoma mediante la tecnología CRISP (2014 – Patente americana nº 8.697.359).

Estas patentes son sólo una pequeña muestra del significado que tienen como puntos históricos en la evolución de la ciencia y la tecnología. Pone de manifiesto, además, la importancia económica de éstas, ya que pueden llegar a suponer para sus inventores la creación de imperios económicos gracias a las ganancias generadas por sus patentes (Dropbox, Facebook, GoPro, Windows o Apple son algunos ejemplos de gigantes tecnológicos junto con las patentes mencionadas anteriormente).

Las cifras relacionadas con las patentes aumentan año tras año. Desde 1990 a 2019 el incremento de solicitudes y concesiones supera el 200%. O, como demuestra un estudio

[1] En 2002 la Cámara de Representantes de Estados Unidos reconoce a Antonio Meucci como inventor original del teléfono (los registros de las patentes habían desaparecido) [107th Congress, H Res 269]

realizado en 2019 por la American Intellectual Property Lawyer's Association (AIPLA), sólo en Estados Unidos los costes derivados de litigaciones para la defensa de una patente pueden ser de media entre 2,3 y 4 millones de dólares.

Las patentes llegan a ser tan valiosas que, en ocasiones, una empresa compra la totalidad de otra únicamente para tener la propiedad de sus patentes, aunque después venda a un coste más económico la empresa comprada (sin incluir las patentes en el contrato de compraventa). Un ejemplo de ello es el caso Google–Motorola[2]. Con este tipo de acciones las empresas pueden no sólo adquirir el conocimiento de la patente, sino, además, adquirir los derechos de explotación –con los beneficios económicos que esto supone tanto propias como en licencias–, obtener ventajas competitivas, evitar su uso por parte de la competencia, etc.

Junto con la importancia como protección, el impulso a la innovación y el desempeño económico tanto de empresas como, indirectamente, de los propios países que las conceden, las patentes tienen un gran valor documental.

Debido a la necesidad de cumplir tres requisitos fundamentales (ser novedad, cumplir la actividad inventiva y tener aplicación industrial) para que una patente sea concedida, los documentos presentados a la Oficina que concede la patente deben recoger detalladamente la descripción y funcionamiento del invento en las diferentes secciones que componen una solicitud de patente (descripción, reivindicaciones e informe de búsqueda, junto con esquemas y dibujos opcionales). Además, deben explicar el problema que resuelven y cómo lo logran, lo que implica que deben recoger todas las evidencias que permitan demostrar que indudablemente la patente cumple los requisitos necesarios para ser concedida.

Esto implica que el documento de una patente contiene una gran cantidad de información muy valiosa que recoge el pasado y presente del conocimiento sobre el que se construye (estado de la técnica), junto con información totalmente inédita. De hecho, según la OMPI, se estima que los documentos de patentes contienen un 70% de información que no se encuentra en ningún otro tipo de publicaciones[3].

El análisis y explotación de los datos contenidos en los documentos de patentes es aplicable a diversos campos, tales como la vigilancia tecnológica, el desarrollo tecnológico, la búsqueda de inversión, la evaluación de la ciencia, tecnología e innovación o las relaciones Industria–Universidad–Gobierno (Triple Hélice). Además, permite cuantificar, tanto a nivel local como global, la evolución y competitividad de los propios países en los que se registran las patentes. Y, en definitiva, es un indicador del desarrollo y evolución industrial de un país o área.

Por lo tanto, debido a su alto valor informacional y las posibles aplicaciones de este conocimiento, el estudio de la información contenida en los documentos de patentes

[2] https://www.forbes.com/sites/quentinhardy/2011/08/15/google-buys-motorola-for-patent-parts

[3] https://www.wipo.int/wipo_magazine/en/2005/01/article_0003.html

resulta de especial importancia. Pese a ello, la mayoría de las bases de datos accesibles online hasta la fecha se limitaban a mostrar los datos bibliográficos de los documentos (título, año, inventor, resumen, etc.) pero no se podía acceder al cuerpo completo del documento, lo que limitaba los estudios que se podían realizar. Con la aparición de bases de datos de patentes a texto completo en la web, como las ofrecidas por algunas de las Oficinas de patentes (i.e.: USPTO, EPO, WIPO, etc.) o bases de datos especializadas (Google Patents, Lens.org, etc.) aparece la posibilidad de realizar un análisis completo de los documentos, permitiendo una mayor explotación de los datos contenidos.

En la información incluida en el documento de una patente, se puede encontrar enlaces referenciando contenidos externos (tanto otras patentes como documentación no-patente), utilizados para dar contexto, corroborar información y/o reivindicar la legitimidad y novedad de la patente. Estas conexiones (patente a patente; patente a no-patente) pueden analizarse para conocer el impacto y uso de las patentes o la evolución del conocimiento científico, del mismo modo que se estudian los propios enlaces contenidos en la Web o las citas en publicaciones científicas.

Para poder realizar un análisis de estas características, el campo de la Cibermetría (Orduña-Malea, E.; Aguillo, 2014) aporta todas las herramientas y metodologías necesarias. Mediante la integración de técnicas informétricas, bibliométricas y cienciométricas, la Cibermetría es el campo encargado de medir, estudiar y analizar cuantitativamente la información contenida en el ciberespacio. Para lograrlo, cuenta con un ecosistema de indicadores y métricas que permite conocer el uso, impacto y visibilidad de los contenidos online disponibles en el espacio red.

Dentro de la Cibermetría, la técnica de análisis de enlaces es ampliamente utilizada para la identificación y evaluación de las relaciones generadas entre documentos mediante el estudio de los hiperenlaces que los vinculan. Estos vínculos permiten interconectar o apuntar al documento que aporta la información necesaria mediante la mención de la ubicación fichero en el que se encuentra dicha información (i.e.: una página web, un archivo de vídeo, imagen o documento PDF). Así mismo, existen diferentes tipos de enlaces, algunos ejemplos son (Orduña-Malea, E.; Aguillo, 2014):

- → Estructurales: permiten navegar un sitio web entre las páginas que lo conforman (i.e.: los enlaces que se encuentran en un menú de navegación),
- → Entrantes: midiendo desde el dominio a.com, aquellos recibidos desde otro dominio web, por ejemplo, b.com
- → Salientes: midiendo desde el dominio a.com, aquellos que van desde a.com a b.com
- → Internos: a nivel de dominio (enlaces desde a.com/1 a a.com/2) y a nivel de documento (de una sección a otra dentro de un mismo documento)

Debido a que se trata de elementos esenciales para la construcción de la Web, favoreciendo la visibilidad y reutilización de contenidos, especialmente debido a que enlazan a aquellos contenidos con los que potencialmente guardan una relación semántica, su estudio y medición permite que sean utilizados como indicadores de visibilidad e impacto: cuanto más enlazado se encuentra un recurso, más importante

(más gente lo ha enlazado) y visible es (aparece en las primeras posiciones de resultados en los buscadores, generando más tráfico web hacia ellas).

A la hora de realizar análisis de enlaces, existen diversas métricas que pueden ser aplicadas a la medición de enlaces (número total, cantidad entrante/saliente, tipo, categoría, etc.), además, la combinación con otras métricas (offline, redes sociales, otras áreas, etc.) permite generar indicadores o marcadores de calidad relativos al grado de relevancia otorgado al contenido al y desde el que se dirigen.

Aplicar el análisis de enlaces tanto a los enlaces entrantes como salientes de los documentos de patentes puede ayudar a entender mejor la forma en la que las patentes son consumidas en el entorno online, donde la forma de comunicar, transmitir y compartir la información se genera bajo actividades y acciones que pueden diferir del mundo offline. Por tanto, al estudiar las patentes en el mundo online, se amplían los horizontes de medición y conocimiento de éstas.

Como se puede observar en la Figura 1, el análisis de enlaces relacionados con patentes ofrece tres vías diferentes:

- → Enlaces entre patentes: conexiones que permiten localizar patentes de temática similar, aunque se encuentren en otra Oficina u área de aplicación
- → Enlaces desde patentes hacia recursos web: uso de recursos de información por parte de los inventores
- → Enlaces desde recursos web hacia patentes: impacto, uso o visibilidad de la patente por parte de la sociedad

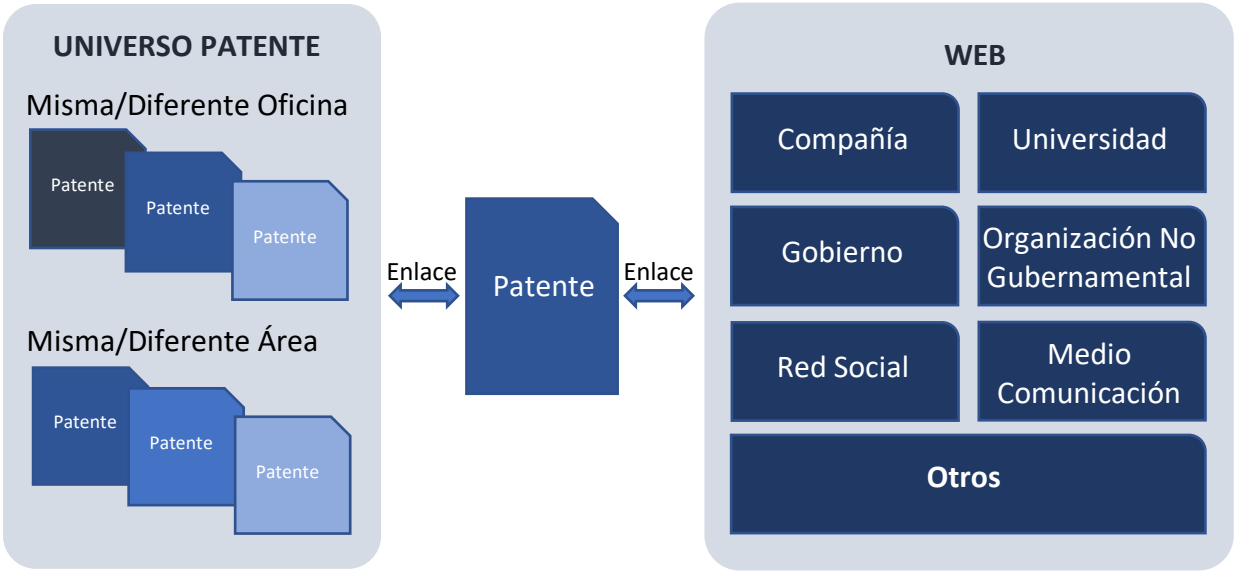


Figura 1: Sistema de relación entre documentos de patentes y los universos de estudio. Fuente: elaboración propia

Por lo tanto, realizando un análisis de enlaces entrantes y salientes a los documentos de patentes se puede lograr dos objetivos: primero, localizar las patentes más relevantes debido a la cantidad y calidad de enlaces que reciben. Segundo, localizar recursos de información de calidad alojados en el entorno Web gracias a su referencia desde documentos de patentes. Siendo ambos objetivos complementarios.

El primer objetivo permite conocer el impacto y visibilidad de las patentes. Al contabilizar los enlaces como menciones y analizar la calidad de esas menciones, es posible calcular cómo de importante es el documento mencionado. Esto puede ayudar a conocer el nivel de transferencia que existe mediante las interacciones y sinergias que se dan entre las patentes y el mundo que las consume.

El segundo objetivo permite localizar contenidos y documentos que contienen información tan relevante (y por ello son recursos de calidad) que los inventores han decidido incluir en la patente (un documento altamente técnico) para aumentar las posibilidades de conseguir la concesión de la invención.

Gracias a este análisis, se puede tener conocimiento de la relevancia de los documentos de patentes como objetos informacionales, se puede localizar recursos de información de calidad, conocer los métodos y la importancia del uso de enlaces en documentos de patentes, su evolución y caracterización. En definitiva, qué tipo de conexión de la información existe utilizando las patentes como nexo entre otros recursos informativos y los consumidores de información.

El análisis de enlaces en patentes permitirá aportar luz sobre el consumo de los documentos y la información que contienen, pudiendo ser útil para una gran cantidad de agentes, como pueden ser los propios inventores, agencias, gobiernos, instituciones científicas, universidades, empresas, abogados, investigadores, ciudadanos, industria, gestores, entre otros.

1.2. Retos en la medición

Actualmente existen nuevas posibilidades de explotación de los datos contenidos en documentos de patentes debido a la aparición de fuentes de información que permiten el análisis online del texto completo de la patente; anteriormente, los estudios realizados (Agrawal & Henderson, 2009; Aristodemou & Tietze, 2018; Y. S. Chen, Shih, & Chang, 2014; E. J. Han & Sohn, 2015; Hegde & Sampat, 2009; Kousha & Thelwall, 2017; Meyer, 2000a, 2000b; Sarin et al., 2020; Tseng, Lin, & Lin, 2007; Venugopalan & Rai, 2015; Wang, Lei, & Lee, 2014; Ye, Huang, & Chen, 2016) sobre documentos de patente se limitaban al análisis de los datos bibliográficos (inventores, años, empresas, palabras clave, …), por ello, hasta la presente tesis, el estudio de los enlaces contenidos en patentes es muy limitado (Orduña-Malea, Thelwall, & Kousha, 2016).

Aunque actualmente es posible acceder al texto completo de los documentos de patentes de forma gratuita y online, la presente tesis se enfrenta a tres problemas metodológicos principales:

Primero, el acceso a lo documentos de patentes. Pese a poder encontrar actualmente los documentos de patentes a texto completo en la Web, la cobertura de las diferentes fuentes de información o la forma de acceso a los documentos (pueden encontrarse en PDF, HTML o formatos descargables como XML) varía entre ellas. Además, para poder acceder a la información sin restricciones, debe buscarse un modo de acceso que evite

posibles pérdidas de datos, problemas de acceso y descarga, y que permita sistematizar el proceso de análisis, pudiendo replicar el método tantas veces como sea necesario.

Segundo, la lectura de los datos. Dependiendo de la fuente y el acceso a los documentos pueden existir errores de lectura y recopilación de información. Por ejemplo, al leer una página HTML y extraer los datos de ella, es necesario asegurarse que la página está totalmente cargada y no queda información oculta. En otros casos, el problema puede estar derivado por el tipo de fichero que se está utilizando; por ejemplo, un documento PDF permite mantener la información en la posición correcta pero la lectura del texto para su extracción puede dar errores (cambios de letras [i.e.: e por c, i por l, etc.]) o la lectura de zonas incorrectas (inclusión de pie de página, error en la lectura de imágenes, etc.).

Por último, el tercer problema se encuentra relacionado con el manejo de los datos. Para que un estudio pueda aportar suficiente información, especialmente en la evolución del uso de enlaces tanto entrantes como salientes a documentos de patentes, se considera necesario el análisis de una gran cantidad de datos, así como una cobertura temporal amplia. Así mismo, en caso de recoger una gran cantidad de enlaces extraídos desde las patentes o desde la Web a patentes, junto con la información contextual necesaria, el volumen de datos puede ser excesivamente grande (gigas de información) lo que puede suponer un problema tecnológico en el manejo, almacenamiento y análisis de los datos.

1.3. Objetivos

La presente tesis plantea las siguientes preguntas de investigación:

- ¿Es posible identificar recursos web de calidad mediante el uso de los enlaces web incluidos en las patentes?
- ¿Es posible identificar recursos web de calidad mediante documentos que enlazan a las patentes?

Con el fin de dar respuesta a las preguntas de investigación anteriormente indicadas, la presente tesis se plantea como objetivo principal **diseñar, aplicar y validar un método orientado a la identificación de recursos de información web de calidad a través de la técnica de análisis de enlaces aplicada a patentes**.

Para lograr este fin, los objetivos específicos planteados a continuación se encuentran divididos según el bloque de análisis mediante el que se trate de resolverlos:

- El Bloque Patent Outlink, busca determinar la viabilidad del uso de patentes para la identificación de recursos online de calidad. Para ello se plantea la aplicación de técnicas de outlinks con el fin de lograr los siguientes objetivos específicos:
  - Determinar el grado de utilización de enlaces web como recursos de información en las patentes.
  - Analizar la evolución de este uso de los enlaces en patentes en el tiempo
  - Comprobar la existencia de diferencias en el uso de enlaces en patentes por áreas de conocimiento

- El Bloque Patent Inlink, busca caracterizar el impacto de la patente, como objeto informacional en Internet (*patent as web genre*). Para ello se plantea la aplicación de técnicas de inlinks con el fin de lograr los siguientes objetivos específicos:
  - Estimar la visibilidad web de las patentes
  - Identificar los recursos web que enlazan patentes
  - Determinar la calidad e impacto web de los sitios web que enlazan a patentes

1.4. Avance metodológico

Para desarrollar los objetivos indicados anteriormente, el trabajo se divide en los siguientes dos bloques de análisis

- Patent Outlinks: Análisis de enlaces de patentes a recursos web
- Patent Inlinks: Análisis de enlaces de recursos web a patentes

De este modo se logra un sistema que permite realizar un análisis bidireccional que aplicará el análisis de enlaces como herramienta cibermétrica.

El primer bloque realiza la extracción de los enlaces contenidos los documentos de patentes concedidos por la Oficina de Patentes y Marcas de Estados Unidos. El segundo bloque, se recopilarán todos los enlaces a patentes posibles mediante Majestic, una herramienta especializada en *link intelligence*.

Para lograr extraer la información necesaria se han desarrollado herramientas ad hoc que permiten realizar la recolección y preparación de los datos recogidos. Además, se ha diseñado un sistema de almacenamiento y análisis de los datos debido a la gran cantidad de gigas de información recopilados.

Ambos bloques se encuentran estructurados siguiendo el mismo esquema y se incluyen tanto los experimentos fallidos como los válidos, tratando de dar una mayor información y aportar conocimiento sobre posibles variaciones del método que puedan existir.

1.5. Estructura del trabajo

La presente tesis doctoral se encuentra estructurada en siete capítulos. Siendo el primero el presente de Introducción en el que se justifica la necesidad de ejecución de la investigación y los objetivos a alcanzar.

Sigue el Capítulo 2 que recoge el Estado de la Cuestión, diferenciando dos grandes bloques de estudio. El primer bloque describe en detalle todo el universo relacionado con las patentes para lograr conocer el ecosistema del que son parte. El segundo bloque describe el área de estudio de la Cibermetría, sus técnicas y aplicaciones, que será la herramienta utilizada para la consecución de la tesis.

El Capítulo 3 describe la Metodología utilizada para recoger, preparar y almacenar todos los datos necesarios para su estudio. Para lograr una descripción detalla, y dado que la tesis en sí misma puede ser divida en dos estudios complementarios, se divide la explicación en dos bloques:

- → Análisis de enlaces desde patentes a recursos web
- → Análisis desde contenidos web a patentes

Cada bloque contiene los pasos necesarios para la selección de fuentes de información, recopilación, extracción y almacenamiento requeridos para el análisis.

El Capítulo 4 recoge los resultados del análisis de los datos recopilados, divido en los mismos bloques descritos en el Capítulo 3.

El Capítulo 5 contiene la discusión de los resultados obtenidos mientras que el Capítulo 6 recoge las conclusiones extraídas tras el desarrollo y análisis de la tesis, junto con las líneas futuras de investigación.

Por último, el Capítulo 7 recopila toda la bibliografía utilizada a lo largo de la tesis.

Al finalizar se encuentran los diversos Anexos que recopilan información útil, así como la extensión de resultados para la facilitar la comprensión de la tesis.

Finalmente, se ponen a disposición distintos Anexos que recopilan tanto los datos brutos de este trabajo como información complementaria.

Capítulo 2
Estado de la cuestión

En este capítulo se profundiza en las áreas principales en las que se cimienta el proyecto, las patentes y la cibermetría, desarrolladas en dos grandes partes.

La primera parte recoge en detalle las patentes, su historia, uso e importancia, de forma que se pueda conocer mejor tanto los propios documentos como el entorno que rodea al mundo de las patentes. De esta forma se obtiene el contexto y perspectiva que permiten entender el alcance del proyecto.

La segunda parte describe el campo de la Cibermetría en general y la técnica de Análisis de Enlaces en particular, para poder conocer todos los métodos y herramientas disponibles para el análisis de la información.

2.1. Objeto de análisis: las patentes

De acuerdo con el diccionario de la Lengua de la Real Academia Española, el término **patente** (*1. Adj. Manifiesto, visible*) deriva del latín *patents (-entis)* y significa 'estar expuesto' o 'ser evidente'. De modo que una patente es un documento que expone o evidencia [presenta o demuestra] una invención.

Las patentes son descendientes naturales de las *Litterae Patentes* (Cartas Abiertas), documentos legales otorgados por un monarca o presidente, que concedían el derecho exclusivo sobre un cargo, derecho, monopolio, título o estatus a su beneficiario. Estos documentos declaran públicamente *"*a quienes la presente vieren y entendieren*"* la voluntad del emisor, generalmente un monarca, de conceder al titular de la concesión unos derechos o beneficios para lo recogido en el documento. Estos derechos podían encontrarse entre una gran variedad de motivos[4], siendo uno de ellos el Real Privilegio de Invención de un inventor a explotar su invento (Casado-Serviño & Sanz-Martínez, 2013).

De acuerdo, tanto con la definición etimológica como con los documentos predecesores, el papel fundamental de una patente es el de mostrar a todo aquel que acceda a la *carta abierta* la información necesaria para entender el funcionamiento de la invención recogida en el documento. A cambio de desvelar los secretos del invento, el titular de la patente obtiene, según los sistemas actuales de protección intelectual, el derecho de *exclusión.* Este derecho permite evitar que terceros agentes exploten comercialmente la invención patentada, siendo su inventor quien debe dar permiso para la reproducción, uso, distribución y/o venta.

La protección que otorga una patente tiene una aplicabilidad temporal y geográfica definida y limitada, que viene dada por el momento y país en el que se realiza la solicitud. Aunque, como se verá en el §2.1.3 se puede registrar una patente de modo internacional. La protección temporal, una vez concedida la patente, comienza en el momento en el que se presenta y concede la solicitud y tiene una duración de 20 años[5] tras los que ésta pasará a ser de dominio público.

Para que se pueda ofrecer esta protección la invención objeto de patente debe cumplir tres requisitos indispensables:

a. Ser una novedad
b. Cumplir con la actividad inventiva
c. Tener aplicación industrial

[4] Hasta 92 documentados en el caso de Gran Bretaña
https://www.whatdotheyknow.com/request/108336/response/284995/attach/html/3/FOI%2076096%20Elibank.doc.html

[5] En el caso de las patentes farmacéuticas o fitosanitarias el periodo de exclusividad es diferente, la protección no comienza en el momento en el que se solicita la patente, primero se deben cumplir una serie de requisitos regulatorios, una vez obtenida la autorización de comercialización aplica el derecho de exclusividad. Esto implica que de los 20 años que dura la protección de una patente finalmente pueden ser menos. Para evitar las perdidas en el retorno de la inversión, los sistemas de patentes ofrecen certificados complementarios de protección que extienden hasta 5 años la duración de la patente, de modo que se garantice un mínimo de 15 años de exclusividad.

Para que un invento se considere una novedad, según la Ley 24/2015, (a) éste no debe encontrarse en el *estado de la técnica*, siendo ésta todo aquello que se ha puesto previamente a disposición del público en todo el mundo, de forma escrita, oral o mediante su uso. Es decir, para que un invento se considere novedoso, no debe encontrarse nada igual en el mercado, ni se debe haber expuesto o divulgado de algún modo su funcionamiento, incluyendo la presentación o comentario en seminarios, cursos, exposiciones, artículos, comunicados de prensa, debates, posters, blogs o la comercialización parcial del mismo.

La actividad inventiva (b) hace referencia a la capacidad del invento de no ser inferido de una manera evidente por un experto en la materia desde el estado de la técnica, dicho de otro modo, ha sido necesaria la aplicación de una gran cantidad materia gris en el desarrollo del invento.

Por último, el invento a patentar debe resolver un problema técnico, teniendo una aplicabilidad o utilidad industrial (c), independientemente del tipo de industria que sea, permitiendo su explotación real.

Aunque un invento cumpla los tres requisitos de patentabilidad mencionados anteriormente éste debe, además, ser considerado una invención patentable. De acuerdo con la Ley 24/2015 en España no se consideran patentables los siguientes supuestos:

- Descubrimientos o teorías científicas, así como los métodos matemáticos
- Obras literarias, artísticas o creaciones estéticas
- Los planes, reglas y métodos para el ejercicio de actividades intelectuales, juegos o actividades económico-comerciales
- Programas de ordenador
- Las formas de presentar información
- Métodos de diagnóstico y tratamientos de tipo terapéuticos o quirúrgicos
- Las invenciones cuya explotación comercial sea contraria al orden público o a las buenas costumbres. En particular:
    - Los procedimientos de clonación de seres humanos
    - Los procedimientos de modificación de la identidad genética germinal del ser humano
    - Las utilizaciones de embriones humanos con fines industriales o comerciales
    - Los procedimientos de modificación de la identidad genética de los animales que supongan para estos sufrimientos sin utilidad médica o veterinaria
- Las variedades vegetales o las razas animales
- Los procedimientos esencialmente biológicos de plantas o animales
- El cuerpo humano en todas sus etapas o una parte de éste
- Secuencias de ácido desoxirribonucleico (ADN) sin indicación de función biológica alguna

2.1.1. Propiedad intelectual

El arte, los inventos y las marcas se encuentran protegidos bajo el paraguas de la Propiedad Intelectual. Según el Real Decreto 1/1996, de 12 de abril, "*la Propiedad Intelectual está integrada por derechos de carácter personal y patrimonial, que atribuyen al autor la plena disposición y el derecho exclusivo a la explotación de la obra, sin más limitaciones que las establecidas en la ley*". Es, por tanto, la protección que se otorga a bienes o activos intangibles, de modo que las creaciones de la mente cuenten con la misma protección que aquellas con propiedad física.

Como se puede ver en la Figura 2, bajo el paraguas indicado anteriormente como Propiedad Intelectual, existen diferentes categorías en función del tipo de derecho que protege a sus creadores: la Propiedad Industrial y los Derechos de Autor.

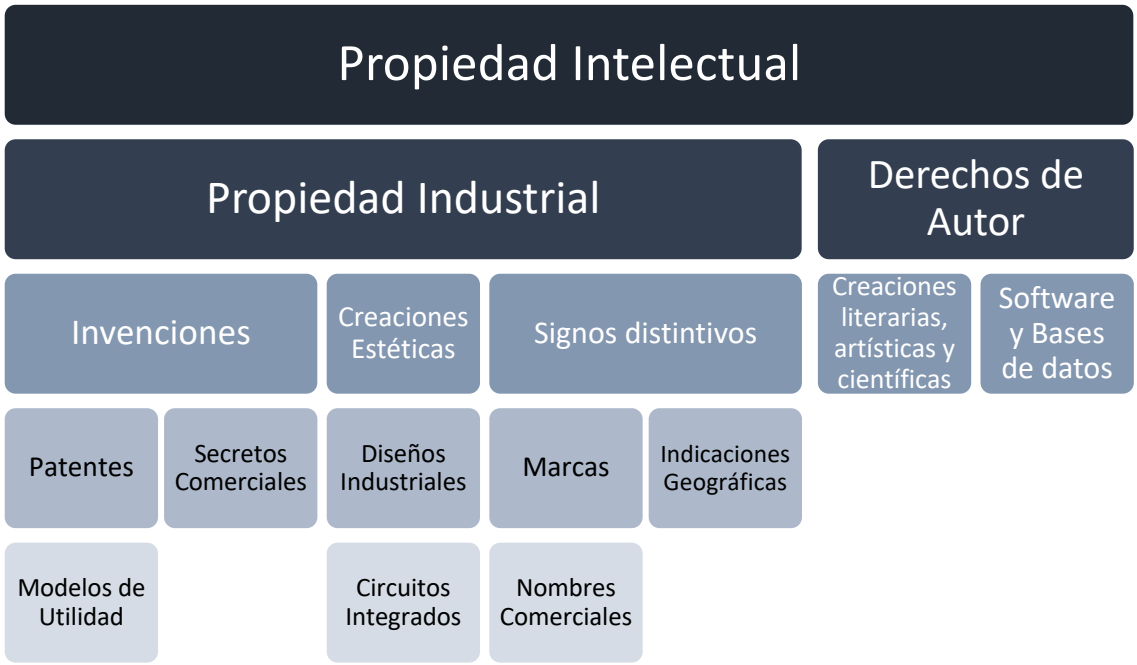


Figura 2: Cuadro ilustrativo sobre las áreas de la Propiedad Intelectual.
Fuente: elaboración propia

Los Derechos de Autor son aquellos que obtienen los creadores sobre sus obras, abarcando la palabra *obras* cualquier tipo de expresión creativa, original, intelectual o artística como son: las obras literarias, la música, la pintura, las esculturas, obras de teatro, películas, fotografías, retransmisiones televisivas, los programas informáticos, las bases de datos, los anuncios publicitarios, los perfumes, los mapas y los dibujos técnicos o arquitectónicos, siendo este listado únicamente una breve recopilación de ejemplos (OMPI, 2016b). Los derechos de autor nacen el momento en que se crea la obra y, generalmente, no requieren de registro, así como tampoco precisan de una marca indicando que se encuentran los derechos reservados o que la obra se encuentra

sujeta a *copyright*. Estos derechos, además, duran la totalidad de la vida del autor y los 70 años posteriores a su fallecimiento[6].

La Propiedad Industrial abarca las invenciones y los diseños industriales, así como las marcas registradas, nombres comerciales, denominaciones de origen y secretos comerciales. Según se explica en el Convenio de París, la propiedad industrial se basa en la protección de las creaciones intelectuales de carácter técnico que transmiten información, principalmente a los consumidores, sobre los productos y servicios disponibles en el mercado. En este caso sí es necesario el registro en una Oficina de modo oficial para lograr proteger la idea y la temporalidad variará en función del tipo registro y país en el que se realice.

Los derechos que conlleva la autoría de una *obra* amparada bajo el derecho de autor son dos: los derechos morales y los derechos patrimoniales. Los derechos de explotación económica, también llamados patrimoniales, permiten al autor percibir una retribución económica por el uso o adquisición por terceros de la obra, es decir, por el derecho a copia. Los derechos morales, habitualmente no cedibles, preservan y protegen la vinculación autor–obra.

En cambio, la propiedad industrial se asienta en un derecho fundamental: el derecho a exclusión (OMPI, 2016a). Su protección tiene como fin último el impedir toda utilización no autorizada de los productos o servicios protegidos[7]. La diferencia fundamental entre los dos sistemas es la monopolización del mercado que se genera. Los derechos de autor permiten el uso de la obra siempre y cuando se reconozca (y se compense al autor si procede), la protección industrial únicamente permite que se use el conocimiento cuando el inventor así lo desee[8].

Para poder entender la importancia de la propiedad industrial y el motivo por el que se escogen las patentes y no otro tipo de salvaguarda para las invenciones, se considera necesario comprender la diferencia que existe entre los diferentes métodos de protección industrial disponibles.

En la Tabla 1 se puede encontrar un resumen esquemático de las diferencias más importantes entre las diferentes posibilidades de protección que existen explicadas a continuación. Debido a la naturaleza y tipo de protección que ofrecen los *Modelos de Utilidad*, *Diseños No Registrados*, *Topografía de Productos Semiconductores* y los *Nombres Comerciales* estos han sido descritos, pero no incluidos en la tabla.

Patentes de invención

Las patentes de invención, objeto principal de la presente investigación, se conceden a las invenciones técnicas. Como se explica con mayor detalle en el §2.1, para que se

[6] Generalmente, ya que esto varía en función del país.

[7] Una obra amparada por los Derechos de Autor si puede ser utilizada por otros siempre y cuando no sea una copia exacta.

[8] Es por este motivo que la protección industrial únicamente se otorga para 20 años y el derecho de autor puede durar más de un siglo.

pueda lograr una patente, el invento debe cumplir 3 requisitos básicos (ser novedad, tener actividad inventiva y aplicación industrial). Se debe solicitar en una oficina de protección intelectual y una vez garantizada la patente, su duración es de 20 años, encontrándose limitada al país o área en el que se concede.

Modelos de utilidad

Los modelos de utilidad son utilizados para proteger invenciones de menor rango inventivo que las patentes, con menores requisitos –se debe cumplir obligatoriamente con el requisito de novedad– y menor tiempo de protección (10 años), reduciendo a su vez los costes. Asimismo, la solicitud es más sencilla, el proceso de concesión más corto (suele durar 4 meses) y el alcance de la protección es similar. Mediante este sistema se pueden proteger utensilios, instrumentos, herramientas, aparatos, dispositivos o partes de estos. [Ley 24/2015, de 24 de julio]

Diseños industriales

Los diseños industriales comprenden tanto dibujos como modelos de aplicación industrial y se utilizan para proteger los elementos creativos, ornamentales y estéticos que determinan la apariencia física de un producto. Para que se conceda la protección el diseño debe ser novedad y tener un carácter singular. La duración de protección es de 5 años renovables cada 5 años hasta un máximo de 25 y funciona del mismo modo que la protección de una patente: concede el derecho exclusivo de explotación al diseñador. [*Ley 20/2003, de 7 de julio*]

Diseños No Registrados

Del mismo modo que existen los derechos de autor para las obras, existe la posibilidad de proteger un diseño sin necesidad de realizar un trámite legal de registro para ello. Al presentar al público un diseño novedoso y original se adquiere el derecho a impedir que otros copien éste. La protección legal es mucho menor y únicamente se puede explotar durante los tres años siguientes a la presentación del producto. En España no tiene aplicabilidad, pero sí en la Unión Europea.

Topografía de Productos Semiconductores (Circuitos integrados)

Dado que integrar una gran cantidad de funcionalidades en un esquema de trazado es complejo y costoso, pero realizar una copia no, los circuitos se encuentran protegidos de modo que no se pierda la inversión realizada en investigación. Dado que no atañen a la estética, ni son novedades, no tienen cabida en el resto de los sistemas de protección. La protección que se obtiene una vez registrado es de 10 años. [*Ley 11/1988, de 3 de mayo*]

Marcas

Como Marca se entiende los símbolos, signos o indicadores que diferencien o distingan a un producto o servicio. Dentro de esta protección se encuentran los nombres, palabras, números, logotipos, colores, olores, formas o sonidos que identifiquen el producto y lo distinga claramente de su competencia. La distinción inequívoca es el único requisito que se debe cumplir en el momento de realizar el registro. Pese a que se puede generar una marca sin necesidad de registrar oficialmente, se considera conveniente ya que facilita posibles litigios. Un registro de marca es indefinido siempre que se renueve cada 10 años. [Ley 17/2001]

Nombres comerciales

Los nombres comerciales permiten identificar a las empresas en el tráfico mercantil, de modo que se distinga del resto de empresas de la competencia. No es necesario que coincidan con la marca (que identifica a los productos o servicios que vende) o la denominación social (que identifica a una persona jurídica). Su temporalidad es la misma que para las marcas (renovable cada 10 años ilimitadamente).

Indicaciones geográficas

Una Indicación geográfica permite señalar las cualidades, calidad y reputación de un producto ligado a su lugar de origen, de modo que los factores locales de producción permiten diferenciar el producto de su competencia. Las Indicaciones geográficas recogen Denominaciones de Origen Protegidas (centradas en las características del producto) e Indicaciones Geográficas Protegidas (centradas en la zona geográfica de producción). Las DOP son más estrictas que las IGP ya que es obligatorio que todo el proceso de producción sea realizado en la zona a la que pertenece. [Reglamento CE 1151/2012]

Secretos comerciales

Son una alternativa a las patentes. Mediante contrato de confidencialidad, se ofrece información no conocida por el público, de modo que se pueda explotar un invento o proceso sin necesidad de realizar una patente (que descubriría el funcionamiento completo). Los secretos comerciales no permiten perseguir si la competencia realiza ingeniería inversa, pero son muy útiles para proteger ventajas competitivas. Para que un secreto comercial sea considerado como tal la información protegida debe ser:

a. valiosa
b. difícil de descubrir
c. encontrarse debidamente protegida.

Los secretos comerciales pueden durar indefinidamente, la única condición es que la información que protegen continúe siendo secreto.

La protección que otorgan todos los derechos de propiedad intelectual descritos anteriormente se refiere a los diferentes componentes intangibles, tanto internos como externos, de las creaciones, por lo tanto, pueden utilizarse conjunta y simultáneamente para proteger los distintos aspectos de una misma invención.

Asimismo, cabe señalar que en el mismo Convenio de Paris en el que describen todos los tipos de protección indicados hasta ahora existe un apartado concreto para la *Competencia Desleal*. Ésta viene a completar la protección otorgada a la propiedad intelectual, de modo que se proteja los conocimientos, tecnología e información que no pueden ser objeto de las protecciones existentes. De este modo se intenta evitar generar confusión o aseveraciones falsas que resulten contrarias a las exigencias de buena fe [Ley 3/1991, de 10 de enero].

A continuación, la Tabla 1 muestra un esquema de las principales diferencias para cada unas de las posibilidades de protección descritas con anterioridad.

Tabla 1: Esquema resumen de la Propiedad Intelectual en España.
Fuente: elaboración propia

	Patentes	Secretos Comerciales	Diseños Industriales	Marcas	Indicaciones Geográficas	Derechos de Autor
¿Qué se protege?	Inventos	Información de negocio	Apariencia exterior	Identificación de personas, productos o servicios	Signo de cualidades o reputación geolocalizadas	Obras creativas, literarias o científicas
¿Cómo se protege?	Solicitud	Contrato	Solicitud	Uso y Solicitud	Solicitud	Inherente
Requisitos	Nuevo, útil, no obvio	Valioso, debe mantenerse como secreto	Nuevo y original	Uso comercial, diferencia inequívoca y característica	Cumplir el pliego de condiciones de la Indicación	Original
Temporalidad	20 años	Indefinido	5 años (renovables cada 5 años hasta 25)	Indefinido (revisión cada 10 años)	Indefinido (revisión cada 10 años)	Vida del autor + 70 años
Acceso al conocimiento	Si	No	Si	No completo	No completo	No completo
Conducta prohibida	Manufacturar, usar o vender	Defraudar; publicar; exponer	Copiar	Generar confusión	Copiar o engañar sobre el origen	Copiar o crear una obra similar
Desarrollo independiente	Prohibido	Permitido	Prohibido	Prohibido	Prohibido	Puede estar prohibido
Coste	Alto	Depende	Moderado	Moderado	Moderado	Nulo

2.1.2. Sistemas de patentes. Historia.

En este apartado se realiza un breve repaso de la evolución histórica de los sistemas de patentes. Este recorrido se inicia en el sur de la península itálica, en el 720 a. C., se asentó en el golfo de Tarento una colonia griega con el nombre de Síbaris. Sus habitantes, los sibaritas, eran conocidos por su interés en la abundancia, el lujo y el confort. En su ciudad no estaba permitido que trabajasen herreros ni carpinteros para evitar el ruido, por este motivo no había gallos ni niños jugando en la calle. Les gustaba vestir bien, sus telas eran tan demandadas que alcanzaban precios astronómicos para la época. Pero lo más destacable era su interés por la comida. Tanto era así, que cuando un cocinero preparaba un plato de especial sabor, se le concedía el derecho de ser el único que lo elaborase durante un año (Frumkin, 1945).

Este derecho exclusivo contiene las bases de las patentes modernas vistas en el §2.1: el plato debía ser una novedad, el autor tenía los derechos sobre su plato y se le ofrecía el derecho de explotación y exclusión durante un plazo determinado otorgando así un

monopolio temporal (Witty, 2017). Aunque se desconoce la protección real ofrecida por el estado a los inventores-chefs, se puede observar una forma primitiva de lo que serían las actuales patentes.

En el año 510 a. C, según Heródoto, Telis convence a la población de Síbaris para derrocar el gobierno oligarca desterrando a 500 de las personas más ricas de la capital y confiscar sus bienes. Los desterrados sibaritas, buscan refugio en Crotona, otra ciudad de la región de Calabria, pero son perseguidos y se solicita a las autoridades de Crotona que sean entregados. Los crotoniatas se niegan y así empieza una guerra que terminará con la destrucción y eliminación, literal[9], de la ciudad de Síbaris (Rutter, 1970).

Del mismo modo que Síbaris fue eliminada, el concepto de patente usado por los sibaritas desapareció. Ni en la antigua Grecia, dónde Platón o Aristóteles estaban en contra de recompensar los avances tecnológicos por su "*degradación de la mente"* y ser perjudiciales para el avance político y social. Y, pese a que Hipódamo de Mileto propuso una ley para que los ciudadanos que ofreciesen descubrimientos ventajosos para su país recibiesen honores, aunque Aristóteles se opuso a la idea (Paden, 2001). Ni durante el Imperio Romano se tiene constancia de que existiera una idea similar. De hecho, en el Imperio Romano los monopolios eran ilegales. Si bien es cierto que los gobernantes podían ofrecer monopolios como gratificaciones, pero el sistema se corrompió y el Emperador Zenón (aproximadamente en el 480 d. C) llegó a proclamar:

> *"Nadie debe ejercer el monopolio sobre ningún […] material, ya sea por su propia autoridad o bajo la autoridad de una norma imperial, pasadas o futuras…" Emperador Zenón*

Eliminando, además, cualquier tipo de privilegio previo que pudiera existir y permitiendo únicamente en casos muy justificados formas de explotación única (Prager, 1950). Por lo tanto, ni griegos ni romanos, pese a sus inmensas contribuciones y legado, favorecieron a los autores intelectuales otorgándoles algún tipo de retribución o reconocimiento por su aportación.

Tras la caída del Imperio Romano y durante toda la Edad Media la religión y el barbarismo impidieron el correcto avance tecnológico. A lo largo de la Alta Edad Media (s. V – X) seguía vigente el pensamiento de Platón y Aristóteles y no se tiene constancia de que existiera algún tipo de salvaguarda para las nuevas invenciones (Frumkin, 1945).

En la Baja Edad Media, con la consolidación de los monarcas, se extiende el uso de las *litterae patents* comentadas en el §2.1, derechos otorgados por los monarcas que se concedían por méritos logrados o servicios ofrecidos. Estos privilegios serán los que más adelante se conozcan como Patentes y se otorguen a los inventores. Con estos privilegios, se busca atraer el talento para su explotación local de forma que permitiese el avance de los gremios, ciudades o reinos con respecto a la competencia.

[9] Según indican los escritos de Heródoto, los crotoniatas desviaron el río Cratis para inundar y eliminar completamente la ciudad de Síbaris.

Y es en este periodo, durante los primeros años de la Baja Edad Media, cuando comienzan a aparecer privilegios significativos para el desarrollo de las actuales patentes (Frumkin, 1945; Prager, 1950):

- → En 1105 se le concede un diploma a Norman Abbot en la baja Normandía para la instalación de un molino
- → En 1234, en Burdeos, a Bonafusus de Santa Columbia se le concede el permiso para la fabricación de telas
- → En 1315 se concede en Bohemia un privilegio para el desalojo de agua en minas
- → En 1330 Felipe IV de Francia le concede a Philippe de Caqueray el privilegio exclusivo para la fabricación de vidrio
- → En 1332 a Bartolomeus Verde se le concede, y se le paga por ello, un privilegio de innovación tecnológica por la construcción de un molinillo de viento en Venecia
- → En 1378 se concede a Mauritius la exclusividad de un sistema de desagüe
- → En 1390 Ulman Stromer, en Nuremberg, logra un privilegio por un molinillo para fabricar papel
- → En 1404, en Praga, se concede a Michael von Deutsch-Brod un privilegio para la fabricación de unas fuentes de agua
- → En 1421 Filippo Brunelleschi obtiene la exclusividad de uso durante tres años de un sistema de su invención para el transporte de cargas pesadas para la construcción de la cúpula de la catedral de Florencia
- → En 1440 a John Shidame se le concede un privilegio para un nuevo sistema de producción de sal
- → En 1444, Antonio Marini logra un privilegio, en Venecia, para la explotación durante seis años de molinos de viento
- → En 1449 se concedió al vidriero flamenco John de Utyman derecho de explotación por su proceso de tintado de vidrio
- → En 1460, en Venecia, se le concede un privilegio a Gugglielmo Lombardus por un tipo de horno que teñía con un menor consumo
- → En 1469 Joannis da Spira logra un privilegio en Venecia para una nueva técnica de impresión de libros
- → En 1472 logra en Venecia Mathio Brancho un privilegio por un molinillo para macerar

Y, finalmente, en 1474, en Italia, se formaliza mediante el Estatuto de Venecia lo que hoy en día se conoce como un sistema de patentes. Los privilegios indicados anteriormente se parecían en mayor o menor medida a las patentes actuales, la mayoría cumplía alguno de los tres puntos imprescindibles de una patente actual, así como ofrecer beneficio y reconocimiento al autor por una cantidad variable de años. Pero la *Parte Veneziana* publicada por el Senado de la Republica de Venecia sienta las bases completas de qué es una patente, la utilidad que debe tener para la sociedad, el estímulo inventivo que supone, y el reconocimiento y compensación económica para el inventor.

Pese a que la ley no fue rigurosamente aplicada inicialmente, ya que se concedieron diferentes periodos de explotación durante los primeros siglos, se considera la base

fundamental de todos los sistemas de patentes con una vida de más de 500 años sin apenas modificaciones. Su éxito fue notable, desde 1474 hasta 1550 se concedieron más de 100 patentes, con un incremento progresivo de patentes otorgadas al año (Nard & Morriss, 2006). Desde Italia el sistema de patentes se importó al resto de Europa: Alemania, Bélgica, Francia, Holanda e Inglaterra crearon los suyos propios.

Entre estos, en la Francia de Enrique II, se introdujo un nuevo punto para la concesión de patentes, el inventor debía indicar una descripción detallada, por lo que en 1555 se publicó la primera especificación en una patente. Siendo, además, los franceses los primeros en 1791 en indicar que cualquier tipo de descubrimiento o invención pertenece a su autor, siendo esta la primera Ley de Patentes moderna del mundo (Adams, 2019).

En Inglaterra, en 1623, durante el reinado de Jacobo I[10], sucedió un hecho significativo que heredarían los sistemas de patentes: la publicación del *Estatuto de Monopolios*. En éste, se eliminaban todos los monopolios otorgados hasta la fecha, se limitaba la temporalidad de los nuevos y, lo más importante, se limitaban los inventos a novedades, que debían encontrarse descritas en detalle y obligando al inventor a hacer realidad el invento patentado. Se buscaba promover la economía e impulsar la industria, se iniciaba la Revolución Industrial. El Estatuto tuvo una duración de 200 años, hasta que en 1852 se instauró la enmienda de la Ley de Patentes, generando el sistema de patentes inglés actual (Nard & Morriss, 2006).

La legislación inglesa influyó notablemente a las colonias americanas, donde las primeras patentes aparecieron en Massachussetts (1641) o Connecticut (1672) amparadas por estatutos similares al inglés. Tras la Guerra de Independencia y necesitando promover la industrialización del país, para evitar los problemas crecientes de discrepancias entre estados, se promueve incluir en la Constitución en 1787 la creación de un sistema de patentes, quedando redactado en 1790 el *Patent Act* y siendo firmado por George Washington el que será el primer estatuto sobre patentes de los Estados Unidos en el que se incluyen los fundamentos del actual sistema de patentes americano y que apareció en 1836 (Nard & Morriss, 2006).

En España, se tiene constancia de privilegios reales especiales otorgados por Fernando III y Alfonso X. Pero el primer privilegio destinado a la protección del inventor se da bajo el reinado de Isabel la Católica, en el año 1478 (Casado-Serviño & Sanz-Martínez, 2013). Aunque no es hasta 1679 cuando se crea la Real Junta General de comercio, predecesora de la actual Oficina Española de Patentes, y desde la que se promovía la actividad artesanal e industrial. Es en 1826 cuando se aprueba el Real Decreto de Privilegios Exclusivos de Invención e Introducción, reconocido como la primera ley española sobre patentes.

[10] Pese a que se considera que durante el reinado de Isabel I (1559 - 1603) se concedieron patentes similares a las actuales, éstas se parecían más a los privilegios de monopolio anteriores. Estos privilegios en un principio pensados para atraer a expertos extranjeros e impulsar la economía mediante inventos, acabaron convertidos en monopolios de comercio ofrecidos a las altas clases inglesas, lo que terminó enfadando a la población.

Finalmente, dado que la protección que ofrecen estos sistemas de patentes se limita geográficamente por el país en el que ha sido otorgada, se busca la internacionalización de los diferentes sistemas, intentando que los inventores se encuentren protegidos en otros países, por ello en 1883 se firma el Convenio de París (Adams, 2019) para la protección de la Propiedad Industrial en general. Además, en 1954 se creó el Convenio sobre la Clasificación Internacional, mediante el cual se genera un sistema alfanumérico para codificar los temas sujetos de patentes. Y, por último[11], en 1970 se establece el Tratado Internacional de Cooperación en materia de Patentes (PCT) que resuelve el problema del patentado internacional que no resolvía el Convenio de Paris, de modo que mediante una única solicitud se reciba protección en varios países.

2.1.3. Proceso de patentado, difusión y publicación

Históricamente, tal y como se explica en el apartado anterior, las patentes eran otorgadas por monarcas o gobiernos al mando del país en el que se realizaba la solicitud de protección. Actualmente, aunque la protección se continúa logrando en el país en el que tiene concedida la patente, la solicitud para comenzar el proceso de patentado de una invención se debe realizar en una oficina de registro de patentes y puede solicitarlo cualquier persona física o jurídica, directamente o mediante agente o representante en la oficina del país en el que desee protección para la invención.

Las oficinas de registro pueden ser de tipo nacional o regional, la diferencia entre una y otra varía en la cantidad de lugares en los que tendría validez la protección una vez otorgada. Mediante una oficina nacional, la patente se encuentra registrada y salvaguardada en el país de solicitud. En una oficina regional, la protección se otorga en todos aquellos países en los que el solicitante indique está interesado en recibir la protección y formen parte del acuerdo de la región.

Actualmente, existen 194 oficinas nacionales[12] y 5 oficinas de patentes regionales:

- → **Organización Africana de la Propiedad Intelectual (OAPI)[13]:** formada por 17 Estados (Benin, Burkina Faso, Camerún, República Centroafricana, Chad, Comoras, Congo, Côte d'Ivoire, Gabón, Guinea, Guinea Ecuatorial, Malí, Mauritania, Níger, Guinea Bissau, Senegal y Togo)
- → **Organización Regional Africana de la Propiedad Intelectual (ARIPO)[14]:** formada por 19 Estados (Botsuana, Ghana, Gambia, Kenia, Lesoto, Liberia, Malawi, Mozambique, Namibia, Ruanda, Somalia, Sudán, Santo Tomé y Príncipe, Sierra Leona, Suazilandia, Tanzania, Uganda, Zambia, Zimbabue)

[11] Con respecto a los convenios internacionales a nivel global, también existen convenios regionales como el que se firma en 1973 sobre la Patente Europea tras la creación de la Comunidad Económica Europea y que unifica la solicitud de patente del mismo modo que el PCT pero a nivel Europeo.

[12] https://www.wipo.int/directory/en/urls.jsp

[13] http://www.oapi.int/

[14] http://www.aripo.org

- → **Organización Euroasiática de Patentes (EAPO)**[15]**:** formada por 9 Estados (Turkmenistán, República de Bielorrusia, República de Tayikistán, Federación Rusa, República de Kazajistán, República de Azerbaiyán, Kirguistán, República de Moldavia, República de Armenia)
- → **Organización Europea de Patentes (OEP)**[16]**:** formada por 38 Estados miembros, 2 de extensión y 4 de validación (Albania, Alemania, Austria, Bulgaria, Bélgica, Chipre, Croacia, República Checa, Dinamarca, España, Estonia, Eslovenia, Eslovaquia, Finlandia, Francia, Grecia, Hungría, Irlanda, Islandia, Italia, Letonia, Liechtenstein, Lituania, Luxemburgo, Macedonia del Norte, Malta, Mónaco, Noruega, Países Bajos, Polonia, Portugal, Reino Unido, Rumanía, Serbia, Suecia, San Marino, Suiza y Turquía. Estados de extensión: Bosnia y Herzegovina, Montenegro. Estados de validación: Camboya, Republica de Moldavia, Marruecos y Túnez)
  - ⇢ **OEP – UE:** en la Unión Europea se propone, además, la posibilidad de solicitar una patente unitaria que otorgue protección en 24 Estados (actuales miembros de la UE, a excepción de España y Croacia, junto con Reino Unido que se ha desvinculado del proyecto) sin requerir la validación individual de estos, simplificando el proceso y los costes asociados.[17]
- → **Oficina de Patentes del Consejo de Cooperación de los Estados Árabes del Golfo (Oficina de patentes CCG)**[18]: formado por 6 Estados (Bahréin, Kuwait, Omán, Qatar, Arabia Saudita y Emiratos Árabes Unidos)

Además, existen varios tratados internacionales en materia de registro y obtención protección de patentes:

- → **Convenio de Paris**[19]**:** se adopta en 1883 y actualmente lo firman 177 Estados, abarca la protección de toda la propiedad intelectual y establece las normas comunes que todos los Estados firmantes deben aplicar. Se centra en tres categorías principales:
  - ⇢ **Trato nacional**: los Estados participantes deben reconocer a ciudadanos de otros estados participantes la misma protección que a los nacionales.
  - ⇢ **Derecho de prioridad**: tras la presentación de una solicitud en uno de los Estados firmantes, el solicitante tiene un plazo de 12 meses para solicitar el registro de patente en el resto de los Estados firmantes manteniendo la fecha de entrega de la primera solicitud y logrando así mantener el titulo de novedad y evitar posibles usurpaciones por presentaciones posteriores.

[15] https://www.eapo.org/

[16] https://www.epo.org/

[17] Se espera que el sistema entre en funcionamiento en 2022. Tanto España como Croacia podrán adherirse a él cuando deseen

[18] www.gcc-sg.org/eng/

[19] https://wipolex.wipo.int/es/text/288515

- ⇢ **Reglas comunes**: en relación con las patentes, el Convenio indica como normas el derecho del inventor a ser mencionado, las patentes concedidas en los diferentes Estados son independientes entre sí (una patente no puede ser denegada, anulada o caducar porque sea así en otro Estado), no es posible denegar una patente porque el producto o su venta se encuentre sujeto a restricciones legislativas y regulariza la concesión de licencias obligatorias para evitar abusos en determinadas condiciones.

→ **Tratado sobre el Derecho de Patentes**[20]**:** recoge las normas mínimas y comunes para la armonización de los procedimientos existentes en el proceso de elaboración de solicitudes de patentes ante las oficinas de registro. Se encuentra firmado por 42 Estados.

→ **Tratado de Cooperación en materia de Patentes (PCT)**[21]**:** actualmente firmado por 153 Estados, es un complemento al Convenio de Paris y únicamente pueden suscribirlo aquellos Estados que previamente han firmado éste. El PCT es lo más parecido a una patente internacional que existe actualmente, ya que permite designar entre todos los Estados contratantes aquellos en los que se debe tramitar la patente para solicitar protección iniciando el proceso con una única solicitud de patente (aunque después se debe lograr la patente por vía nacional). El procedimiento PCT consta de dos fases: internacional y nacional. Dado que se trata de una de las formas de lograr la protección de una patente, el proceso exacto se detallará en este mismo capitulo.

Por último, existen otros tratados internacionales que permiten unificar o normalizar diferentes aspectos de la protección que debe otorgar una patente, siendo los más relevantes:

→ **Tratado de Budapest**[22]**:** estipula el reconocimiento de las autoridades internacionales de depósito de microorganismos, con fines en el procedimiento en materia de patentado, suprimiendo el requisito de depositar muestras ante cada una de las autoridades en las que se busca protección. Se encuentra suscrito por 83 Estados.

→ **Arreglo de Estrasburgo relativo a la Clasificación Internacional de Patentes**[23]**:** establece la Clasificación Internacional de Patentes (CIP) dividiendo la tecnología en ocho secciones divididas en más de 70.000 subsecciones siendo representada mediante un código formado por números y letras que debe figurar en todos los documentos de patente. La clasificación de una patente la indica la oficina de registro y es un elemento indispensable para la recuperación de documentos. Actualmente la firman 62 Estados y se realiza una revisión anual para su actualización.

[20] https://wipolex.wipo.int/es/text/288997

[21] https://wipolex.wipo.int/es/text/488124

[22] https://wipolex.wipo.int/es/text/283785

[23] https://wipolex.wipo.int/es/text/291859

Gracias a la normalización y estandarización que ofrecen los convenios y tratados existentes, pese a que no es posible lograr una patente internacional per se, el proceso de solicitud es menos costoso, tanto económicamente como en plazos, y permitiendo salvaguardar las invenciones de un modo mucho más fiable para los inventores. Esta estandarización conlleva que los pasos a realizar para lograr la protección son muy parecidos en prácticamente la totalidad de sistemas de protección (aunque todos ellos tienen sus particularidades). En la Figura 3 se puede observar un esquema generalizado de las diferentes etapas que conforman el procedimiento de solicitud de una patente, aunque resumido, es el proceso por etapas utilizado por la gran mayoría de sistemas de protección, tanto nacionales como regionales e, incluso, el internacional (PCT).

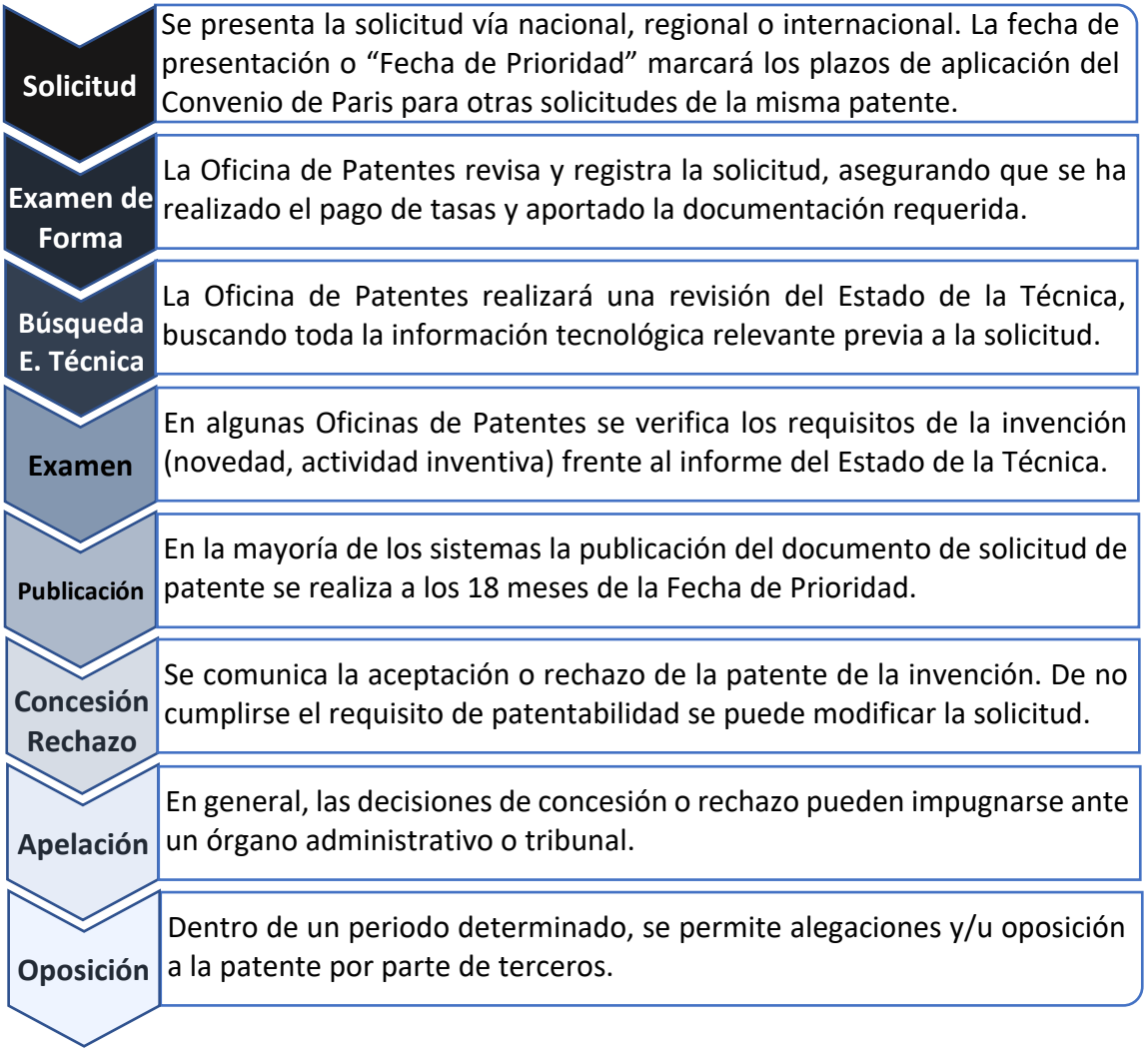


Figura 3: Procedimiento generalizado de una solicitud de patente
Fuente: elaboración propia

Además, se debe tener en cuenta que, una vez concedida una patente, independientemente del modo mediante el que se haya logrado, la protección se encuentra supeditada al cumplimiento de una serie de requisitos:

- → **Caducidad**: existen tres casos en los que una patente puede caducar si: las tasas anuales no son pagadas dentro de las fechas limites indicadas por las oficinas en las que se encuentren registradas, el tiempo de duración de la protección ha finalizado o el titular de la misma renuncia a ella.
- → **Derecho de prioridad**: si la solicitud se ha registrado por primera vez en un Estado adscrito al Convenio de Paris, como se menciona anteriormente, puede solicitar durante los 12 meses posteriores a la Fecha de Prioridad protección en otros Estados adscritos al Convenio, ya que de este modo se conserva el criterio de novedad.
- → **Explotación**: es obligatorio explotar comercialmente la invención patentada. Para ello puede realizarlo el propio titular de la invención o trasladar la explotación a un tercero, para esto existen dos métodos:
  - ⇢ **Cesión:** mediante la venta y transferencia de la propiedad de la invención.
  - ⇢ **Licencia:** concediendo los derechos de explotación temporalmente a cambio de una contraprestación.

Ligado a todo el procedimiento se encuentran los costes asociados a la solicitud de una patente. Existen cuatro tipos diferentes de costes asociados, tanto al proceso de mantenimiento como al de su concesión, ya que, una vez lograda para mantener la patente se debe realizar pagos anules.

- → **Tasas administrativas**: incluyen tanto la presentación de la solicitud en la(s) oficina(s), examen inicial para comprobar que el trámite se puede iniciar, designación de los Estados en los que se desea patentar, la concesión y publicación, y la validación (tras la concesión vía europea donde los Estados designados deben validar la patente).
- → **Costes del procedimiento**: mediante comunicaciones con las Oficinas y examinadores durante el proceso. En caso de que se utilice un experto o gabinete especializado, o abogados para la redacción o trámite legal de la patente los gastos se encontrarían en esta categoría.
- → **Costes de traducción**: en caso de solicitar una patente en un país con un idioma diferente, tanto si es por vía nacional (la traducción se necesita en el momento de la solicitud) como si es por vía regional/internacional (la traducción es necesario al llegar a la fase nacional).
- → **Costes de mantenimiento**: mantener activa una patente tras su concesión requiere de un pago de tasas anuales durante los 20 años de protección. En caso de que se utilicen agentes para la tramitación o gestión de la patente entrarían en esta categoría.
- → **Costes de derecho**: en caso de que se deba defender la patente de posibles infracciones o vulneraciones, invalidar posibles patentes que reivindiquen puntos similares, etc.

Calcular el coste exacto de patentar una invención es complejo ya que esto depende de muchos factores (la vía seleccionada, la cantidad de países en los que se realice la solicitud, el número de reivindicaciones o paginas, la velocidad seleccionada, posible cantidad de comunicaciones con la Oficina, el coste de agentes externos) (Sánchez,

Hortal, & Cuesta, 2015). Según el estudio realizado por Everis en 2015, el coste medio de lograr y mantener una patente durante 20 años en las principales oficinas nacionales del mundo (Europa, Brasil, Canadá, China, India, Israel, Japón, Rusia, Corea del Sur y Estados Unidos) en 2011 era de 10.412 euros. En caso de solicitar una patente por vía europea y lograr cobertura en 13 Estados el coste medio podía ascender a 102.926 euros, el aumento se debe sobretodo a los costes de traducción y validación.

Asimismo, las duraciones en los procesos también varían dado que cada Estado y región tienen sus particularidades debido a normas y legislación propias, puesto que no se encuentra en el alcance de esta tesis realizar un análisis de bajo nivel de los métodos de solicitud, a continuación se detalla brevemente el funcionamiento de los tipos de solicitudes existentes para que se tenga un conocimiento general de los diferentes sistemas de protección.

Solicitud Nacional – Caso de estudio: España

Como se ha indicado anteriormente, los registros nacionales permiten recibir protección en el país en el que se solicitan. En el caso de España, la oficina encargada de realizar todos los trámites relativos a la recepción, estudio y concesión de patentes (y el resto de modalidades de propiedad intelectual) es la Oficina Española de Patentes y Marcas (OEPM). Fundada como tal en 1992, recoge el testigo del Real Conservatorio de Artes y Oficios creado en los Reales decretos de 1810, 1820 y 1824 (Casado-Serviño & Sanz-Martínez, 2013).

En este caso, debido al compromiso español con la normalización internacional, las fases de tramitación son las mismas que las indicadas en la Figura 3, dado que los tiempos específicos de trámite, aunque acotados en rangos, sí suelen variar entre países, la Figura 4 muestra una línea temporal por meses para la concesión de una patente en España. La duración del proceso depende de las correcciones y contestaciones y los tiempos que se empleen en ellas. En España, el plazo de publicación medio es de 18 meses [Orden ETU/296/2017].

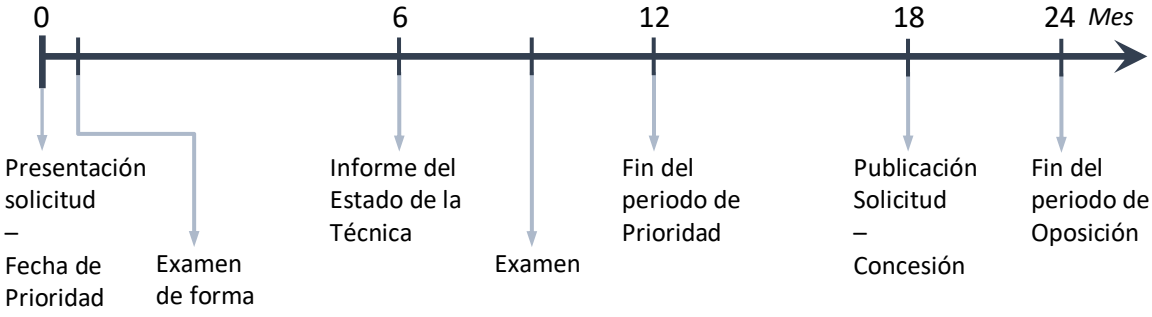


Figura 4: Representación temporal del proceso de solicitud de una patente en España
Fuente: elaboración propia

Con relación a las tasas e importes que se derivan del proceso para lograr una patente, el coste total varía (en el momento de escribir la presente tesis, diciembre 2020) entre los 3.000 y 5.000 euros. La cuantía final depende de diversos factores: si es presentada vía online (15% de descuento), si se pueden aplicar otros descuentos (PYMEs, personas físicas emprendedoras o universidades públicas) y de los costes de preparar la documentación mediante la contratación de un experto o gabinete especializado en PI

(rango 1.000 – 3.000 euros). Respecto a los costes fijos de la vía no electrónica[24], los más importantes se muestran en la Tabla 2, a los que hay que añadir las anualidades en materia de tasas para el mantenimiento de la protección. Estas anualidades se pagan a partir del tercer año y el importe aumenta anualmente comprendido en el rango 18,48€ – 490€ (el importe se encuentra sometido a la revisión de los Presupuestos Generales del Estado).

Tabla 2: Importe de los trámites no electrónicos más importantes en el proceso de solicitud de una patente en España
Fuente: elaboración propia

Trámite	Clave	Importe
Solicitud de registro de patente	IT01	100,38€
Solicitud de informe sobre el Estado de la Técnica	IT04	684,65€
Solicitud de examen sustantivo	IT22	389,77€
Solicitud de resolución urgente de expediente	IT03	47,39€
Por cada prioridad reivindicada en materia de patentes	IT06	19,65€

Solicitud Regional – Caso de estudio: Europa

La patente europea (que no la patente única europea) permite mediante una solicitud designar los Estados europeos a los que aplicar para lograr una protección de patente.

En este caso, la oficina administradora es la Oficina Europea de Patentes (OEP – EPO en inglés), fundada en 1973[25], con sede en Múnich, encargada de gestionar el sistema de patentes europeo que se rige bajo el Convenio de Múnich. Las solicitudes deben presentarse en inglés, francés o alemán, aunque si el Estado en el que se presenta tiene un idioma oficial diferente, se puede realizar la solicitud en el idioma del país y después presentar una traducción.

Las etapas del proceso[26] (Figura 5) son las mismas que las indicadas en la Figura 3, con la particularidad de que la línea temporal es ligeramente diferente a la fase nacional indicada en la Figura 4 y que al finalizar con la concesión de la patente, los Estados en los que se solicita protección deben validarla, tras el pago de las tasas nacionales y la presentación de la traducción de la patente en aquellos Estados que lo requieran.

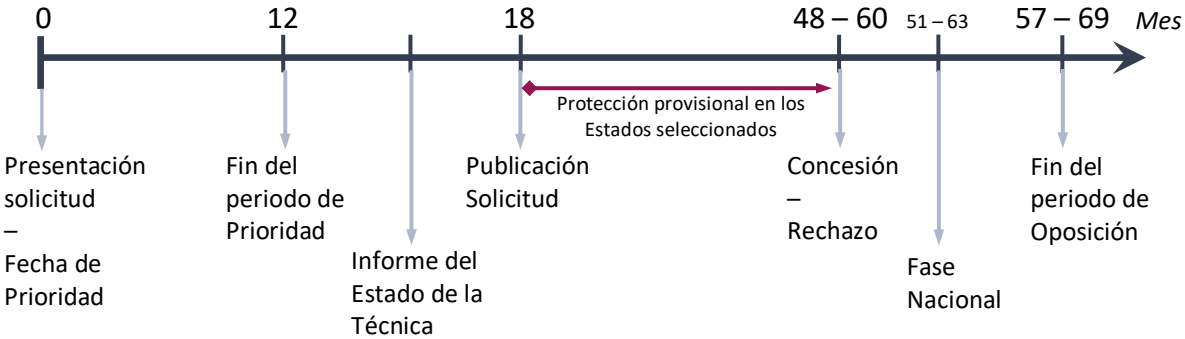


Figura 5: Representación temporal del proceso de solicitud de una patente vía Europea
Fuente: elaboración propia

[24] https://www.oepm.es/export/sites/oepm/comun/documentos_relacionados/Tasas/2020_PATENTES.pdf

[25] https://www.epo.org/about-us/timeline.html

[26] https://www.epo.org/service-support/faq/own-file.html

Durante el proceso de solicitud de patente europea, también deben pagarse distintas tasas hasta el momento de concesión (depósito, búsqueda, Estados seleccionados, examen, concesión, reivindicaciones e impresión), así como las tasas anuales de renovación en cada uno de los Estados. El coste aproximado del proceso[27] hasta alcanzar la etapa de Concesión es de 6.000 euros, el coste total –dependiendo de una gran cantidad de factores (nº de Estados designados, etapas necesarias, traducciones, etc.)– puede partir como valor inicial desde los 15.000 euros. A lo que habría que añadir las costas de un gabinete o experto en caso de requerir sus servicios.

Del mismo modo que sucede en la vía nacional, existen diferentes tipos de descuentos que pueden ayudar a disminuir el precio final del proceso. La Tabla 3 muestra los importes más importantes de cada trámite. Las anualidades en esta vía comienzan en el tercer año siendo 490 euros, hasta los 1.640 euros del último año

Tabla 3: Importe de los trámites no electrónicos más importantes en el proceso de solicitud de una patente en Europa Fuente: elaboración propia

Trámite	Clave	Importe
Solicitud de registro de patente	001	260,00€
Solicitud de informe sobre el Estado de la Técnica	003	1.775,00€
Solicitud de examen sustantivo	006	1.700,00€
Por cada Estado seleccionado	005	610,00€
Por cada prioridad reivindicada en materia de patentes	016	245,00 – 610,00€

Solicitud Internacional: PCT

Este sistema es la forma más parecida a lograr una patente internacional, aunque en sí mismo no es un procedimiento de concesión de patentes, ni sustituye a las concesiones nacionales, pero si permite facilitar el sistema de solicitud en varios países simultáneamente unificando la tramitación de la protección en los Estados seleccionados.

La oficina encargada de realizar el proceso de gestión es la Organización Mundial de la Propiedad Intelectual (OMPI – WIPO)[28], fundada en 1967 y administradora de 26 tratados, entre ellos el PCT. La solicitud se puede presentar frente a la OMPI o la oficina nacional correspondiente.

El procedimiento se divide en dos fases[29] (similares a las de la vía europea) descritas a continuación, y cuyos tiempos se pueden observar en la Figura 6.

[27] https://www.epo.org/applying/fees/fees.html

[28] https://www.wipo.int/about-wipo/

[29] http://www.oepm.es/export/sites/oepm/comun/documentos_relacionados/Publicaciones/Folletos/Proteccion_Internacional_Invenciones.pdf

→ **Internacional**: en la que además del IET se puede solicitar un Informe de Búsqueda Complementaria para realizar una búsqueda de documentos de patente más amplia y un Informe de Examen Preliminar Internacional, para conocer la opinión de la oficina sobre la patentabilidad. Siendo interesante designar para estos informes a la oficina que estratégicamente sea más relevante para la invención ya que ofrecen fundamentos sólidos para poder evaluar correctamente la patentabilidad de la invención.

→ **Nacional**: fase en la que se formaliza la solicitud en los Estados designados y se busca su concesión.

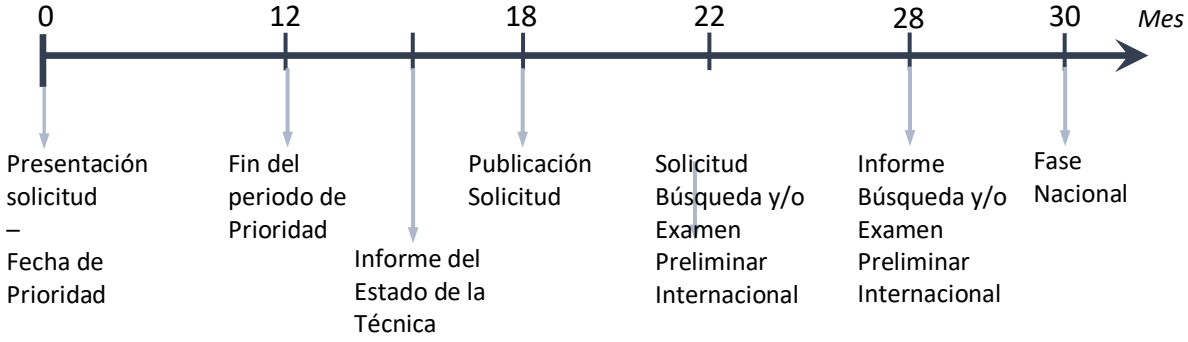


Figura 6: Representación temporal del proceso de solicitud de una patente vía PCT
Fuente: elaboración propia

Los costes de seguir esta vía son muy similares a los de la vía europea (valor inicial aproximado de 15.000 euros), se debe tener en cuenta que mediante este sistema no se realiza pagos de solicitud de patente en cada uno de los Estados designados, únicamente de las traducciones que fueran necesarias, tasas y mandatarios locales. Las tasas principales relativas al PCT son las mostradas en la Tabla 4.

Tabla 4: Importe de los trámites no electrónicos más importantes en el proceso de solicitud PCT
Fuente: elaboración propia

Trámite	Importe
Solicitud de presentación internacional	1.217,00€
Solicitud de informe sobre el Estado de la Técnica	1.775,00€
Tasa de admisión en la Oficina receptora	1.700,00€
Solicitud de examen preliminar	1.830,00€

Otro tipo de solicitudes: Solicitud Acelerada Internacional (PPH)

Además, existen otras vías, como por ejemplo la Solicitud Acelerada Internacional (Patent Prosecution Highway – PPH[30]), que permite acelerar los procedimientos de concesión de patentes, gracias al intercambio de información entre oficinas mediante acuerdo bilaterales y multilaterales, y reducir los costes de gestión evitando posibles suspensos. De este modo, se evita la duplicidad de esfuerzos en materia de presentación de solicitudes y documentos, mediante la reutilización del trabajo realizado en otras Oficinas. Por ejemplo, España tiene acuerdos PPH con China, Colombia, Finlandia, Japón, Marruecos, México, Perú, Rusia, Taiwán y Turquía.

[30] https://www.jpo.go.jp/e/toppage/pph-portal/globalpph.html

2.1.4. El documento de la patente

Más allá de lo que implica una patente en tanto en cuanto a sistema de protección para nuevas invenciones, el propio documento en sí mismo es una fuente relevante de información. Como se ha visto en el §2.1.3 lograr una patente no es sencillo, se debe realizar una gran inversión económica, humana y temporal, que permita reunir toda la información necesaria para *convencer* al examinador que el invento *merece* la protección. Y es todo ese conjunto de información recopilada en el informe la que, a distintos niveles, concede la importancia al propio documento de la patente.

Durante el proceso de patentado existen dos momentos en los que se generan un documento de patente como tal: el momento de solicitud y el momento de concesión. La diferencia fundamental entre el primer y segundo documento es la documentación extra que acompaña al segundo (informes, observaciones, comentarios, etc.) y que permiten entender el alcance total de la protección. Para simplificar el análisis, se va a proceder al estudio de un documento de patente concedido.

Dentro de un documento de patente, en las diferente partes y secciones que lo componen, se puede encontrar información relativa a:

- → Información técnica (dibujos y descripción)
- → Información jurídica (reivindicaciones de la patente)
- → Información comercial (datos del solicitante, fechas, país, etc.)

Para que esta información sea de fácil acceso, facilitando el intercambio de información, su difusión y uso comercial, la OMPI tiene publicadas unas normas (WIPO Standards) que permiten armonizar la estructura y contenido de los documentos de patentes.

Aunque cada país publica los documentos siguiendo sus propias normas legislativas, la mayoría tratan de seguir las recomendaciones, por lo que con ligeras modificaciones en la disposición de la información en los apartados, en todos los países el documento de la patente está compuesto por el mismo tipo de información, siguiendo la estructura de informe estándar conformada por tres secciones indispensables (portada, descripción y reivindicaciones) y dos opcionales e independientes (sección de dibujos, figuras, esquemas, etc. y/o el informe sobre el estado de la técnica).

Estas secciones de forma obligatoria deben incluir, al menos en una ocasión, los siguientes datos estandarizados [31]:

- → El código de la oficina de propiedad industrial u organización que publica (ST.3)
- → El número de publicación (ST.6)
- → El código de tipo de documento (ST.16)
- → La fecha de publicación (ST.2)

[31] Se incluyen entre paréntesis las normas indicadas por la OMPI para la estandarización de la nomenclatura a las que pertenecen los datos listados

Siendo, entre éstos, el número de publicación el dato más relevante ya que permite la identificación unívoca de la patente en todo el mundo. Para lograr esto se encuentra formado, según las recomendaciones de la OMPI, por la concatenación alfanumérica de:

- **Código de País** (Country Code – CC) (ST.3): mediante el uso de dos caracteres se identifican tanto los países como las organizaciones mundiales que tramitan solicitudes de patentes. Algunos ejemplos:
  - España: ES
  - Estados Unidos: US
  - OMPI (WIPO en inglés): WO
  - OEP (EPO en inglés): EP
- **Número de serie**: serie normalmente numérica que puede contener símbolos de separación como contra barras o puntos para la separación de los dígitos.
- **Código del tipo de publicación** (ST.16): uno o dos caracteres que indican el tipo de documento publicado. Algunos ejemplos:
  - Patente: A (solicitud), B (concesión)
  - Modelo de utilidad: U (solicitud), Y (concesión)
  - Traducción: T
  - Medicamento: M
- **Código numérico** (opcional):
  - 1 – 7: disponibles para la codificación de cada oficina de patentes
  - 8: reservado internacionalmente para indicar una corrección en la primera página
  - 9: reservado internacionalmente para indicar una corrección en el documento completo.

Este sistema de identificación mediante códigos numéricos tiene dos problemas importantes:

- La disparidad de significados de códigos dependiendo de la oficina
- La disparidad de significados de códigos en una misma oficina debido a la evolución de ésta a lo largo del tiempo.

Debido a esto y, dado que la norma ST.16 no deja de ser una recomendación de la OMPI, es importante tener acceso al documento *Inventario de los tipos de documentos de patente por orden de oficinas de propiedad industrial emisoras* que recoge las prácticas y usos de cada país[32].

La primera página de un documento de patente, y que hace las veces de portada, contiene los datos bibliográficos (asociados en gran medida a la información comercial) y que permiten identificar inequívocamente a la propia patente. Para facilitar la

[32] https://www.wipo.int/export/sites/www/standards/es/pdf/07-03-02.pdf

comprensión de los datos, su interoperabilidad y difusión, éstos se encuentran identificados mediante códigos INID[33] recogidos en la norma ST.9 de la OMPI[34].

Los datos bibliográficos se agrupan en conjuntos identificativos englobados en los siguientes apartados:

- → (10) Identificación de la patente
- → (20) Solicitud de la patente
- → (30) Datos relativos a la prioridad en virtud del Convenio de Paris o del Acuerdo sobre los Aspectos de los Derechos de Propiedad Intelectual del Comercio
- → (40) Fechas
- → (50) Información técnica
- → (60) Referencias a patentes previas
- → (70) Identificación de personas relacionadas con el documento
- → (80)(90) Datos relativos a convenios internacionales (excepto lo recogido en (30))

En la Figura 7 se puede ver la primera página de una patente española a modo de muestra. Se ha dividido mediante globos numerados las diferentes secciones para ayudar a su explicación. Cada sección contiene la siguiente información junto con sus códigos INID[35]:

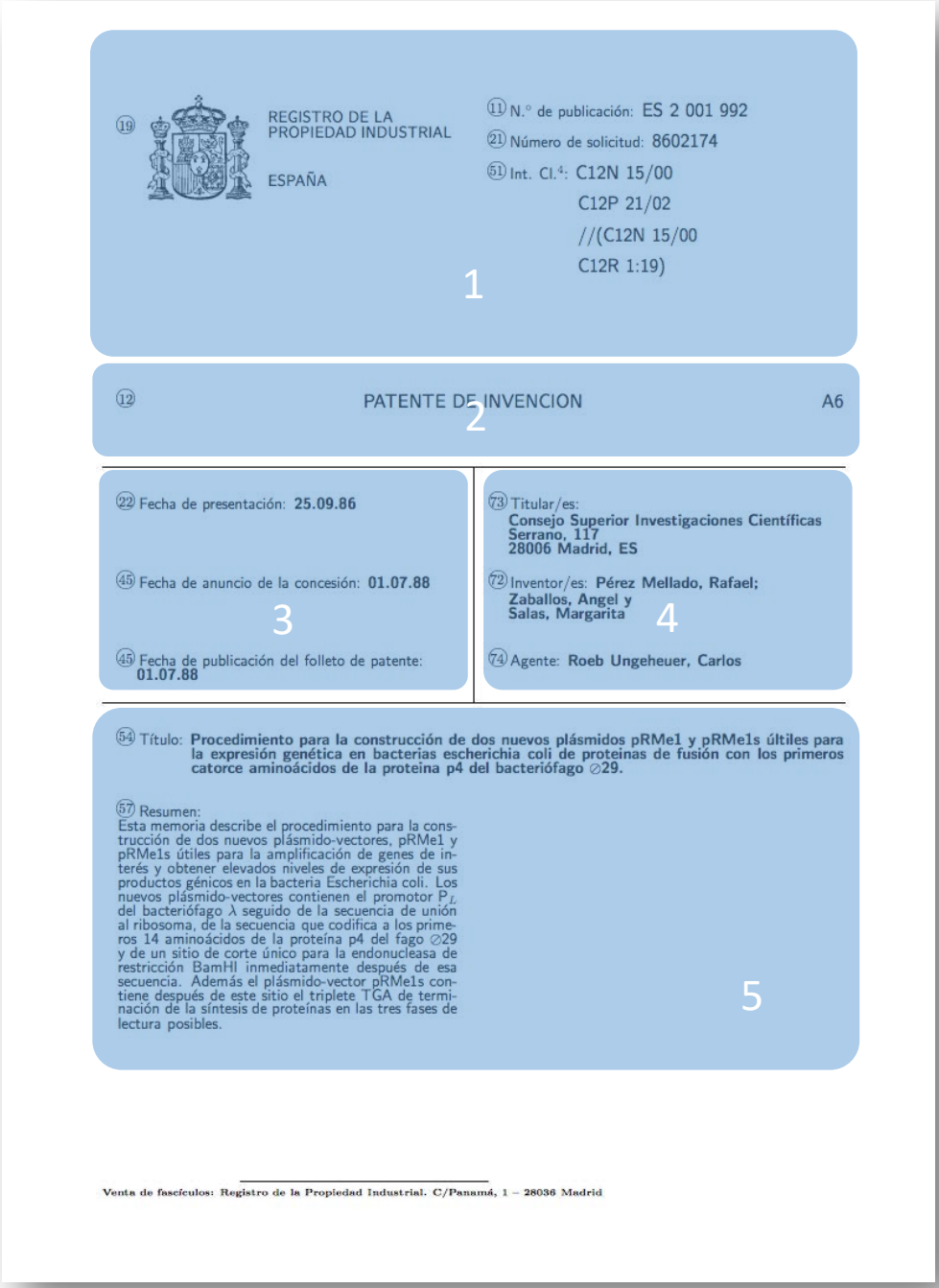
1. (19) Oficina u organización de registro y país de aplicación de la patente
   (11) Nº de la patente
   (21) Nº de solicitud
   (51) Códigos de la Clasificación Internacional
2. (12) Designación del tipo de documento
3. (27) Fecha de presentación
   (45) Fecha de concesión y exposición al público
4. (73) Nombre del titular, cesionario o propietario de la patente
   (72) Nombre del inventor
   (74) Nombre del agente
5. [54] Titulo de la invención
   [57] Resumen

Los códigos INID indicados en cada una de las secciones, y que se muestran en la patente, corresponden a la numeración indicada por la OMPI para la identificación de elementos bibliográficos de las patentes. Ésta es internacional y se sigue en todos los sistemas de patentes del mundo.

[33] Identificación Numérica Internacionalmente acordada en materia de Datos bibliográficos

[34] https://www.wipo.int/export/sites/www/standards/es/pdf/03-09-01.pdf

[35] Los códigos entre paréntesis (…) son de obligada aparición, aquellos entre corchetes […] son opcionales



(19) REGISTRO DE LA PROPIEDAD INDUSTRIAL

ESPAÑA

(11) N.º de publicación: ES 2 001 992

(21) Número de solicitud: 8602174

(51) Int. Cl.[4]: C12N 15/00
C12P 21/02
//(C12N 15/00
C12R 1:19)

1

(12) PATENTE DE INVENCION A6

2

(22) Fecha de presentación: **25.09.86**

(45) Fecha de anuncio de la concesión: **01.07.88**

3

(45) Fecha de publicación del folleto de patente: **01.07.88**

(73) Titular/es:
Consejo Superior Investigaciones Científicas
Serrano, 117
28006 Madrid, ES

(72) Inventor/es: **Pérez Mellado, Rafael;**
Zaballos, Angel y
Salas, Margarita

4

(74) Agente: **Roeb Ungeheuer, Carlos**

(54) Título: **Procedimiento para la construcción de dos nuevos plásmidos pRMe1 y pRMe1s últiles para la expresión genética en bacterias escherichia coli de proteinas de fusión con los primeros catorce aminoácidos de la proteina p4 del bacteriófago ∅29.**

(57) Resumen:
Esta memoria describe el procedimiento para la construcción de dos nuevos plásmido-vectores, pRMe1 y pRMe1s útiles para la amplificación de genes de interés y obtener elevados niveles de expresión de sus productos génicos en la bacteria Escherichia coli. Los nuevos plásmido-vectores contienen el promotor P_L del bacteriófago λ seguido de la secuencia de unión al ribosoma, de la secuencia que codifica a los primeros 14 aminoácidos de la proteína p4 del fago ∅29 y de un sitio de corte único para la endonucleasa de restricción BamHI inmediatamente después de esa secuencia. Además el plásmido-vector pRMe1s contiene después de este sitio el triplete TGA de terminación de la síntesis de proteínas en las tres fases de lectura posibles.

5

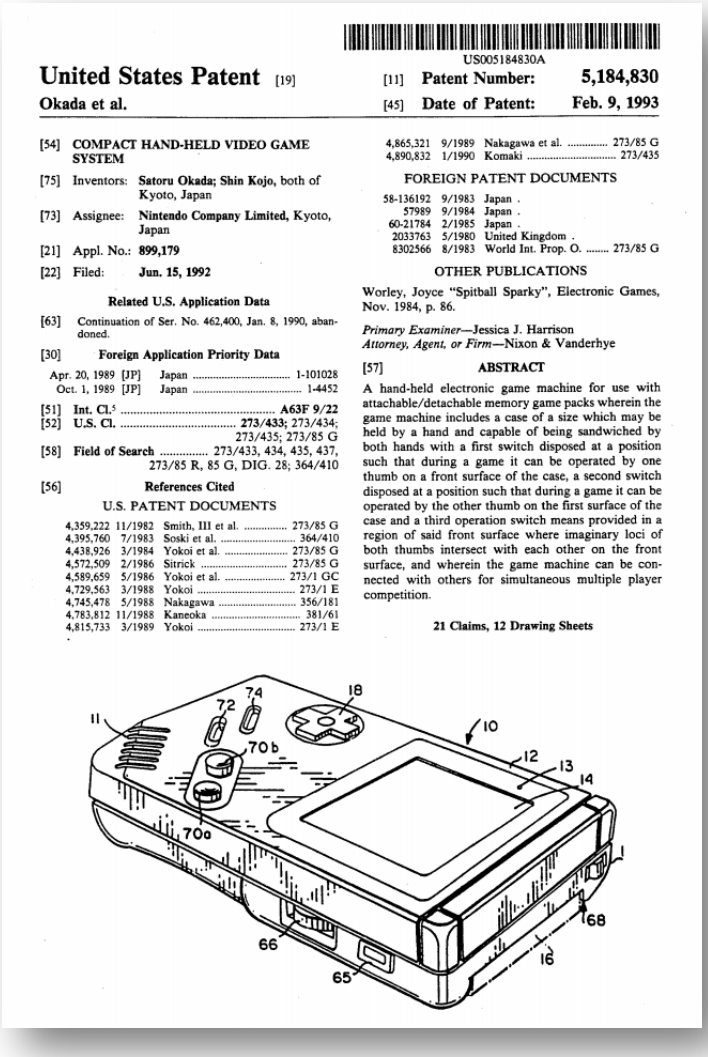
Venta de fascículos: Registro de la Propiedad Industrial. C/Panamá, 1 – 28036 Madrid

Figura 7: Patente Española Nº 2.001.992

Como se puede ver en la Figura 8, en algunos sistemas de patentes –el estadounidense en este caso– la primera página incluye datos bibliográficos [56] referentes a las citas tanto de otros documentos de patentes (nacionales e internacionales), como de otras publicaciones (llamadas citas no patentes, como libros o artículos científicos). En el caso del sistema español, las citas no patentes se encuentran en la sección de descripción.

Tras la primera página se accede al contenido descriptivo de la patente, desarrollado en las siguientes secciones principales:

- → **Sección de descripción**: explicación de la invención
- → **Sección de reivindicaciones**: definición de los límites del derecho de explotación exclusiva
- → **Sección de dibujos** (opcional)



US005184830A

United States Patent [19]

Okada et al.

[11] **Patent Number: 5,184,830**

[45] **Date of Patent: Feb. 9, 1993**

[54] **COMPACT HAND-HELD VIDEO GAME SYSTEM**

[75] Inventors: **Satoru Okada; Shin Kojo**, both of Kyoto, Japan

[73] Assignee: **Nintendo Company Limited**, Kyoto, Japan

[21] Appl. No.: **899,179**

[22] Filed: **Jun. 15, 1992**

Related U.S. Application Data

[63] Continuation of Ser. No. 462,400, Jan. 8, 1990, abandoned.

[30] **Foreign Application Priority Data**

Apr. 20, 1989 [JP] Japan 1-101028
Oct. 1, 1989 [JP] Japan 1-4452

[51] **Int. Cl.**[5] **A63F 9/22**
[52] **U.S. Cl.** ..... **273/433**; 273/434; 273/435; 273/85 G
[58] **Field of Search** ..... 273/433, 434, 435, 437, 273/85 R, 85 G, DIG. 28; 364/410

[56] **References Cited**

U.S. PATENT DOCUMENTS

4,359,222	11/1982	Smith, III et al.	273/85 G
4,395,760	7/1983	Soski et al.	364/410
4,438,926	3/1984	Yokoi et al.	273/85 G
4,572,509	2/1986	Sitrick	273/85 G
4,589,659	5/1986	Yokoi et al.	273/1 GC
4,729,563	3/1988	Yokoi	273/1 E
4,745,478	5/1988	Nakagawa	356/181
4,783,812	11/1988	Kaneoka	381/61
4,815,733	3/1989	Yokoi	273/1 E
4,865,321	9/1989	Nakagawa et al.	273/85 G
4,890,832	1/1990	Komaki	273/435

FOREIGN PATENT DOCUMENTS

58-136192	9/1983	Japan .	
57989	9/1984	Japan .	
60-21784	2/1985	Japan .	
2033763	5/1980	United Kingdom .	
8302566	8/1983	World Int. Prop. O.	273/85 G

OTHER PUBLICATIONS

Worley, Joyce "Spitball Sparky", Electronic Games, Nov. 1984, p. 86.

Primary Examiner—Jessica J. Harrison
Attorney, Agent, or Firm—Nixon & Vanderhye

[57] **ABSTRACT**

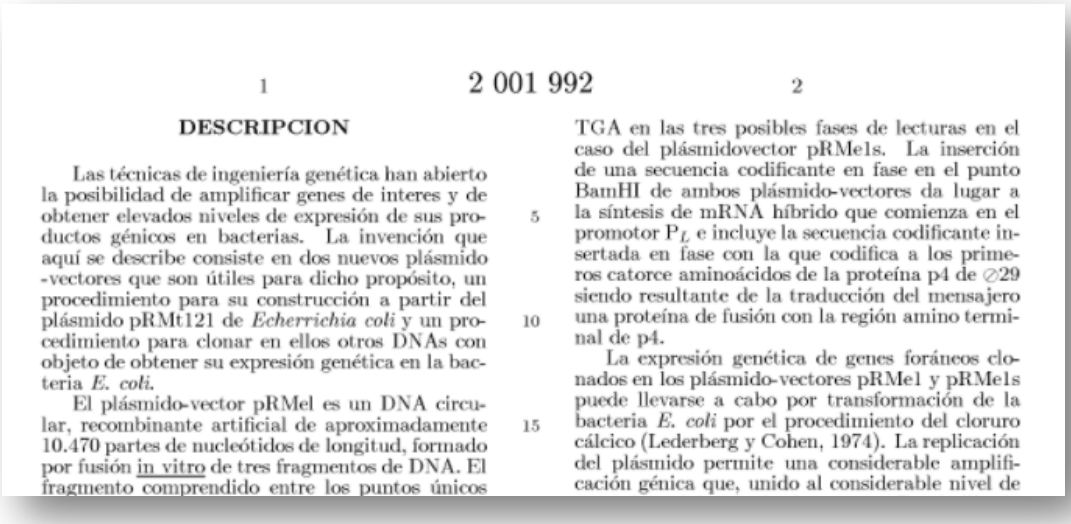
A hand-held electronic game machine for use with attachable/detachable memory game packs wherein the game machine includes a case of a size which may be held by a hand and capable of being sandwiched by both hands with a first switch disposed at a position such that during a game it can be operated by one thumb on a front surface of the case, a second switch disposed at a position such that during a game it can be operated by the other thumb on the first surface of the case and a third operation switch means provided in a region of said front surface where imaginary loci of both thumbs intersect with each other on the front surface, and wherein the game machine can be connected with others for simultaneous multiple player competition.

21 Claims, 12 Drawing Sheets

Figura 8: Patente Estadounidense Nº 5.184.830

En la sección de descripción se realiza una explicación detallada y pormenorizada de la invención. Incluyendo qué problema resuelve, cómo lo resuelve, cómo se ejecuta, cómo se utiliza, qué beneficios tiene y cómo se diferencia de inventos que pudieran parecerse.

Esta sección puede estar compuesta por tantas páginas como sea necesario, y siempre debe permitir entender con suficiente detalle (requisito para lograr la patente) el invento en si mismo, encontrándose redactado de la forma más clara y concisa posible.



1 2 001 992 2

DESCRIPCION

Las técnicas de ingeniería genética han abierto la posibilidad de amplificar genes de interes y de obtener elevados niveles de expresión de sus productos génicos en bacterias. La invención que aquí se describe consiste en dos nuevos plásmido -vectores que son útiles para dicho propósito, un procedimiento para su construcción a partir del plásmido pRMt121 de *Echerrichia coli* y un procedimiento para clonar en ellos otros DNAs con objeto de obtener su expresión genética en la bacteria *E. coli*.

El plásmido-vector pRMel es un DNA circular, recombinante artificial de aproximadamente 10.470 partes de nucleótidos de longitud, formado por fusión in vitro de tres fragmentos de DNA. El fragmento comprendido entre los puntos únicos TGA en las tres posibles fases de lecturas en el caso del plásmidovector pRMe1s. La inserción de una secuencia codificante en fase en el punto BamHI de ambos plásmido-vectores da lugar a la síntesis de mRNA híbrido que comienza en el promotor P_L e incluye la secuencia codificante insertada en fase con la que codifica a los primeros catorce aminoácidos de la proteína p4 de ∅29 siendo resultante de la traducción del mensajero una proteína de fusión con la región amino terminal de p4.

La expresión genética de genes foráneos clonados en los plásmido-vectores pRMe1 y pRMe1s puede llevarse a cabo por transformación de la bacteria *E. coli* por el procedimiento del cloruro cálcico (Lederberg y Cohen, 1974). La replicación del plásmido permite una considerable amplificación génica que, unido al considerable nivel de

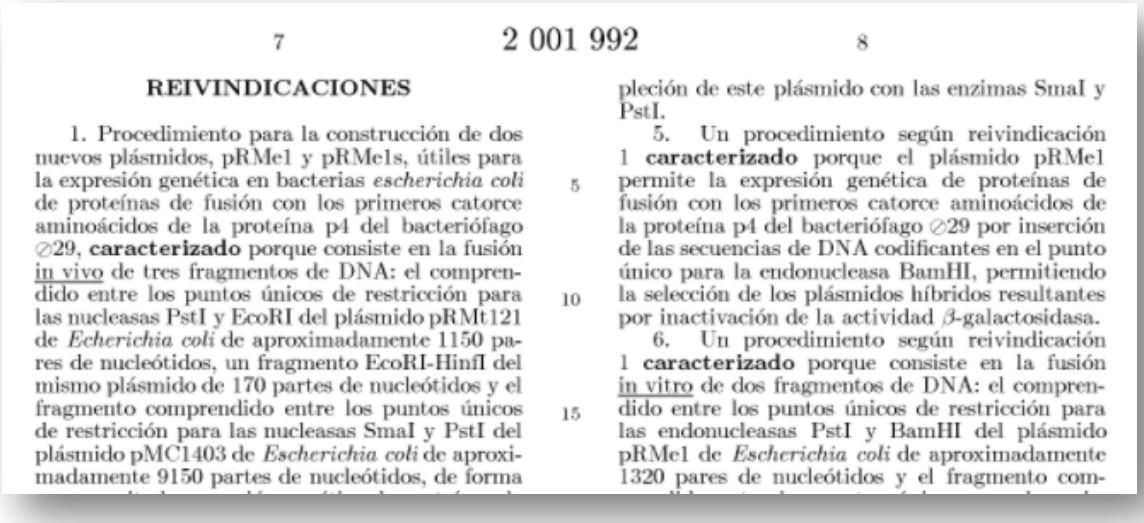
Figura 9: Sección descripción de la patente ES 2.001.992

Como se puede apreciar en la Figura 9, en la parte superior se encuentra el número de documento (esto se repite en todas las páginas que conforman el documento de la patente). Además, la sección de descripción cuenta con las siguientes subsecciones:

- → **Sector de la técnica**: ámbito de aplicabilidad y delimitación de la invención.
- → **Estado de la técnica**: recopilación de todos los antecedentes, variantes o soluciones que permitan entender la necesidad de la solución planteada y la novedad de la invención frente a lo ya conocido.
- → **Explicación de la invención**: mediante el detalle completo de las características de la invención que se plantea como solución al estado de la técnica indicada en el apartado anterior. Debe ser clara y minuciosa, permitiendo la reproducción por parte de un experto en la materia.
- → **Explicación del modo de realización**: descripción pormenorizada de –al menos un modo de– funcionamiento y ejecución de la solución. Se trata únicamente de información de los aspectos técnicos.
- → (Opcional) **Explicación de los dibujos**: mediante su descripción detallada.
- → **Explicación de la aplicación industrial**: en caso de no ser evidente la aplicabilidad industrial (requisito indispensable para la patentabilidad).

En el caso de ser una invención sobre un procedimiento microbiológico, la descripción deberá contener toda la información que el solicitante tenga sobre el mismo, así como depositar (no más tarde de la fecha de solicitud de la patente) en una institución autorizada, una muestra de dicho microorganismo siguiendo los convenios

internacionales. Indicando en la descripción la institución[36] en la que se encuentra el depósito y el número o clave de identificación.



7 2 001 992 8

REIVINDICACIONES

1. Procedimiento para la construcción de dos nuevos plásmidos, pRMe1 y pRMe1s, útiles para la expresión genética en bacterias *escherichia coli* de proteínas de fusión con los primeros catorce aminoácidos de la proteína p4 del bacteriófago ∅29, **caracterizado** porque consiste en la fusión in vivo de tres fragmentos de DNA: el comprendido entre los puntos únicos de restricción para las nucleasas PstI y EcoRI del plásmido pRMt121 de *Echerichia coli* de aproximadamente 1150 pares de nucleótidos, un fragmento EcoRI-HinfI del mismo plásmido de 170 partes de nucleótidos y el fragmento comprendido entre los puntos únicos de restricción para las nucleasas SmaI y PstI del plásmido pMC1403 de *Escherichia coli* de aproximadamente 9150 partes de nucleótidos, de forma pleción de este plásmido con las enzimas SmaI y PstI.

5. Un procedimiento según reivindicación 1 **caracterizado** porque el plásmido pRMe1 permite la expresión genética de proteínas de fusión con los primeros catorce aminoácidos de la proteína p4 del bacteriófago ∅29 por inserción de las secuencias de DNA codificantes en el punto único para la endonucleasa BamHI, permitiendo la selección de los plásmidos híbridos resultantes por inactivación de la actividad β-galactosidasa.

6. Un procedimiento según reivindicación 1 **caracterizado** porque consiste en la fusión in vitro de dos fragmentos de DNA: el comprendido entre los puntos únicos de restricción para las endonucleasas PstI y BamHI del plásmido pRMe1 de *Escherichia coli* de aproximadamente 1320 pares de nucleótidos y el fragmento com-

Figura 10: Sección indicaciones de la patente ES 2.001.992

La Figura 10 muestra la sección de reivindicaciones, que aparece en el documento de la patente tras la de descripción, y que permite definir claramente el objeto para el que se solicita la protección, siendo la parte jurídicamente más relevante de la patente dado que sólo se protege lo indicado en esta sección. Las reivindicaciones deben encontrarse sustentadas en la información contenida en la sección de descripción y los dibujos (si los hubiera).

Las reivindicaciones deben contener:

- → **Preámbulo**: introduciendo el objeto de la invención y todas las características técnicas conocidas que definen los elementos a proteger.
- → **Parte caracterizadora**: especificando pormenorizadamente todas las características técnicas nuevas a proteger. Esta parte se encuentra diferenciada por el uso de las expresiones "caracterizado por", "que comprende" (utilizadas cuando existen elementos técnicos extra) y "que consiste en" (que excluye otros elementos). Existen dos tipos de reivindicaciones, independientes o esenciales (definen la invención de forma general y no hacen referencia a otras reivindicaciones) y dependientes (contienen las características adicionales de las reivindicaciones de las que dependen).

[36] En España existen dos sedes: la Colección Española de Cultivos Tipo, con sede en la Universidad de Valencia, y el Banco Español de Algas, con sede en la Universidad de Las Palmas, Gran Canaria.

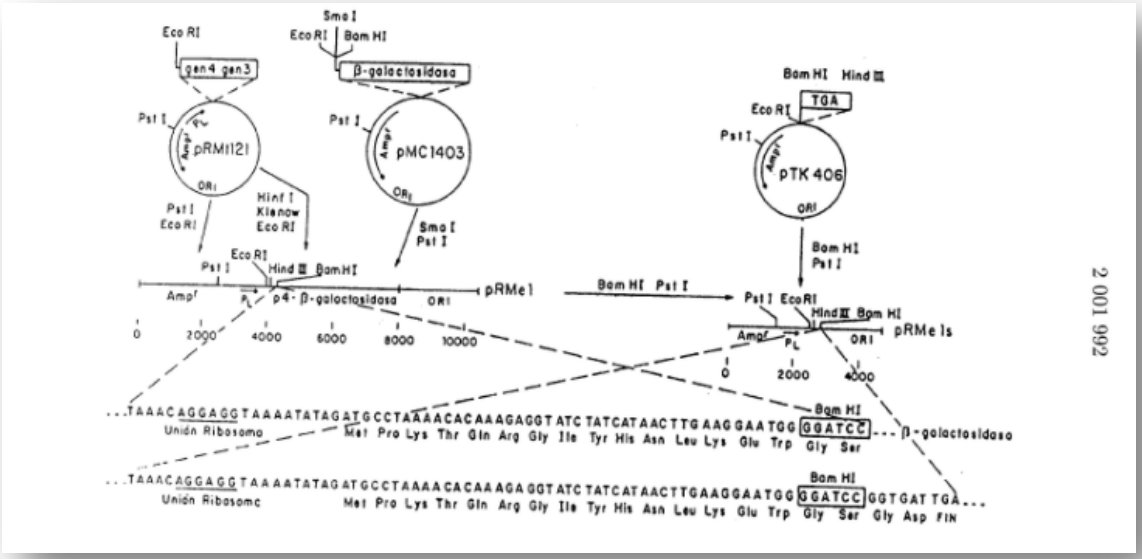


Figura 11: Dibujo de la patente ES 2.001.992

Por último, en un documento de patente podemos encontrar la sección opcional de dibujos como el que se muestra en la Figura 11. Ésta se utiliza para facilitar la comprensión de la solución mediante el uso de la representación gráfica y esquemática del invento. Se pueden utilizar tantos dibujos como sea preciso siempre que se encuentren referenciados en la sección de descripción.

Además, existen otros documentos como el Informe sobre el Estado de la Técnica para Información Tecnológica[37] y una opinión escrita (emitido por el OEPM, de tipo preliminar y no vinculante) relativos a la solicitud, así como posibles comunicaciones o enmiendas que permiten delimitar perfectamente el alcance de la invención.

Es la unión de toda la documentación que conforma una patente, las secciones (desde la portada hasta los dibujos) junto con los informes técnicos que pueda resolver la oficina en la que se presenta la solicitud, lo que confieren la importancia al propio documento de la patente. Se trata de un documento que contiene información científico-técnica, jurídica y comercial de gran relevancia, que recopila literatura de tipo no patente para sustentar esa información, y que representa un punto fijo en la evolución del conocimiento del mismo modo que puede serlo un libro o un artículo de investigación.

[37] En otros sistemas de patentes las solicitudes también se encuentran informadas se forma similar, aunque los informes y formas de adjuntar al documento de la patente varían.

2.1.5. La economía de las patentes: Uso y explotación

Según los datos publicados por la OMPI la evolución, tanto en las solicitudes de protección (Figura 12) como en las concesiones en los últimos 20 años (Figura 13), ha seguido una tendencia al alza con un incremento en ambos casos que supera el 200% entre el primer y último año.

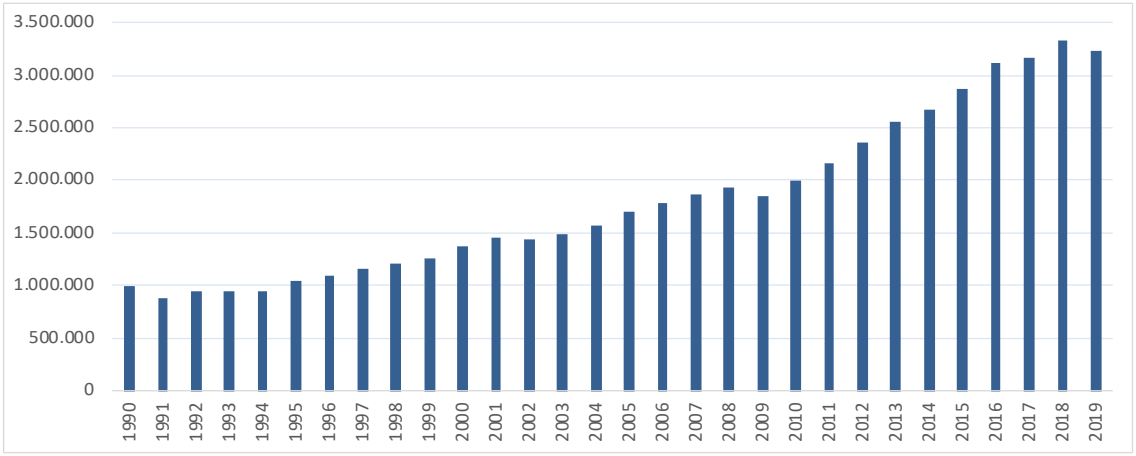


Figura 12: Evolución de solicitudes de patente a nivel mundial 1990 – 2019 [Fuente: WIPO IP Statistics Data Center]

Se debe tener en cuenta para poder poner en valor el incremento de solicitudes de patentes que las invenciones requieren de grandes inversiones tanto en investigación como en desarrollo, sumando a esto la inversión necesaria que la solicitud de una patente y su concesión requieren, junto con años extra de gestiones y desembolsos económicos muy importantes para las empresas como se ha visto en el §2.1.3.

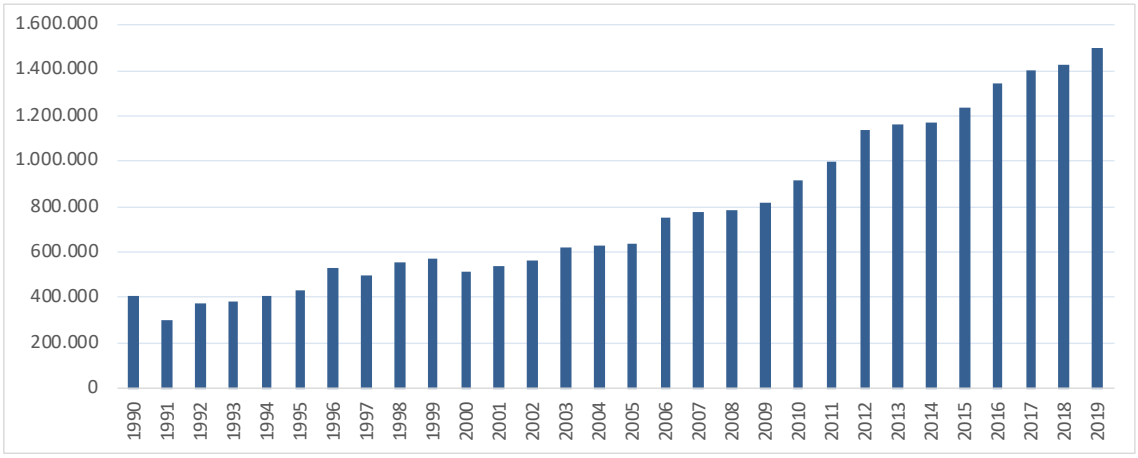


Figura 13: Evolución de patentes concedidas a nivel mundial 1990 – 2019 [Fuente: WIPO IP Statistics Data Center]

El incremento global observado se da en la mayoría de Oficinas del mundo, en 2019 las cinco Oficinas más importantes por cantidad de solicitudes recibidas fueron China (1,4 Millones), Estados Unidos (621.432), Japón (307.969), Corea del Sur (218.975) y EPO (181.479) (WIPO, 2010). Pese a que existe un ligero descenso de solicitudes en 2019 (-3% total debido a un menor número de solicitudes principalmente en China, Japón, Rusia y Reino Unido), es el año que más patentes se conceden, alcanzando la cifra total de 1.500.900 en todo el mundo. Además, pese a que una patente se puede mantener

hasta 20 años después de la fecha de solicitud no todos los titulares de patentes mantienen sus derechos hasta el final de la vida útil de la patente. Pese a ello, en 2019 la OMPI estima que 14,9 millones de patentes se encontraban vigentes en 127 Oficinas, siendo esto un incremento del 7% con respecto al año anterior.

Aunque se podría esperar que únicamente las grandes empresas se vieran beneficiadas por el uso de las patentes, en Europa en 2019 el 18% y el 10% de las patentes fueron solicitadas por PYMES o individuales, y universidades y organizaciones públicas de investigación respectivamente. Es un dato a remarcar que 1 de cada 5 solicitudes fueran realizadas por PYMES ya que esto denota que las pequeñas y medianas empresas consideran útiles las patentes (EPO, 2019).

Estas cifras no hacen más que apoyar los diversos motivos por los que la gente solicita la protección de sus invenciones. Estos motivos se pueden agrupar en tres categorías (Battke, Schmidt, Stollenwerk, & Hoffmann, 2016; De Rassenfosse, Palangkaraya, & Webster, 2016; Gifford, 2004; Giménez, 2018; Jensen, Palangkaraya, & Webster, 2015; Kani & Motohashi, 2012; Langinier & Moschini, 2002; McDonald, 2015; Nemet & Johnson, 2012; Oehmke, 2006; Park, 2008; Sampat, 2018) principales que se encuentran desarrolladas en el resto del capitulo:

- → Económicos
- → Valor corporativo
- → Transferencia

Motivos económicos

Hasta ahora uno de los principales acicates comentados en los anteriores capítulos es el de buscar la protección para evitar la usurpación de la misma. De este modo, el titular de la patente puede ejercer su derecho exclusivo para poder recuperar la inversión en investigación realizada mediante la explotación de la idea, directamente o mediante licencias o concesiones. Esto permite impulsar la solicitud de las propias patentes, ya que sin los beneficios que aporta la titularidad de una patente es más difícil ganar batallas legales para evitar posibles fraudes (robo de ideas, espionaje industrial, etc.).

Junto con los motivos principales expuestos anteriormente, se puede encontrar los descritos a continuación, que resumen las ventajas que puede llegar a tener la concesión de una patente para una empresa en el ámbito económico:

- → **Preservar la libertad de operación**: al ostentar el monopolio temporal de explotación de la invención, la empresa puede extraer beneficios de no tener competidores en el mercado.
- → **Licenciar**: como se ha explicado anteriormente, mediante el uso de licencias para la explotación vía un tercero, la empresa puede generar beneficios sin necesidad de incurrir en ciertas desventajas.
- → **Vigilancia tecnológica**: controlando las tendencias de publicación de la competencia se puede conocer las posibles futuras publicaciones de las empresas, lo que permitirá adelantarse y ofrecer a los clientes una contrapartida.

- → **Infracciones de terceros**: evitando la posible usurpación o copia de la patente, de modo que se evite la pérdida de ingresos por una existencia ilícita en el mercado.

Estos motivos económicos tienen una base de aplicación muy importante en las pequeñas y medianas empresa (PYMEs), ya que ellas especialmente tienen un 21% más de probabilidades de experimentar un periodo de crecimiento tras la búsqueda de la protección intelectual (EUIPO, 2019). Es por esto por lo que el 32% de las empresas considera de "gran importancia" las patentes para proteger y generar valor económico en la empresa, existiendo un 60% de empresas que consideran que sus derechos de protección intelectual han tenido un impacto "muy positivo" o "positivo".

Valor corporativo

Lograr una imagen positiva de la empresa es algo que todas las compañías buscan, se plantean estrategias de específicas del negocio buscando sobresalir frente a la competencia gracias a sus características únicas; para alcanzar este objetivo tener una estrategia de protección industrial trabajada y lograr concesiones de patentes puede ayudar aportando valor al portafolio de la entidad.

Del mismo modo que en la Academia se busca alcanzar la excelencia mediante el engranaje de investigación, docencia y transferencia, las empresas pueden aumentar su valor tanto con activos intangibles (explotación, licencias, cesiones, etc.) como con pasivos intangibles (la información contenida en las patentes). Actualmente, los activos intangibles suponen más del 80% del valor del mercado de las empresas S&P 500, cuando hace tan solo cuarenta años suponían únicamente el 17% (Aon, 2019).

De este modo, las empresas pueden situarse a la vanguardia del sector en el que se desarrollen y posicionarse como pioneros o expertos en su invención. Gracias a esto, pueden bloquear el acceso al mercado de la competencia, mejorando su posición en el mercado, en las negociaciones que pueda requerir sus negocios e incluso en la búsqueda y consecución de financiación[38].

Transferencia de conocimiento

La transferencia de conocimiento es el proceso mediante el cual se transmite los conocimientos científicos y/o tecnológicos, tecnologías, saber hacer (o know-how) desde una organización a otra. Este proceso se aplica especialmente en el transvase de conocimiento que se da entre la Universidad y las Empresas, pero no está limitado únicamente a este flujo de información ya que puede tener cualquier tipo de organización a ambos lados de los extremos (universidades, centros de investigación, laboratorios, centros tecnológicos, todo tipo de empresas, gobiernos, etc.).

[38] https://www.wipo.int/wipo_magazine/es/2008/05/article_0001.html

Este proceso se considera especialmente crítico y necesario para el impulso de la innovación y economía –del mismo modo que sucede con las patentes– y tiene dos tipos de forma de comunicación[39]:

- → **De tipo oficial**: cesión de licencias o derechos, contratos de colaboración, acuerdos de transferencia, contratos de investigación o consultoría, franquicias, start-ups o spin-offs.
- → **De tipo informal**: movilidad de capital humano, publicaciones y comunicaciones científicas, docencia, conferencias y seminarios, etc.

Teniendo en cuenta la propia naturaleza de las patentes (el inventor recibe protección a cambio de indicar detalladamente cómo se logra el invento), éstas son un punto clave en los dos tipos de transferencia (Benson & Magee, 2015) ya que, bien por motivos económicos, bien para dar a conocer la invención, permiten divulgar el conocimiento que contienen.

Gracias a esta divulgación, se puede realizar búsquedas para dar con respuestas a problemas técnicos mediante soluciones alternativas, es decir, pueden encontrarse soluciones para los problemas de un sector mediante la aplicación de las invenciones logradas en otro.

Uno de los movimientos más importantes dentro de la transferencia de conocimiento, es el descrito por Etzkowitz y Leydersdorff conocido como el modelo Triple Hélice. Este modelo busca impulsar las interacciones entre la Academia, la Industria y los Gobiernos para fomentar el desarrollo económico y social (E. G. Campbell, Powers, Blumenthal, & Biles, 2004; Czarnitzki, Hussinger, & Schneider, 2012; Etzkowitz & Leydesdorff, 1995; Gkoumas & Christou, 2020; Karytinos & Ingham, 2015; Leydesdorff, 2012; López Jiménez & Dittmar, 2019; Meyer, Siniläinen, & Utecht, 2003). El modelo Triple Hélice señala que la sinergia entre las tres vertientes permite transferir el conocimiento otorgando valor a la hibridación de las entidades y borrando los límites de sus roles básicos. De este modo, se impulsa la participación de las Universidades en actividades comerciales mediante la creación de patentes y sus licitaciones (entre otras). El modelo fue posteriormente revisado y ampliado por otros autores.

Junto con las motivaciones para solicitar la protección de una patente, se puede encontrar el otro extremo del uso y explotación de las patentes: su uso en materia de inteligencia competitiva mediante la previsión tecnológica y la patentometría.

Previsión tecnológica (Technological forecasting)

Del mismo modo que se aplican modelos de predicción económica, previsiones de mercado e incluso la previsión del tiempo, las patentes pueden ser usadas como herramientas para intentar prever el futuro tecnológico gracias a su contenido. Pese a que al mercado un producto final radicalmente diferente o innovador puede llegar sin que los usuarios lo esperen, éste es la consecución de años de investigación y desarrollo a varios niveles de toda la tecnología que lo compone. Por ello, realizar un seguimiento

[39] https://www.wipo.int/about-ip/es/universities_research/ip_knowledgetransfer/faqs/

de control de los movimientos que existen en la concesión de patentes puede ayudar a entender o visionar el futuro del campo estudiado (R. S. Campbell, 1983; H. Chen, Zhang, Zhu, & Lu, 2017; H. Chen et al., 2017; Kaya Firat, Madnick, & Lee Woon, 2008; Sungjoo Lee, Yoon, & Park, 2009; Li, Xie, Jiang, Zhou, & Huang, 2018; Noh, Song, & Lee, 2016; Segev & Kantola, 2012; Trappey, Trappey, Wu, & Lin, 2012; Yoon & Park, 2004; Zhou, Zhang, Porter, Guo, & Zhu, 2014).

La aplicación de esta técnica requiere de expertos y modelos de previsión complejos, así como de un conocimiento tanto de las patentes y el funcionamiento de los sistemas de protección, como de las áreas de aplicación. Los métodos de previsión se pueden clasificar según la aplicación en 9 categorías:

- → Opiniones de expertos
- → Análisis de tendencias
- → Métodos de monitorización e inteligencia
- → Métodos estadísticos
- → Modelado y simulación
- → Escenarios
- → Métodos de evaluación, decisión y económicos
- → Métodos descriptivos y de matrices
- → Creatividad

La técnica no es perfecta y dependerá de la construcción de los modelos y la información recogida obtener resultados de calidad, pero aun con los mejores expertos y modelos ésta puede fallar debido a cambios inesperados (por ejemplo, la aparición de una pandemia mundial que paraliza la producción normal en todo el mundo).

Patentometría

La aplicación de análisis estadísticos y matemáticos a la información de las patentes se conoce como patentometría. Esta técnica, subdisciplina de la bibliometría, se centra en el estudio de los indicadores, generales o específicos, de los diferentes tipos de datos de patentes. La tipología de estos indicadores varía según su aplicación:

- → De actividad: número de patentes, distribución, países, etc.
- → Relacionales: vínculos entre empresas o investigadores
- → Relaciones de segunda generación: coocurrencias, palabras clave, título, resumen, etc.
- → Agrupación: familias de patentes, clustering, mapeo tecnológico, etc.

La patentometría es utilizada principalmente para establecer relaciones entre ciencia y técnica (Altuntas, Dereli, & Kusiak, 2015; H. Chen et al., 2017; Harhoff, Narin, Scherer, & Vopel, 1999; Meyer, 2000a; Narin, 1994; Sarin et al., 2020), pero tiene multitud de aplicaciones:

- → Identificación de tecnologías, autores, organizaciones, países y sectores
- → Identificación de competencia, socios, productos
- → Evolución y rastreo de la técnica
- → Caracterización de organizaciones

- → Evaluación de la ciencia y todos sus actores
- → Definición de procesos de difusión
- → Definición del ciclo de vida tecnológico

La consecución de este tipo de análisis requiere de herramientas y modelos aplicados a varios niveles. Primero para la monitorización de la información, ya que, aunque las patentes se encuentren en bases de datos con sus propias estructuras, éstas deben ser entendidas y debe ser posible su análisis, ya que cada base de datos tiene sus particularidades. Además, actualmente el volumen de información es tan grande que son necesarios espacios de almacenamiento del orden de petabytes.

Ese mismo volumen de información que puede dificultar el almacenamiento plantea un problema en el momento de gestión de la información, tanto para su preparación como para su posterior análisis, ya que serán necesarias herramientas y equipos especializados capaces de realizar grandes cargas de trabajo.

Por último, pese a que el análisis de los resultados puede parecer sencillo, su aplicación requiere del conocimiento del entorno del sector y la tecnología, por lo que será necesaria la contextualización mediante un experto.

Es por estos motivos que se trata de un área explorada de forma tibia, ya que en la actualidad únicamente se recopilan en Scopus[40] 80 documentos que contienen la palabra "*Patentometrics*"[41] en el título, resumen o palabra clave. La Figura 14, realizada mediante la herramienta VOSviewer, muestra una representación de coocurrencias de palabras clave, de este modo se puede apreciar las áreas cubiertas por la patentometría. Como se puede ver en la imagen se encuentra altamente relacionada con tres áreas principales:

- → **Economía y ciencia**: Análisis de patentes, ratio de crecimiento, ingeniería, inventor, logros, competitividad, tecnología
- → **Patentes e invenciones**: análisis de redes sociales, desarrollo tecnológico, minería de datos, ingeniería industrial.
- → **Bibliometría**: cibermetría, índice-h, bibliometría, innovación tecnológica, flujo del conocimiento.

Un análisis patentométrico puede ser muy útil, tanto como análisis único como combinado con otras técnicas, para la explotación de la información contenida en las patentes. Las empresas se pueden beneficiar de la aplicabilidad tanto del *technological forecasting* como de la patentometría para posicionar tecnológicamente mediante el uso de la inteligencia competitiva y traducir el conocimiento en capital.

[40] https://www.scopus.com/

[41] Los datos son una infraestimación, ya que la base de datos Scopus es restrictiva y, además, los autores pueden haber utilizado otros términos para la referencia al uso de la técnica.

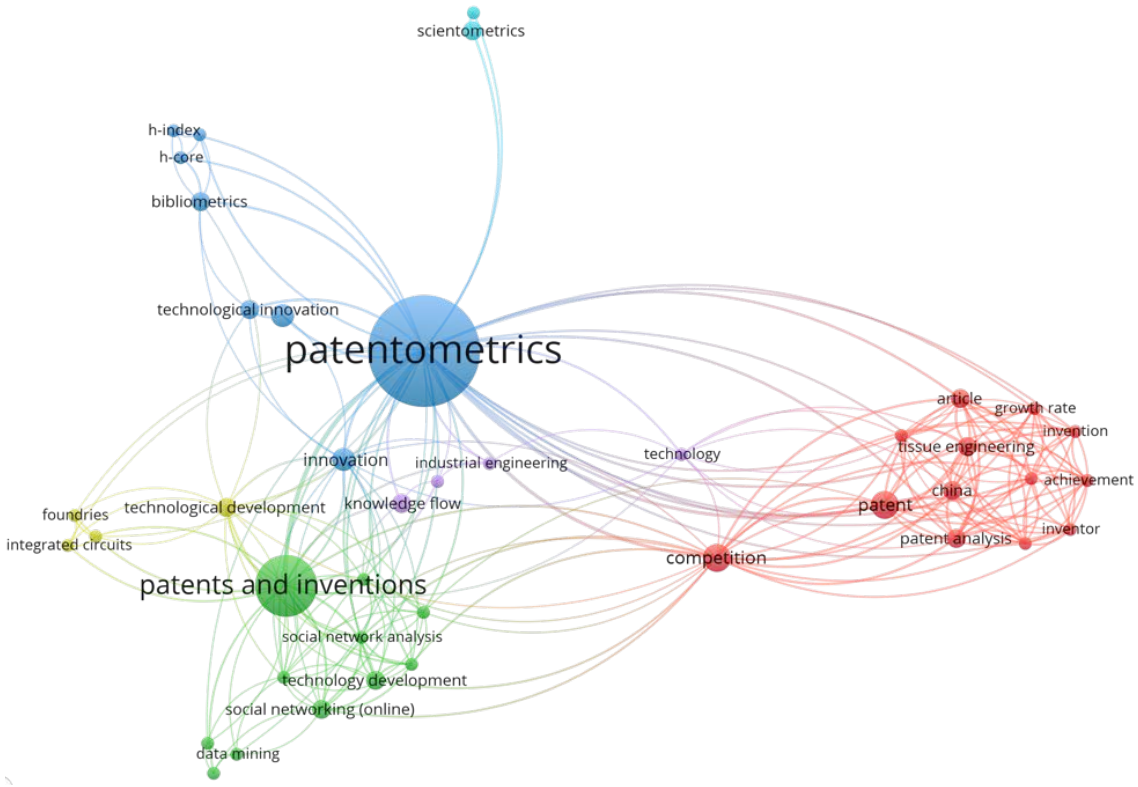


Figura 14: Análisis de concurrencias de palabras clave en documentos relacionados con "Patentometrics"
Fuente: Scopus. Visualización mediante VOSviewer (vosviewer.com)

2.1.6. Explotación de datos de patentes

Hasta ahora se ha focalizado la importancia de las patentes en su primera función: proteger al inventor y su idea; Pero como se comenta en el §2.1, las patentes tienen una segunda función: poner a la disposición de todo el mundo el conocimiento generado con la invención. Como se explica en el capitulo anterior, con esto se busca estimular la innovación y contribuir a mejorar el crecimiento económico.

Pero los documentos de patentes son contenedores, no sólo de la información relativa a una invención, sino también son un vínculo con el conocimiento pasado sobre el que se cimientan, generando un flujo de conocimiento mediante la utilización de citas y referencias para justificar las descripciones y reivindicaciones; junto con las menciones a patentes (incluidas por la Oficina en los informes que acompañan a la documentación) con las que pueden existir vínculos por el contenido.

Esto implica que partiendo desde una sola patente y realizando un estudio de las citas a otras patentes y tipos de documentos referenciados se pueda lograr una fuente de información, que permita en sí misma o complementando una búsqueda típica en fuentes bibliográficas tradicionales (artículos, revistas, libros, etc.), alcanzar a tener una imagen completa de la invención y todo su desarrollo. Las patentes son realmente útiles en la búsqueda de información, tanto es así, que el 80% de las patentes recogen información exclusiva y de calidad académica que no se encuentra publicada en otros lugares (Singh, Chakraborty, & Vincent, 2016).

Asimismo, para que una patente pueda ser concedida ésta debe cumplir con el criterio de novedad; y la novedad, pese a que la patente sea una protección territorial en el país en el que se concede, se aplica en todo el mundo. Por lo tanto, para que las Oficinas puedan comprobar el criterio de novedad de una patente, y el resto de población localizar el conocimiento concentrado en ellas para sus diversos usos deben encontrarse a disposición de todos en diversas fuentes de información.

Existen multitud de bases de datos de patentes que permiten realizar las búsquedas necesarias para lograr multitud de aplicaciones prácticas. Junto con todo lo descrito en el §2.1.5 se debe tener en cuenta que el conocimiento incluido en una patente permite principalmente:

- → Determinar la patentabilidad de las futuras invenciones
- → Estimar el valor de una patente
- → Evitar duplicidad de esfuerzos en investigación y desarrollo
- → Evitar infracciones de uso de patentes
- → Lograr información sobre las líneas de trabajo de la competencia
- → Explotar la información de patentes no concedidas, con aplicación en otros países o caducadas
- → Identificar tendencias y tecnologías emergentes
- → Búsqueda de competidores o socios

Debido a la importancia de que las patentes se encuentren disponibles al servicio de la sociedad, la mayoría de las Oficinas nacionales, regionales, así como la OMPI, tienen sus propias páginas y bases de datos ofreciendo acceso a la información de las patentes. En el Anexo I se recoge un listado de enlaces a las bases de datos nacionales y regionales.

El principal problema que presentan estas bases de datos es la limitación geográfica en sí misma, no existe una base de datos que contenga todas las patentes debido a que no existe una Oficina que haya concedido todas las patentes, por lo que será necesario visitar varias para lograr la máxima cobertura posible. En caso de que el procedimiento se encuentre realizado mediante la vía PCT (§2.1.3), la OMPI ofrece en su base de datos (Patentscope[42]) acceso a todas las colecciones de patentes solicitadas "internacionalmente".

Junto con el problema de la cobertura, se encuentra el problema del idioma en caso de que no haya traducciones de los documentos presentados por el titular de la patente, aunque algunas bases de datos y herramientas contienen traductores especializados. Además, existen aplicaciones o bases de datos de tipo comercial, que ofrecen otras herramientas y servicios que pueden complementar a la información de las propias patentes, aportando un alto valor con información relativa a informes empresariales, licencias, etc.

Existen multitud de bases de datos con contenido de patentes que pueden catalogarse en función de las características o tipología de la base de datos, pero existen una serie

[42] https://patentscope.wipo.int/

de diferencias que se deben tener en cuenta para poder localizar las información necesaria para la realización de estudios concretos (Alvarez Gil, L.; Albarez Gonzalez, M.; Contreras Villavicencio, 2016; Archontopoulos, 2004; Arias, 2003; de Rassenfosse, Dernis, & Boedt, 2014; Manglano & Zulueta, 2008; Raturi, Sahoo, & Tiwari, 2012; White, 2010; WIPO, 2012):

- → Nacionales o internacionales: dependiendo de si la base de datos cuenta únicamente con patentes de un Estado (nacional) o varios (internacional).
- → Públicas o privadas: dependiendo de si es necesario o no realizar algún pago por el acceso a la información.
- → Bibliográficas: contienen la información relativa a los datos de la patente recogidos en los códigos INID.
- → Documentales: contienen la propia patente y documentos asociados. Algunas bases de datos recogen únicamente la patente, en otras se recoge también los exámenes, solicitudes, comunicaciones, etc.
- → Texto completo: no todas las patentes que contienen la solicitud o concesión de la patente publican el texto completo de la misma. En los últimos años se han popularizado, permitiendo la consulta del documento al completo, incluyendo imágenes y dibujos.
- → Temáticas: bases de datos especializadas en un área de conocimiento o aplicación.
- → Cobertura temporal: no todas las bases de datos contienen todas las patentes publicadas. Muchas Oficinas realizan esfuerzos para ofrecer la mayor cantidad posible. Por ejemplo, en algunos casos las patentes antiguas se pueden encontrar fotografiadas.
- → Actualizaciones: las bases de datos se pueden actualizar a diario, semanal, quincenal, mensual, bimestral, trimestral o anualmente. Algunas bases de datos únicamente indican "*regularmente*".
- → Funcionalidades: cada base de datos puede tener su propio sistema de búsqueda. En algunos casos se trata de formatos complejos que requiere de conocimiento técnicos (i.e: lenguajes especializados como SQL).
- → Modos de acceso: además de contar todas ellas con interfaz web, algunas también disponen de API para el acceso mediante herramientas externas o descargas masivas, incluso en formatos accesibles como XML.

Existen multitud de bases de datos, cada una con sus propias particularidades. A continuación, se puede encontrar un listado con las más populares (el listado completo de bases de datos nacionales y regionales se encuentra en el §Anexo I):

- → Espacenet (EPO) https://worldwide.espacenet.com/
- → Patentscope (OMPI) https://patentscope.wipo.int/search/es/search.jsf
- → USPTO (Oficina EEUU) http://patft.uspto.gov/netahtml/PTO/search-bool.html
- → INVENES (Oficina España) http://invenes.oepm.es

Además, como se indica anteriormente, existen proveedores privados de bases de datos, la mayoría con información agregada, que ofrecen acceso a las patentes. La Tabla 5 recoge algunas de estas bases de datos, descritas con la categorización explicada con

anterioridad en relación con las patentes que recopilan. Todas las bases de datos incluidas en la tabla:

- → contienen (directa o indirectamente) el texto completo de los documentos de patentes
- → son de tipo internacional
- → contienen todas las áreas o temáticas de patentes
- → sus coberturas temporales vienen delimitadas por las Oficinas desde la que se extrae la información

Tabla 5: Bases de datos de patentes propietarias
Fuente: elaboración propia

Fuente	URL	Acceso	Funcionalidades	Modo de acceso
Derwent	https://clarivate.com	Pago	Indicadores propios, Informes y datos brutos exportables	Web, API
Dialog	https://dialog.com	Pago	Indicadores propios, Paneles informativos personalizables	Web
Questel-Orbit	https://www.orbit.com	Pago	Combinación de fuentes, paneles de visualizacion personalizables, consultoria	Web
STN International	https://www.stn-international.com	Pago	Diferenciación por idioma y área	Web
Scopus	https://www.scopus.com	Pago	No	Web, API
Dimensions	https://www.dimensions.ai	Pago	Referencias	Web, API
Lens	https://www.lens.org	Gratis	Información complementaria	Web, API pago
Google Patents	https://patents.google.com	Gratis	No	Web

Es cierto que las bases de datos no son perfectas y existen una serie de problemas que se deben tener en cuenta en el momento de realizar una búsqueda:

- → **Desambiguación**: en los nombres de inventores, organizaciones, etc. Del mismo modo que sucede en otros ámbitos, pueden existir varias formas de mencionar a una misma entidad.
- → **Duplicados**: en caso de estudiar desde diferentes bases de datos en las que existan registros de diversos países se debe tener en cuenta que la misma patente puede aparecer con diferentes números.
- → **Errores de traducción**: en ocasiones pueden aparecer problemas con traducciones realizadas automáticamente por el sistema.

- **Errores de reproducción**: en caso de que la información sea extraída mediante ficheros XML o lectura de API, se debe tener en cuenta que pueden existir errores de lectura (sobre todo en los documentos anteriores a 2010) ya que los documentos son leídos directamente por el sistema (desde imagen o pdf) y los sistemas pueden cometer errores.

Estrategias de búsqueda:
Queda patente que la información recogida en las bases de datos resulta de especial interés, incluso cuando no se está realizando un estudio sobre patentes. Pero como se ha visto, existen multitud de bases de datos y sistemas por lo que es necesaria una estrategia de búsqueda adecuada. Los criterios más relevantes para realizar una búsqueda de patentes son:

- **Palabras clave**: la mayoría de bases de datos permiten el uso de operadores para refinar la búsqueda mediante las palabras más importantes. Ejemplos: "AND", "OR", "NOT", mediante anidado con paréntesis o con comillas para buscar exactamente.
- **Temática**: mediante el sistema de clasificación internacional se puede buscar todas aquellas patentes que se encuentren en el mismo grupo.
- **Fechas**: las patentes tienen varias fechas importantes en su vida (solicitud, concesión, publicación). Se debe tener cuidado ya existen varios formatos de fecha utilizados.
- **Número de patente**
- **Nombre de inventor u organización**: aunque se debe tener cuidado con los posibles errores que producirse (i.e.: abreviaciones, erratas, reproducción, etc.)
- **Campos específicos**: algunas bases de datos permiten buscar seleccionando el campo exacto en el que debe aparecer la palabra clave.
- **Citas**: todos los documentos citados en una patente deben ser clasificados según su importancia y tipología. Las diferentes categorías pueden ser utilizadas conjuntamente para ofrecer más información. A continuación, se indican las más relevantes para realizar una búsqueda (existiendo otras como P, L, E, O, etc.):
  - **Categoría X**: documento que por sí mismo anticipa la invención a patentar y por lo tanto no se puede considerar nueva.
  - **Categoría Y**: en combinación con otro documento anticipa la invención y un experto podría deducirla.
  - **Categoría A**: documento que provee estado de la técnica y no es perjudicial para la demostración de novedad o actividad inventiva.

2.2. Herramientas de análisis: Análisis de enlaces

Actualmente, según los datos recogidos por la Unión Internacional de Telecomunicaciones (UIT) de las Naciones Unidas, aproximadamente el 51% de la población mundial (casi 8.000 millones de personas) se encuentran conectadas a Internet (International Telecommunication Union, 2020). A esa cantidad se puede

agregar las mostradas en la Figura 15 que recoge la cantidad de información que se envía usando la red de redes cada segundo[43].

Toda la información contenida en internet –incluyendo las patentes– se puede transformar en conocimiento al analizarla. La cibermetría aporta las técnicas y procedimientos necesarios para realizar un análisis, partiendo de fundamentos informétricos, utilizando técnicas tanto cuantitativas como cualitativas, desde una perspectiva procedente de las ciencias sociales. Aplicando técnicas cibermétricas a la información contenida en internet se puede entender, conocer y predecir el uso y comportamiento de la información y sus usuarios

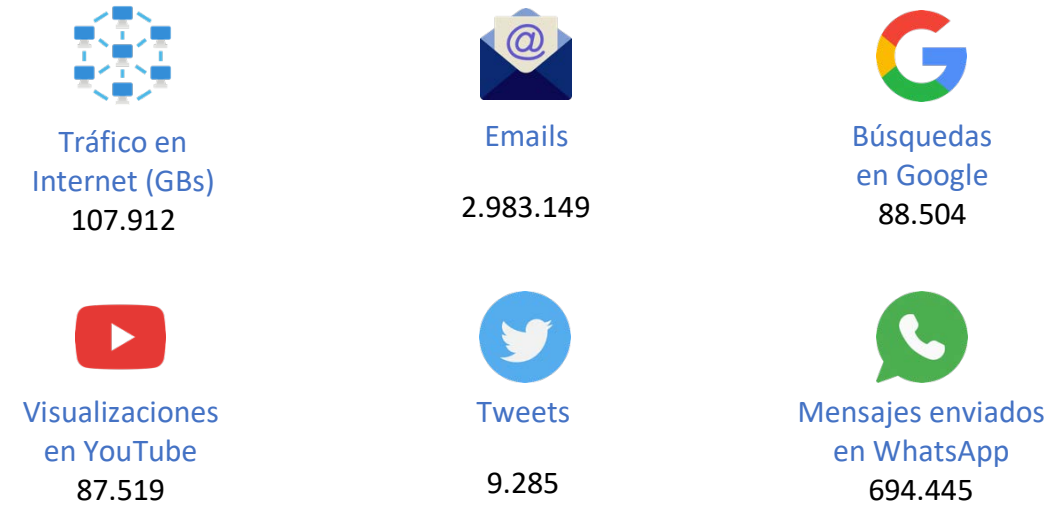


Figura 15: Cantidad de información enviada mediante Internet en un segundo
Fuente: datos: InternetLiveStats, visualización propia

Para ello, en la presente sección del capitulo se explica la disciplina, sus características y herramientas, buscando su aplicabilidad al análisis de las patentes, especialmente mediante el uso del análisis de enlaces, una técnica que permite a varios niveles la extracción de conocimiento desde la web y en especial sobre los recursos más relevantes.

2.2.1. Cibermetría

La velocidad a la que cambia el mundo es cada vez mayor; Gordon More en 1965 (rectificada en 1975) publicó la que actualmente sería conocida como la ley de Moore (Moore, 1965)(Orduña-Malea et al., 2016) en ésta se formula que cada dos años la capacidad de los transistores se duplicaría. Esta ley fue más tarde expandida por Kurzweil –denominada Ley de rendimientos acelerados (Kurzweil, 2004)– y con ella se

[43] https://www.internetlivestats.com/

describe el crecimiento exponencial pudiéndose aplicar también al progreso en almacenamiento, tecnológico, redes de conexión e incluso al progreso social y cultural.

Esto es lo que sucede con la expansión de la World Wide Web, su crecimiento desde que Tim Berners-Lee la crea en 1989 se puede equiparar al *Big Bang* que origina el universo:

- → 1989: Se inicia el desarrollo del protocolo de red para sistemas de información HTTP
- → 1991: Aparece HTML como lenguaje de marcado para la comunicación en la Web
  - ⇢ Arranca Gopher, un protocolo comunicación para la búsqueda, distribución y recuperación de documentos en la web
- → 1992: Se publican ViolaWWW y Mosaic, los primeros navegadores web gráficos
- → 1993: Aparece World Wide Web Wanderer, el primer robot para generar un índice de la web llamado Wandex.
  - ⇢ Se publica el W3Catalog, considerado el primer motor de búsqueda
- → 1994: Se desarrolla la versión dos de HTML y aparecen NetScape y Opera
  - ⇢ Aparecen Yahoo! Directory, World-Wide Web Worm, WebCrawler y Lycos
- → 1995: Se lanza Internet Explorer y aparecer HTMLv3
  - ⇢ Surgen Yahoo! Search, LookSmart y Altavista
- → 1996: Aparecen JavaScript, Java y Flash, lenguajes para dinamizar el contenido web
- → 1997: Aparecen el lenguaje XML y Google Search
- → 1998: nueva versión HTML5 y CSS2. MSN lanza MSN Search
- → 1999: Se lanzan Ajax y AlltheWeb y MSN Messenger
- → 2000: Aparece Baidu
- → 2002: Yahoo! Compra Inktomi y Overture Services Inc., haciéndose así con AlltheWeb y Altavista y combinando todas las tecnologías en Yahoo! Search.
- → 2003: Aparecen Safari y MySpace
- → 2004: Se lanza Firefox y nacen Facebook y Flickr
- → 2005: Lanzamiento de Bebo, Qzone y Reddit
- → 2006: Aparece Twitter, VK y Tumblr
- → 2008: Boom

Desde este año se produce la gran expansión a nivel mundial de la red. El ADSL permite conexiones más potentes, se propaga el uso de terminales móviles con acceso a internet con el paso de la conexión GPRS al 3G, aparece la web 2.0 con mejores sistemas de contenido y búsqueda, y comienzan a usarse las redes sociales. Se pasa de un tráfico mensual de menos de 10.000 Petabytes al mes en 2008, a casi 280.000 Pb/mes en 2018[44]. En 2008 existían 172.000.000 páginas web[45], en el momento de escribir estas líneas (diciembre 2020), el número es de 1.826.000.000, lo que supone un incremento

[44] https://www.cisco.com/c/en/us/solutions/executive-perspectives/annual-internet-report/index.html

[45] https://www.internetlivestats.com/total-number-of-websites/

del 1.061%, y que concuerda con la evolución de la Ley de rendimientos acelerados mencionada anteriormente.

Paralelamente al crecimiento de la Web se daba el nacimiento y evolución de la Cibermetría como disciplina. Como su propio nombre indica, se basa en la aplicación de técnicas *métricas* al *ciber*espacio o entorno web. Según la definición propuesta por y Orduña-Malea y Aguillo (2014), la cibermetría se define como:

"El estudio y caracterización del espacio red a partir del análisis de sus elementos constitutivos (especialmente en los aspectos relacionados con su creación, estructura, topología, difusión, interrelaciones, evolución, consumo e impacto) mediante técnicas cuantitativas de investigación social."

La cibermetría aplica los principios de la *Informetría*, *Bibliometría* y *Cienciometría* al estudio de la Web, siendo el enfoque común en todas ellas el estudio cuantitativo de la información académica en sentido amplio, para conocer los procesos de creación, difusión y consumo de contenidos científicos, más allá de un artículo de revista impreso (Björneborn, 2004). Se genera así un ecosistema de indicadores y métricas que permite entender el universo *online,* y su proyección en el mundo *offline,* a diferentes niveles. Para comprender mejor la interrelación existente en las diferentes disciplinas, así como la profundidad y necesidad de la Cibermetría, Björneborn realiza una representación gráfica (que además incluye la *Webmetría,* y Stefanie Haustein actualiza para incluir la *Altmetría* (Haustein, 2015)) que se puede observar en la Figura 16.

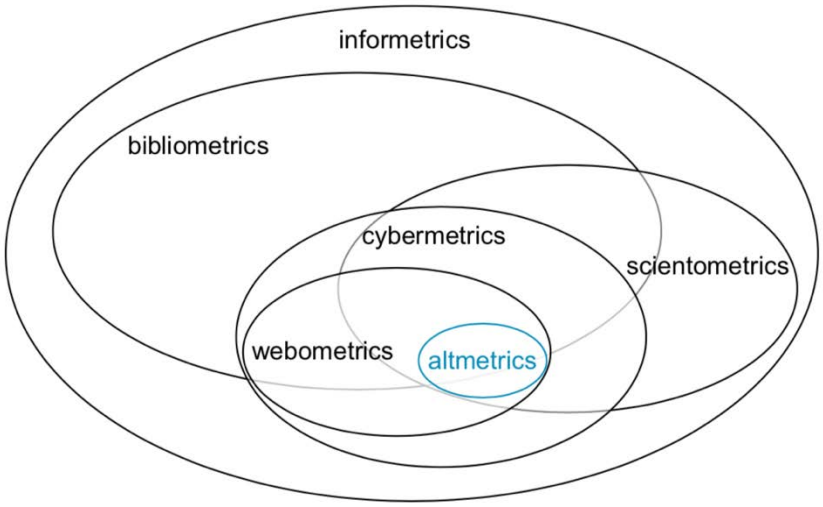


Figura 16: Relación entre disciplinas (Fuente: Haustein, adaptado de Björneborn, 2004)

Cada una de las áreas, pese a solaparse entre ellas, permiten el estudio en profundidad de diferentes aspectos, dado que se considera importante para una mejor comprensión de la Cibermetría, a continuación se define brevemente cada una de las disciplinas:

→ **Informetría**: subdisciplina de las ciencias de la información, realiza el estudio cuantitativo de la producción, almacenamiento, recuperación, diseminación y uso de la información, independientemente de su forma y origen. Busca la

comprensión de los procesos de información mediante el desarrollo de modelos matemáticos y teorías (Wolfram, 2000).

- → **Bibliometría**: realiza el estudio o medición de los diferentes formatos de comunicación (libros, artículos, textos, documentos o información). Para el análisis se utilizan y relacionan los datos de autores, publicaciones científicas, citas y lecturas, de forma que permita calcular el impacto y visibilidad de la entidad estudiada (autor, organización, publicación, grupo de investigación, países, etc.) (Alonso Arroyo, 2004).
- → **Cienciometría**: cuantifica y analiza mediante técnicas fundamentalmente cuantitativas las comunicaciones realizadas en ciencia, tecnología e innovación teniendo como centro del análisis los documentos publicados (Leydesdorff, L. & Milojevic, 2012).
- → **Webmetría**: estudio de los aspectos cuantitativos de la Web, teniendo como objeto de estudio el número y tipo de enlaces, la estructura de la web y su uso, buscando entender el uso y construcción de los recursos de información en la web, sus estructuras y tecnologías siendo estudiados mediante una aproximación bibliométrica e informétrica (Björneborn, 2004).
- → **Altmetría**: tras la aparición de la Web 2.0 aparece el estudio de la creación, diseminación e impacto de información y recursos mediante el uso de redes sociales como forma alternativa (y agregada) a la medición tradicional de impacto mediante citas (Priem, Taraborelli, Groth, & Neylon, 2011; Torres-Salinas, Cabezas-Clavijo, & Jiménez-Contreras, 2013).

Queda patente que, pese a que todas las disciplinas aplican estudios cuantitativos a la información, cada una se centra en una rama y logra un alcance diferente, es por esto que todas las métricas derivadas de estos estudios se encuentran agrupadas bajo el paraguas de las *iMetrics* (Maltseva & Batagelj, 2020; Milojević & Leydesdorff, 2013).

Del mismo modo que para la tecnología existe la Ley de rendimientos acelerados, en cienciometría se utiliza la Ley del crecimiento exponencial de Price (de Solla Price, 1976), que explica el aumento del uso de las *iMetrics*, en 2010 la cantidad de artículos publicados cuadriplicaba la cifra del año 2000 (Milojević & Leydesdorff, 2013).

Como se puede observar en la Figura 17, la relación existente entre disciplinas es alta, ya que todas las co-citas de palabras clave recogidas en los documentos relativos a *iMetrics* publicados entre 1978 y 2014 (5.944 artículos analizados) se encuentran muy agrupadas e interrelacionadas (Khasseh, Soheili, Moghaddam, & Chelak, 2017). Dentro de este análisis se puede observar como el término “patente” se encuentra en la posición 14 (de 30) con una frecuencia de 266 apariciones, lo que denota interés por la aplicación de las *iMetrics* al mundo de las patentes.

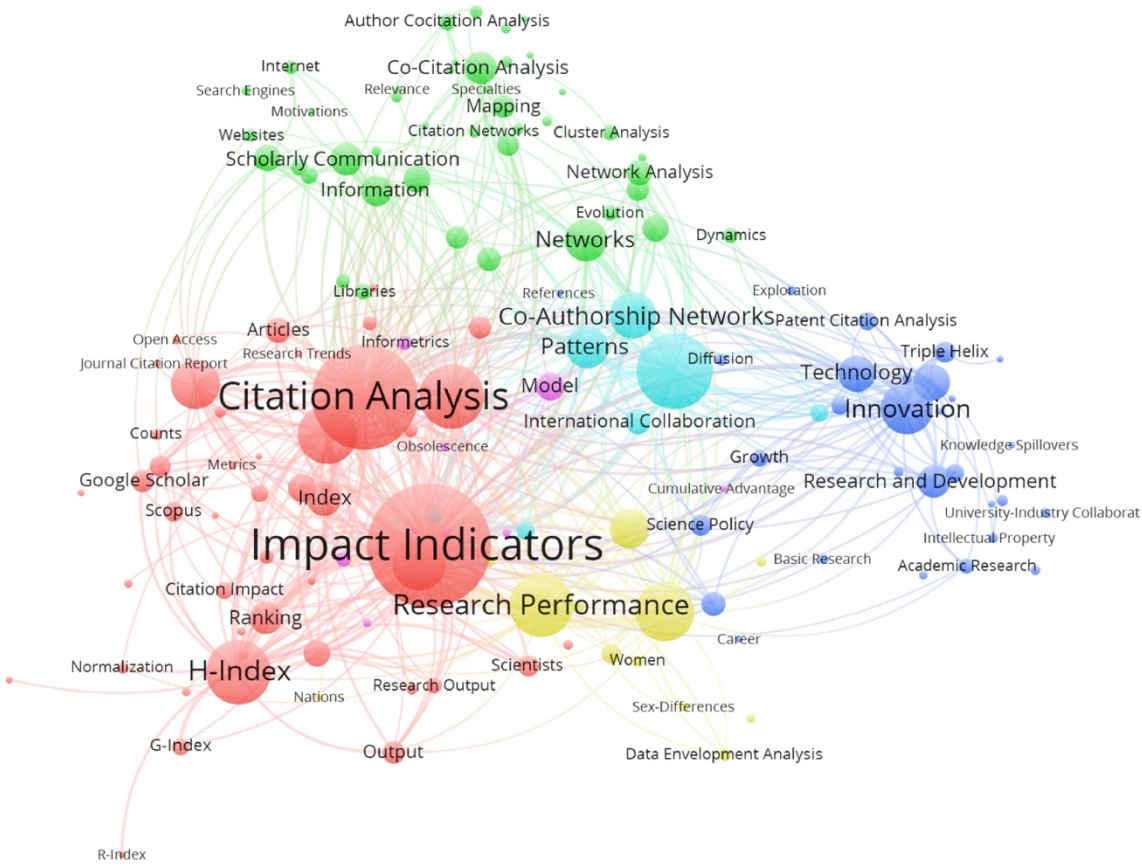


Figura 17: Red de las 155 palabras clave más citadas en iMetrics entre 1978 y 2014 (Khasseh et al., 2017)

Con todas las definiciones y aplicaciones vistas, se debe centrar el foco en la cibermetría para poder entender mejor su ámbito y aplicación. Teniendo en cuenta lo visto hasta ahora, y considerando el espacio red como eje central y los contenidos como unidad de análisis, la cibermetría se puede dividir en tres áreas de trabajo (Orduña-Malea, E.; Aguillo, 2014):

- → **Descriptiva**: se centra en estudiar la propia disciplina, analizando su definición y modelización mediante indicadores, su naturaleza y las unidades de medida.
- → **Instrumental**: se centra en el estudio de las fuentes de información, analizando su funcionamiento, cobertura y limitaciones, así como métodos de extracción, análisis y visualización de la información contenida.
- → **Aplicada**: se centra en la realización de análisis focalizados en un objeto de estudio y su entorno.

Según Orduña-Malea y Aguillo, las tres áreas se encuentran interrelacionadas como muestra la Figura 18, de forma que la cibermetría instrumental aporta precisión al área descriptiva y la aplicada aporta contexto e interpretación. Las áreas instrumental y aplicada se relacionan creando un marco de estudio para el objeto analizado.

Como se ha visto hasta ahora se trata de una disciplina compleja, por ello lo siguientes apartados detallan las partes más importantes para comprender su aplicación en la presente tesis.

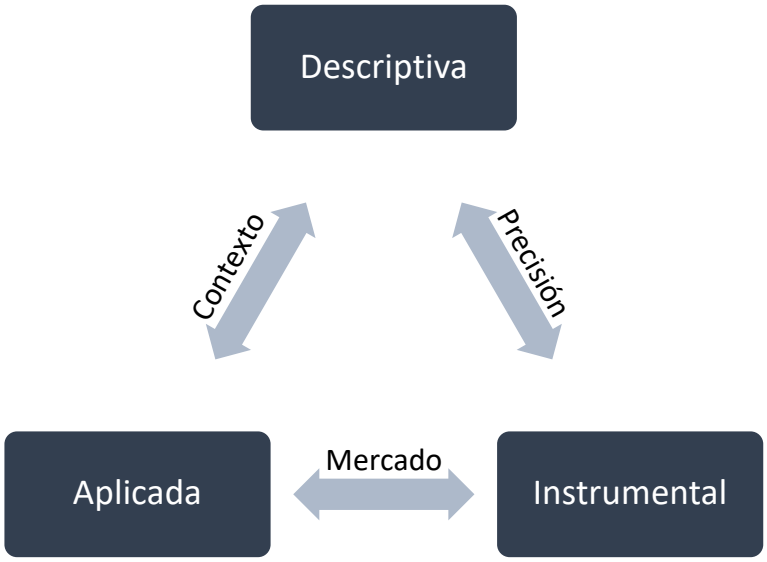


Figura 18: Interrelaciones entre las diferentes áreas de trabajo (Orduña-Malea, E.; Aguillo, 2014)

2.2.1.1. Ámbito de aplicación

Se ha indicado anteriormente que la Cibermetría busca estudiar y entender todos los componentes internos del espacio red. Para lograr comprender exactamente cómo se estudia, primero debemos saber qué se estudia y dónde están sus límites.

Primero se debe abordar la diferencia en la terminología, generalmente se utilizan los términos ciberespacio, internet y web, de forma sinónima, pero existen matices y diferencias entre todas ellas.

- **Internet**: es la abreviación de interconnected network, un conglomerado de redes descentralizadas a nivel mundial interconectadas mediante protocolos estandarizados de comunicación. Cuando se hace mención a Internet (con mayúscula) se hace referencia a la red de redes[46].
- **Ciberespacio**: es el entorno virtual que se genera gracias a la existencia de Internet.
- **Web**: abreviación de World Wide Web (o www) es la colección de información y documentos (paginas y portales web, ficheros, archivos, etc.) que se encuentran en internet.

De este modo se puede sobreentender que Internet es la infraestructura que contiene los documentos que conforman la web y que dan lugar al ciberespacio. Isidro F. Aguillo (Aguillo, 2009) publica un esquema conceptual que representa la diferencia entre las tres áreas (Figura 19).

[46] Internet con minúscula hace referencia a una única red.

<table>
<tr><td colspan="6">CIBERESPACIO</td></tr>
<tr><td rowspan="6">INTERNET</td><td rowspan="5">CONTENIDOS</td><td rowspan="5">INTERNET PÚBLICO</td><td colspan="3">Correo, foros públicos</td></tr>
<tr><td colspan="3">Sistemas de intercambio P2P</td></tr>
<tr><td rowspan="2">webespacio</td><td colspan="2">web visible</td></tr>
<tr><td>web invisible</td><td rowspan="2">internet invisible</td></tr>
<tr><td colspan="2">Infranet</td></tr>
<tr><td colspan="3">Datos sobre consumo</td></tr>
<tr><td colspan="4">Intranets</td></tr>
<tr><td>INTERNET FÍSICA</td><td colspan="4">Estructura, topología, tráfico, demografía, geografía</td></tr>
<tr><td colspan="6">CONTENIDOS ELECTRÓNICOS FUERA DE INTERNET</td></tr>
</table>

Figura 19: Representación de Internet de contenidos e Internet físico.
Fuente: (Aguillo, 2009)

La cibermetría permite medir y analizar todo aquello que se encuentra representado en la Figura 19 (no sin ciertas limitaciones y problemas que serán descritos a lo largo del capitulo) siendo las áreas de mayor relevancia el internet físico y el espacio web. Como se puede apreciar en la Figura 19, dentro del espacio web existen dos subconjuntos de espacios: la web visible y la invisible.

Se entiende por web invisible la conformada por todos aquellos documentos que no son accesibles, bien sea por el desconocimiento de su ubicación o porque debido al tipo o naturaleza de su contenido (i.e.: páginas dinámicas). Del mismo modo que para que los ordenadores que se encuentran en la red se comuniquen es necesario el protocolo TCP/IP, que les permite localizarse entre ellos mediante identificadores únicos (IP), los documentos que se encuentran en la web deben tener su propio número de serie. De este modo, cuando se desee localizar un fichero (independientemente de si se trata de una página web o de un documento pdf) y se introduzca su nombre en el navegador, aparecerá éste y no otro (Kleinrock, 2010; Leiner, B. M. et al, 2009).

En la prehistoria de la web, cuando los indizadores o buscadores de documentos no existían, para poder acceder a un fichero era necesario conocer exactamente dónde se encontraba alojado. Con la expansión del universo se hizo totalmente necesario el uso de herramientas para poder acceder a la información. De este modo, aparecieron los motores de búsqueda (Seymour, Frantsvog, & Kumar, 2011) que permiten mediante la búsqueda de palabras clave acceder a la información.

Debido a la naturaleza propia de algunos documentos (volumen de datos extremadamente grande, información fugaz, pasarelas o contraseñas, etc.) o simplemente porque los motores de búsqueda estándar no dejan de ser bases de datos que almacenan la localización (el identificador único del documento) dónde éste se encuentra y su capacidad es limitada, estos no se encuentran indizados, por lo tanto la web invisible es un lugar de difícil acceso para el usuario común (Bojo Canales et al.,

2004). En 2010 se estimó que el tamaño de la Deep Web (término en inglés) era de 7.500 Terabytes (mientras que la zona visible era de 19 Tb) (Iffat & Sami, 2010).

En relación con el indicador único para cualquier documento indicado con anterioridad, del mismo modo que cada maquina conectada a la red recibe un nombre inequívoco y único con los documentos sucede lo mismo. Para que se de en este caso se le otorga el nombre de la máquina en la que se aloja (IP) seguido de la ruta de acceso y su nombre. Dado que los humanos no trabajamos bien comunicándonos únicamente mediante números, ésta se puede traducir por un nombre, dando así paso a la aparición de los dominios. Un nombre de dominio se conforma habitualmente por dos partes:

- → **TLD**: Top Level Domain, se encuentra tras el punto y se trata de una extensión (.com, .es, .gob, .edu,…)
- → **SLD**: Second Level Domain, el nombre del alojamiento (google.com, upv.es, w3c.org, ...)

Actualmente existen 1.584 TLDs, que se pueden dividir en los siguientes tipos:

- → **gTLD**: genéricos (.top, .loan, .club, .online, .tokio, .london, .nyc, …)
- → **sTLD**: patrocinados [agencias privadas u organizaciones] (.com, .gov, .mil, …)
- → **ccTLD**: geográficos (.es, .uk, .fr, .jp, .in, …)
- → **IDN**: codificación no latina (.닷컴, .クラウド, ...
- → **grRLD**: restringidos (.example, .local, .localhost, .test, …)

En la Tabla 6 se puede encontrar los 30 TLDs más utilizados y sus porcentajes de uso[47]:

Tabla 6: 30 TLDs más utilizados a nivel global 2020
Fuente: elaboración propia

TLD	%	TLD	%	TLD	%	TLD	%	TLD	%	TLD	%
.com	51,9	.in	1,6	.ca	1,0	.co	0,9	.id	0,5	.za	0,4
.ru	6,2	.au	1,6	.jp	0,9	.fr	0,8	.tw	0,5	.eu	0,4
.org	4,5	.uk	1,6	.tr	0,9	.info	0,6	.cn	0,5	.es	0,4
.net	3,3	.de	1,3	.vn	0,9	.pl	0,5	.nl	0,5	.mx	0,4
.ir	1,7	.ua	1,1	.br	0,9	.it	0,5	.io	0,4	.ch	0,4

En el §2.2.1 se ha visto como Internet, la Web y el Ciberespacio se encuentran constantemente evolucionando y expandiéndose. De nuevo en esta ocasión aplica la Ley de crecimiento exponencial, no importa si es al contenido o la estructura que lo sostiene. Esto representa un problema, ya que su dinamismo y constante evolución implican que las mediciones que se realizan en un momento pueden no coincidir con la misma solicitud realizada un segundo más tarde. Por lo tanto, cualquier estudio relativo a las *iMetrics* debe entenderse como una fotografía de un momento concreto. Así mismo, se debe tener en cuenta que no se puede realizar un estudio completo sobre todo el contenido del ciberespacio y que únicamente se trabaja con el contenido que se encuentra en la web visible.

[47] https://w3techs.com/technologies/overview/top_level_domain

2.2.1.2. Modelos

Existen diversas formas de dar forma o agrupar al contenido que se encuentra en la web para ser medido. A continuación, se detallan diversos modelos categorizados en función del tipo de medición o caracterización cibermétrica que se desee realizar(Orduña-Malea, E.; Aguillo, 2014):

- → **Conceptual**: trata de establecer los límites del espacio red y sus elementos, así como su visualización, a partir de su definición teórica
- → **Cuantitativa**: busca cuantificar el tamaño total del espacio web mediante la cuantificación disgregada de cada uno de sus elementos
- → **Evolutiva**: trata de analizar la estabilidad, persistencia, permanencia y desaparición de los elementos contenidos en el espacio web
- → **Topológica**: trata de estudiar la topología del espacio web mediante la cuantificación de infraestructuras, servicios, contenidos y usuarios.

La Tabla 7 recoge un resumen de los diferentes componentes y sus unidades de medición para cada una de las diferentes aproximaciones que se pueden realizar en la medición del ciberespacio. Conociendo las diferentes aproximaciones que existen para la medición de los elementos que componen la web, es necesario indicar las diferentes métricas e indicadores que se pueden aplicar. Estos indicadores se encuentran agrupados según la naturaleza e intencionalidad de cada uno de ellos (Orduña-Malea, E.; Alonso-Arroyo, 2017).

Tabla 7: Tabla resumen de caracterización según tipo de medición (Orduña-Malea, E.; Aguillo, 2014)

Aproximación	Qué mide	Unidades
Cuantitativa	Estructuras físicas	Servidores, ordenadores, dispositivos
	Estructuras Lógicas	Dominios web, Hosts, sitios web
	Estructuras de Comunicación	Cables, redes, disponibilidad, puntos de acceso, tráfico, velocidad de conexión
	Servicios	Navegadores, aplicaciones
	Contenidos	Ficheros, metadatos, información
	Usuarios	Cantidad, parámetros (cultura, lenguaje, edad, …)
	Audiencia	Cantidad de usuarios, parámetros (procedencia, edad, dispositivo, …)
Evolutiva	Tipología	Dominio web
	Área temática	Contenido por disciplinas
	Naturaleza	Según la tipología de la página
	Edad	A mayor edad, menos modificaciones en las webs
	Tamaño	A mayor número de páginas en un sitio web o indexadas por motores de búsqueda, mayor cambio
Topológica	Infraestructura	Redes de ordenadores
	Contenidos	Redes de páginas web
	Usuarios	Redes de usuarios
	Etiquetas	Redes de tags

En relación con los indicadores web aplicados a patentes específicamente, dado que el análisis de la patente se hace a título de documento aplicarían los mismos indicadores que se muestran en la Tabla 8, con especial relevancia aquellos relacionados con visibilidad y consumo.

Tabla 8: Resumen clasificación de métricas e indicadores online
Fuente: Basada en (Orduña-Malea; Alonso-Arroyo, 2017)

Naturaleza	Tipo de métrica	Tipo de indicador	Unidad	Ejemplos de Métricas
Impacto	Visibilidad	Mención	Enlaces	Número de enlaces entrantes y salientes (directos o indirectos), número de menciones textuales, palabras clave
	Consumo	Uso	Visitas	Número de accesos web, clicks, visionados, descargas
	Satisfacción	Opinión	Calificaciones	Likes, marcadores, reviews, etc.
	Difusión	Alcance	Seguidores	Número de seguidores, visualizaciones, contestaciones, reTweets, recomendaciones
	Interés	Atención	Comentarios	Número de comentarios, quoted reTweets, réplicas, etc.
	Autoridad	Optimización	Clasificación	Posicionamiento en buscadores horizontales y verticales
	Conectividad	Topología	Densidad	Análisis de redes (métricas a nivel de nodo y a nivel de red)
Presencia	Productividad	Tamaño	URLs indexadas	Peso de ficheros, número de ficheros, número de URLs, dominio, subdominio, subdirectorio, tipo de fichero, formato o naturaleza
Tiempo	Antigüedad	Edad	Días	Tiempo de exposición
Forma	Diseño	Calidad	Usabilidad	Core Web Vitals (LCP, FID y CLS), enlaces rotos, etc.
Combinados	Multidimensional	Multidimensional	Multidimensional	Web Impact Factor

2.2.1.3. Metodologías

Los métodos de recogida, análisis y visualización de la información dependen del tipo de información, fuente y herramienta que se esté utilizando y en función del resultado que se desee obtener. Debido a que no está en el alcance de la presente tesis realizar una explicación pormenorizada de todos los métodos, la T recoge un esquema de metodología propuesto por (Orduña-Malea, E.; Aguillo, 2014), posteriormente se explicará aquellos métodos que sí tienen aplicación en el desarrollo de la tesis.

Tabla 9: Métodos de medida, análisis y visualización
Fuente: basada en Orduña-Malea; Aguillo, 2014

Pasos	**Métodos**		
Selección del tipo de fuente	Motores de búsqueda	Plataformas específicas	Ficheros de transacciones
Seleccionar fuente	Google, Bing, DuckDuckGo, Baidu, etc.	Facebook, Twitter, Mendeley, etc.	Digital Analytics, Alexa, Similar Web, etc.
Consulta Fuente	Motores/Plataformas		Ficheros Weblogs
	Directo (comandos) Indirecto (consulta API)		Indirecto (aplicaciones) Indirecto (paneles)
Captura de datos	→ Hit count estimates → Query splitting → API → Web scraping → Descarga directa		→ Análisis descriptivo → Análisis de sesiones → Análisis de uso
Analizar datos	Estadística descriptiva; métodos de inferencia, etc.		
Visualizar datos	Diagramas (redes simples y geográficas; noded-positioned), Rankings, DataViz		

Para la presente tesis resulta relevante la explicación del funcionamiento de los motores de búsqueda y plataformas específicas.

2.2.1.3.1. Motores de búsqueda

Como se explica en el §2.2.1, los motores de búsqueda marcaron un antes y un después en el acceso a la información, ya que gracias a ellos se democratiza la búsqueda de información. Las webs y ficheros pasan a encontrarse a unos clicks de distancia y eso facilita y aumenta el nivel de penetración de la web en la sociedad (Seymour et al., 2011).

Su funcionamiento es como una gran base de datos, en la que se almacena clasificada la información recopilada y de la cual se extrae tras una consulta del usuario (una búsqueda) la información. Un motor de búsqueda divide su funcionamiento en tres pasos:

- → **Rastreo**: mediante el uso de bots o arañas, los sistemas buscan información en la red para nutrir la base de datos. Las arañas son programas informáticos que van saltando de enlace a enlace y leyendo los documentos en los que aterrizan.
- → **Indexación**: el siguiente paso, una vez localizado el fichero, es categorizarlo y almacenarlo en la base de datos. La categorización es el punto crítico, ya que una correcta permitirá acceder más fácilmente a la información correcta cuando se realice una búsqueda. La categorización se realiza mediante palabras clave detectadas en el cuerpo del documento, sus metadatos, descripciones, etc.
- → **Búsqueda**: activada por el usuario, desde la página de acceso del motor de búsqueda, escribe la consulta y activa los algoritmos de búsqueda que analizarán el contenido y buscarán en la base de datos los documentos relevantes para ella.

Para lograr una experiencia de búsqueda optima y ofrecer siempre los mejores resultados, los algoritmos trabajan en perfeccionar las respuestas en función de la interacción del usuario con los resultados. Por este motivo, los algoritmos de búsqueda son el secreto mejor guardado de las compañías, ya que de la calidad de los resultados depende el uso del motor.

Existen tres tipos de motores de búsqueda[48]:

- → **Jerárquicos**: los más similares a la descripción anterior, ofrecen una interfaz amigable desde la que consultar a una base de datos. Actualmente lo más famosos son Google o Bing.
- → **Directorios**: se trata de directorios de enlaces que apuntan a páginas con motores de búsqueda concretos. Open Directory Project es un ejemplo.
- → **Metabuscadores**: son replicadores de búsquedas, el usuario inserta su duda en una interfaz y el metabuscador la replica en múltiples buscadores. Por ejemplo: DogPile o Metacrawler.

En España, el 96,47% de los usuarios utiliza Google como motor de búsqueda, el porcentaje restante se encuentra repartido entre Bing (2,07%), Yahoo! (099%), DuckDuclGo (0,22%), Ecosia (0,11%) y otros (0,13%)[49]. Si se toman las cifras mundiales, pese a disminuir ligeramente, Google mantiene el 91,38% de cuota de mercado, por lo que este será el buscador que se utilice para la aplicación de las técnicas necesarias.

Los motores de búsqueda, como se señala con anterioridad, presentan una serie de limitaciones. Además, de los problemas de indización y alcance indicados, se presentan problemas de relativos a las políticas de empresa; al tratarse de compañías privadas,

[48] Actualmente Directorios y Metabuscadores se encuentran en desuso

[49] https://gs.statcounter.com/search-engine-market-share/all/spain/2020

pueden realizar las modificaciones en sus algoritmos que consideren pertinentes, modificando el sistema de indización y búsqueda posterior, e insertando el sesgo que pueda interesarles (E. S. Han & Goleman, D.; boyatzis, R.; Mckee, 2008; Mowshowitz & Kawaguchi, 2005).

Se debe tener encuentra siempre el factor humano, aunque sea de modo inherente y sin voluntad propia del equipo de desarrollo, los algoritmos se pueden encontrar limitados por esto.

Además, debido a la diversidad cultural, de lenguaje, forma de comunicación e incluso localización geográfica, se pueden dar resultados sesgados. La mayoría de buscadores intentan ofrecer al usuario respuestas más centradas en sus gustos o necesidades analizando su historial de búsqueda previo o las cookies[50], por lo que en la búsqueda de información totalmente personalizada se puede ofrecer resultados que contengan un sesgo (Vaughan & Thelwall, 2004).

Google y Google Patents

Google[51] es el buscador de la compañía Alphabet Inc., y nace en 1998 de la mano de Larry Page y Sergey Brin. Su principal atractivo es la sencillez de uso, únicamente contiene un cajetín en el que el usuario puede introducir su consulta y tras unos segundos recibirá los resultados de búsqueda.

Este motor se especializa en la atención al usuario, por este motivo cuenta con un ecosistema de herramientas y aplicaciones que se pueden utilizar para lograr una búsqueda más personalizada. Algunas de estas herramientas son el conversor de divisas, traductor, seguimiento de paquetes o información del tiempo. Una de las últimas funcionalidades añadidas permite obtener la respuesta a la pregunta realizada resaltada en amarillo en una web totalmente independiente. Además, en caso de tener una cuenta de usuario en Google y utilizar su navegador Chrome, al tener acceso a nuestros gustos e intereses, los resultados de la búsqueda serán perfilados mediante esta información.

El buscador, además de indexar los documentos, ofrece la posibilidad de realizar búsquedas mediante operadores a modo de búsqueda avanzada, de modo que un usuario con mayores conocimientos pueda tratar de localizar la información más ajustada a su búsqueda.

En el caso de Google Patents[52], se trata de un buscador especializado y centrado en la localización de documentos de patentes. Actualmente muestra patentes a texto completo de 17 oficinas de patentes y cuenta con más de 87 millones de documentos,

[50] Ficheros generados en los navegadores tras realizar visitas a páginas web que guardan la información de navegación buscando facilitar y mejorar la experiencia de navegación.

[51] https://www.google.com

[52] https://patents.google.com/

tanto solicitudes como concesiones, en su base de datos. La Tabla 10 recoge el número de documentos concedidos para cinco de las Oficinas que se encuentran en la base de datos.

El servicio fue lanzado en 2006 y permite acceder al contenido del documento tanto en su versión online como en el documento en formato .pdf. Ofrece la posibilidad de realizar búsquedas específicas, limitando los resultados por año, oficina, término o tipo de documento, entre otros.

Desde la compañía se busca ofrecer servicios añadidos que permitan mejorar la experiencia de usuarios. En el caso de Google Patents, aunque no existe una API que facilite la recuperación de información a gran escala, mediante el uso de Google Cloud Platform (un servicio de almacenamiento en la nube para grandes volúmenes de datos) se puede acceder desde 2019 a colecciones completas, para facilitar la búsqueda de documentos.

Tabla 10: Número de documentos concedidos en la base de datos de Google Patents

Oficina	Número documentos concedidos
USPTO	11.811.438
EPO	2.020.295
WIPO	4.664.943
Japón	5.823.918
España	901.622

Los resultados de Google Patents son, a su vez, accesibles entre los resultados al realizar una búsqueda en otros servicios de la compañía (Google motor de búsqueda o Google Scholar) en caso de ser relevantes para los resultados.

2.2.1.3.2. Plataformas específicas

Existen otras plataformas específicas ampliamente utilizadas en estudios cibermétricos (Orduña-malea, 2012), que ofrecen información sobre los recursos contenidos en la web. Se trata de herramientas de tipo *panel-based*, es decir, mediante un panel de gestión se puede acceder a los datos o información que contienen. Funcionan igual que los motores de búsqueda, mediante el uso de robots o arañas recorren la web para identificar y almacenar información relativa a los enlaces. A diferencia de los motores de búsqueda utilizan está información para generar métricas de los propios portales y ofrecerlas a los usuarios de las herramientas.

Algunas de las herramientas que ofrecen son relativas al tráfico generado por los portales analizados, las palabras clave más demandas, el posicionamiento de páginas web y conteo de enlaces (entre otras).

Existe una cantidad considerable de herramientas, pero las más importantes –y que además permiten el análisis de enlaces, tema central de la tesis, en profundidad– son

AHrefs[53], Moz Link Explorer[54] y Majestic[55]. A continuación se detalla brevemente los puntos diferenciadores de cada una de ellas:

- AHrefs: herramienta de pago con cuatro niveles de acceso (Lite, Estándar, Avanzado y Agencia). Se encuentra centrada en el análisis SEO y cuenta con cinco secciones principales: backlinks explorer, explorador de palabras clave, explorador de contenido, análisis de la competencia y alertas. Permite crear proyectos para controlar la evolución de los dominios analizados.
- Moz Link Explorer: herramienta de pago con cuatro niveles de acceso (Estándar, Medio, Grande, Premium). Permite controlar los enlaces entrantes, palabras clave, enlaces rotos, competencia, textos ancla y nichos. Se encuentra enfocada en la mejora del SEO de un portal web mediante estudios de auditoría.
- Majestic: es la única especializada en el análisis de enlaces. Cuenta con tres niveles de acceso (Lite, Pro y API). Ofrece una gran cantidad de métricas relacionadas con los enlaces de los dominios o subdominios que recopila. En todos los niveles de acceso permite la descarga de la información, tanto en datos brutos como en informes y gráficas. Sus indicadores más importantes son:
  - Citation Flow: indicador sobre 100 que puntúa el número de backlinks
  - Trust Flow: indicador sobre 100 que puntúa la calidad de los backlinks
  - Topical Trust Flow: que indica la importancia del dominio según la temática del portal web enlazante.

2.2.1.4. Técnicas

Existen diversas técnicas que pueden ser utilizadas para realizar análisis de tipo cibermétricos. A continuación se detalla algunas de ellas:

- Análisis de audiencia web: la medición de los usuarios, sus acciones en la web y su comportamiento puede ser recopilado para su posterior análisis mediante el uso de herramientas públicas (herramientas panel-based como Google Analytics, Alexa, etc.) o privadas (logs de registro en servidores).
- Análisis de sitios de redes sociales: las acciones de los usuarios en las redes sociales, las formas de expresión (mediante texto o imágenes) y redes generadas por los usuarios mediante la creación de contactos puede ser analizada desde las diferentes redes sociales.
- Search analytics: técnica que permite recopilar datos relativos al uso de palabras clave, tendencias, intención de búsqueda, resultados de búsqueda, etc. para entender mejor al usuario. Algunas herramientas que ofrecen este servicio son Google Trends, Moz o SEMRush.

[53] https://ahrefs.com/

[54] https://moz.com/link-explorer

[55] https://es.majestic.com/

- Web sentiment analysis: el análisis de sentimiento web permite monitorizar la conversación relacionada con una persona, palabra, negocio, etc. para comprobar el ánimo de los usuarios respecto de esta.
- Extracción de datos:
  - Web crawling: para recopilar documentos web se pueden programar robots o arañas ad-hoc que permitan recorrer la web (en general o un portal en particular) del mismo modo que lo hacen los motores de búsqueda para localizar los documentos buscados (o descubrirlos).
  - Web scraping: una vez localizados los documentos se puede extraer la información necesaria de éstos localizando aquella que sea útil y exportándola a otro fichero para poder analizarla. Esto se realiza mediante herramientas (públicas o propietarias) que recorren el documento, localizan, extraen y almacenan esa información.
  - APIs o Descarga directa: algunas fuentes ofrecen la posibilidad de descarga de datos directamente desde sus portales, bien mediante el uso de API o directamente desde el panel de navegación.
- Link Analysis: es una técnica de análisis de datos utilizada para evaluar las relaciones generadas mediante el uso de enlaces dentro de la web. Esta técnica será la utilizada en la presente tesis, por lo que se detalla en profundidad en el siguiente apartado.

2.2.2. Link Analysis

En análisis de enlaces se centra en la explotación de la información que pueden ofrecer estos marcadores. Se trata de una técnica ampliamente utilizada en multitud de campos, desde la informática hasta la sociología.

Desde un punto de vista informático, un hiperenlace[56] es un elemento que se encuentra en un documento electrónico y que hace referencia a otro recurso, ofreciendo una puerta de acceso al mismo. Equiparándolo al mundo real, un enlace sería número de teléfono y a su vez línea de comunicación.

Los enlaces forman parte de intrínseca de internet, ya que sin ellos la localización de cualquier tipo de documento sería imposible sin conocer previamente su dirección tal y como se explica en el §2.2.1.1. Sin enlaces, el universo se encontraría compuesto simplemente por rocas (documentos) errando por la nada de la web, ellos son los que generan los tirones gravitacionales necesarios para generar sistemas solares (sitios web), formar galaxias (clústeres de organizaciones, redes sociales, etc.) y dan sentido al universo en sí mismo.

Es cierto que los motores de búsqueda ayudan a localizar los ficheros sin conocer su localización, pero para que una araña sea capaz de encontrar un documento, guardar su URL y almacenarla categorizándola, es necesario localizar esa URL y para ello los bots necesitan los enlaces para poder saltar de un documento a otro.

[56] https://es.wikipedia.org/wiki/Hiperenlace

En la web, los hiperenlaces se encuentran generados –habitualmente– mediante el lenguaje de programación HTML (en caso de que se utilicen dentro de ficheros como Word o pdf, el sistema de marcado es diferente). Para generar un enlace se introduce entre las etiquetas de marcado de vinculo la referencia que se desea realizar junto con una descripción:

cadena de caracteres

Siendo:

- **<a>**: las etiquetas del lenguaje que indican la existencia de un vínculo
- **HREF**: es la abreviatura de *Hipertext Reference*, indica que ese campo equivale al enlace
- **URL**: *Uniform Resource Locator*, se trata de la dirección (el indicador único) del elemento que se quiere vincular
- **Cadena de caracteres**: puede tratarse de una palabra, texto o imagen explicativo de hacia dónde va el enlace

Además, dentro de la etiqueta de apertura <a> se pueden incluir otras características como *class*, *id*, *title* o *target* para ofrecer más información, cambiar el estilo o realizar acciones.

Los enlaces se pueden categorizar según el motivo por el que se han incluido en el documento o la funcionalidad que aportan (Orduña-Malea, E.; Aguillo, 2014) recopilan la siguiente caracterización:

- **Finalidad**:
  - **Organizativos**: permiten organizar un sitio web ("siguiente página", "subir")
  - **Por contenido**: se encuentran relacionados con la información indicada ("más información")
- **Destino**:
  - **Intrínsecos**: relacionan diferentes apartados de un sitio web. Por ejemplo, desde el menú superior a una sección interna de la página
  - **Internos**: señalan documentos (ficheros o páginas) que se encuentran en el dominio web
  - **Externos**: señalan documentos (ficheros o páginas) fuera del dominio web en el que se encuentran
  - **Rotos**: *404 Not found* indica que el elemento que antes se encontraba en ese enlace ha dejado de existir, faltan argumentos en el enlace o ha cambiado de dirección.
- **Rol**:
  - **Entrantes**: se realizan desde una fuente a otra, siendo el objeto de estudio el receptor
  - **Salientes**: se realizan desde una fuente a otra, siendo el objeto de estudio la primera
  - **Autoenlaces**: enlaces donde fuente y destino coinciden

- **Naturaleza**:
  - **Selectivos**: se estudian en función de un dominio
  - **Ponderados**: dependiendo de la fuente el enlace adquiere un valor u otro
  - **Generales**: aquellos que no son selectivos ni ponderados

Gracias a estas definiciones es posible entender mejor los propios enlaces, pero para poder estudiarlos es necesario conocer las áreas de trabajo y aplicación que tienen. Dentro del análisis de enlaces, existen dos áreas que permiten estudiar y entender, así como aplicar a la búsqueda de resultados el análisis de enlaces, descritos a continuación (Thelwall, 2009):

Evaluación del impacto de los enlaces

En bibliometría se asume que la cantidad de veces que se cita un artículo es un indicador de su valor o impacto. Aplicando este concepto al análisis de enlaces, mayor impacto o valor tendrá un documento cuantas más veces sea enlazado. Por lo tanto, está técnica busca conocer el número de enlaces que se encuentran dirigidos a un documento web mediante la recolección de datos en el ciberespacio (Orduña-Malea, E.; Aguillo, 2014).

Para recopilar los datos necesarios, se puede usar, por ejemplo, Google como motor de búsqueda. Pese a que las búsquedas con métodos avanzados se encuentran muy limitadas, se puede realizar una búsqueda similar a:

"www.url.com" –site:url.com

Con la instrucción indicada, el motor de búsqueda va a buscar en toda su base de datos aquellos documentos que no se encuentren en el dominio url.com pero que si contengan el texto exacto indicado entre comillas. De este modo, el motor ofrecerá un número de resultados que será una aproximación al número total de documentos que contienen el enlace y no son de tipo interno.

Estos resultados deben de ser tomados con cautela, ya que los motores de búsqueda priorizan velocidad de cálculo frente a calidad de resultados, por lo que pueden salir números más altos de lo normal (Font-Julian, Ontalba-Ruipérez, & Orduña-Malea, 2018). Existen, además, otras fuentes de información, como se indica en el §2.2.1.4 que pueden ofrecer estos datos

Este tipo de análisis se puede realizar a diferentes niveles:

- **Dominio**: url.com
- **Subdominio**: www.url.com o sub.url.com
- **Subdominio y directorio**: sub.url.com/categoria
- **Página individual**: url.com/categoría/explicación.html

Los resultados que ofrecerá el análisis pueden ayudar –entre otros– a:

- Comprender mejor el valor e impacto de los documentos analizados
- Localizar los recursos mejor valorados (y peor) y realizar análisis de diferencias
- Identificar los tipos de sitios que más enlazan a determinado recurso
- Identificar los países enlazadores mediante un análisis de TLD

Mapas de relaciones de enlaces

Dicen que una imagen vale más que mil palabras, por ello los mapas de relaciones de enlaces son muy útiles para poder obtener una fotografía más amplia del entorno estudiado. Además, permiten localizar subáreas o subconjuntos interconectados, que de otro modo podrían pasar desapercibidos.

Las Figura 14 y Figura 17 son ejemplos de visualización de un mapa de enlaces. En estos mapas, normalmente los nodos representan las páginas, sitios web o recursos y las flechas o líneas que los unen la dirección y cantidad de enlaces realizados/recibidos. La parte más crítica de este sistema es lograr una correcta visualización, para que toda la información se encuentre disponible, visible y permita generar conocimiento. Thelwall (Thelwall, 2009) propone las siguientes técnicas:

- → **Diagrama de red simple**: los nodos se encuentran enlazados mediante flechas sencillas. Para facilitar la comprensión, los nodos muy interconectados deben mostrarse cercanos, mientras que los que tengan menos relación se deben separar. Se pueden realizar con herramientas de creación gráficas sencillas.
- → **Diagrama de nodo posicionado**: representan nodos y flechas como en el caso anterior, pero con la particularidad de que la localización de los nodos en el mapa debe realizarse mediante la aplicación de herramientas especializas. Para ello se puede utilizar Gephi o XLGraph.
- → **Diagrama geográfico**: la interconexión entre nodos se realiza como en los anteriores diagramas, pero en este caso los nodos se sobre-posicionan en un mapa, mostrando en nodo en la localización a la que pertenece.

La representación de los nodos y enlaces puede aportar, además, información respecto a los nodos y sus conexiones. Cuánto más grande sea el nodo, más enlaces recibe, y cuanto más marcada, pintada o gruesa sea la línea de enlace, más menciones existirán.

2.2.2.1. Uso de los enlaces

Los enlaces se encuentran diseñados para permitir la navegación y el acceso de documentos en la web, pero esto no siempre implica que su uso sea el correcto o adecuado, o que tenga una importancia transcendental, aunque en un principio se considere implícita. La propia naturaleza del enlace es similar a las citas de los artículos científicos, otorgan o dan crédito al elemento que señalan. Por lo tanto, su uso debería ser realizado con mesura y cautela.

Sin embargo, esto no siempre es así ya que en multitud de ocasiones el uso de los enlaces se realiza de modo incorrecto o no apunta a la información relevante correspondiente.

Los enlaces permiten aumentar la visibilidad de los documentos que se encuentran en la web, además, su utilización puede ayudar al descubrimiento de información.

Permiten obtener credibilidad, del mismo modo que en la documentación científica se cita para referenciar a las autoridades, los hiperenlaces demuestran que existen fuentes fiables que respaldan la información que se muestra.

Facilitan la usabilidad, poniendo al alcance de un click la información que se desea compartir.

Pese a que el uso de enlaces es fundamental para la divulgación y visibilidad de los documentos, independientemente de su tipología y uso, los motivos por los que son utilizados no son fácilmente extrapolables mediante su análisis.

2.2.3. Cibermetría en patentes

La cibermetría aplicada al uso de las patentes, aunque se ha visto en el §2.2.1 que, existiendo producción científica y aplicabilidad, se cuenta con un amplio margen de maniobra para desarrollar ciertas técnicas en el área ya que los estudios se centran sobre todo en el análisis de citas entre patentes o a trabajos científicos, pero ignorando de algún modo los enlaces (Barberá-Tomás, Jiménez-Sáez, & Castelló-Molina, 2011; Criscuolo & Verspagen, 2008; Glänzel, Moed, Schmoch, & Thelwall, 2019; Iversen, 2000; Kang & Sohn, 2016; Karvonen & Kässi, 2013; Sanghoon Lee & Kim, 2017; Meyer, 2000a; Michel & Bettels, 2001; Sarin et al., 2020; Sterzi, 2013; Trajtenberg, 2006; Yang et al., 2015).

Debido a la complejidad existente hasta la fecha para trabajar con documentos de patentes, la técnica de Análisis de enlaces aplicada a la cibermetría no se encuentra explotada completamente ya que únicamente se encuentra el trabajo desarrollado por (Orduña-Malea et al., 2016). En dicho trabajo se realiza la primera aproximación al análisis de enlaces en patentes. Los resultados muestran que existen muchos enlaces y diversos recursos sobresalen (Wikipedia, YouTube, Archive.org). No obstante, el trabajo encuentra distintas limitaciones: únicamente se aplica a un conjunto muy limitado de patentes, de las cuales se extrae y analiza únicamente aquellos enlaces relativos a un conjunto limitado de universidades. Además, se utiliza Google Patents como fuente de información desde la que extraer los enlaces, lo que implica que pueden existir patentes ocultas ya que el buscador puede no mostrar todos los resultados que concuerden con la búsqueda realizada. Por lo tanto, pese a que el trabajo busca realizar un análisis similar, el enfoque y alcance es un mucho menor a la proyección de esta tesis, quedando demostrada la necesidad de establecer una metodología para realizar estos análisis de forma masiva y rigurosa.

La presente tesis busca, gracias a la existencia de nuevas fuentes de información sobre patentes a texto completo, herramientas de análisis y procesamiento, arrojar un poco de luz en la posibilidad de utilizar el análisis de enlaces para entender las formas de comunicación en los documentos de patente y la búsqueda de recursos de información valiosos.

Capítulo 3
Metodología

En este capítulo se detalla la metodología seguida durante todo el proceso de desarrollo de la tesis, incluyendo los experimentos y pruebas fallidas ya que aportan información relevante para comprender y mejorar los procesos. Además, se incluye un esquema resumen del modelo que permite recabar toda la información necesaria para el desarrollo del Capítulo 4: Resultados, para así tratar de responder a las preguntas planteadas como punto de partida del presente trabajo.

Para la definición del método de análisis se han diferenciado dos grandes bloques:

- → Patent Outlink: Análisis de enlaces incluidos en Patentes a contenidos Web
- → Patent Inlink: Análisis de enlaces incluidos en contenido Web a Patentes

De este modo, como muestra la Figura 20, se separa la localización de enlaces dentro de las patentes de la búsqueda de las patentes más relevantes, de forma que, aunque el método final pueda ser utilizado de forma unificada, permita de igual manera un proceso de análisis separado.

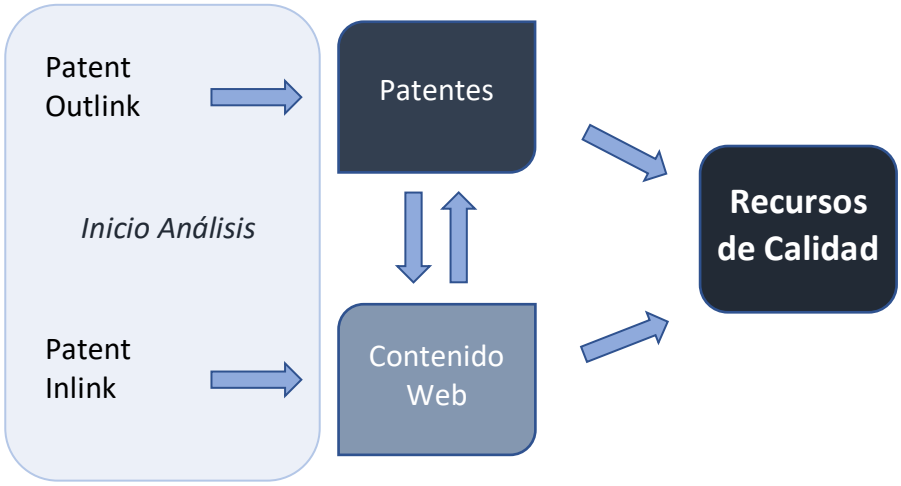


Figura 20: Esquema del flujo de información en el proceso metodológico por bloques

Esta división por bloques permite la realización de los estudios de modo complementario. El Bloque Patent Outlink permite localizar los recursos de información contenidos en las patentes, mientras que el Bloque Patent Inlink permite conocer el impacto y visibilidad web de las patentes mediante la localización del contenido web que las enlaza –potencialmente de calidad–, permitiendo el estudio independiente en caso de ser oportuno.

Para poder realizar un análisis de estudio aplicado y comprobar la viabilidad de los resultados en el Capítulo 4, en este punto cabe indicar que para la selección de corpus de patentes a analizar se han escogido aquellas que cumplen con los siguientes requisitos:

- → Oficina: Estados Unidos (USPTO)
- → Estado: Concedida
- → Año de concesión: 2008 – 2018

Esto supone un total de 3.133.247 de patentes en un periodo de 10 años.

La decisión de obtener la muestra entre las patentes concedidas en Estados Unidos desde 2008 a 2018 se basa en diversos criterios:

- → Como se indica en el §2.1.5 la USPTO es la segunda Oficina con mayor volumen de solicitudes y concesiones, siendo la primera la Oficina China. Debido a la internacionalización del sistema norteamericano (en US se conceden un 33% de patentes a residentes, frente al 73% en China), así cómo posibles problemas idiomáticos, se decide seleccionar la USPTO.
- → Se opta por seleccionar el documento de patente ya concedida, debido a que es el documento estable, se encuentra aceptado por la Oficina tras revisiones y exámenes por parte de personal especializado, por lo que contienen información muy valiosa relativa a comentarios y aportaciones de referencias.
- → La selección temporal de 10 años permitirá obtener datos relativos a la evolución.
- → No se realiza ninguna diferencia por áreas, lo que permitirá estudiar posibles diferentes entre ellas.

El proceso de trabajo en los dos bloques de estudio es el mismo, consta de tres etapas que aglutinan todos los pasos (Figura 21) a realizar para la obtención de los datos a analizar:

1. Selección de la fuente de información
2. Diseño y desarrollo del proceso de recogida de datos
3. Diseño de un sistema de limpieza, preparación y almacenamiento de los datos

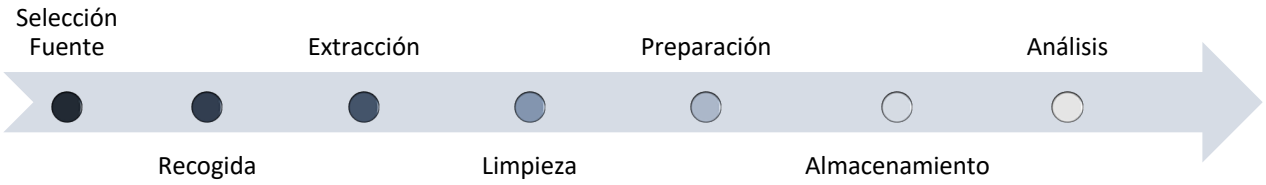


Figura 21: Proceso de ejecución de la metodología. Fuente: elaboración propia

3.1. Análisis de enlaces incluidos en Patentes a Contenido Web

En un documento de patente, tal y como se ha visto en el §2.1.4, existe una gran cantidad de información repartida en diferentes secciones. Hasta ahora, tal y como se indica en el §2.1.5, los estudios de análisis de la información contenida en patentes se han centrado en el análisis de citas o en palabras clave (*Technological forecasting* o Patentometría). Pero el presente estudio busca realizar un análisis de todos los enlaces que puedan aparecer en un documento de patente, independientemente de la sección de la que se trate.

Para ello primero se ha realizado un análisis manual de las posibles localizaciones de enlaces en el documento. Se comprueba que los enlaces se localizan en tres de las secciones que conforman una patente:

- Citas
- Descripción
- Reivindicaciones

Una vez conocidas las posibles localizaciones de enlaces dentro de un documento, se pasa a realizar la selección de la fuente de información.

En el §3.3 se puede encontrar la Figura 24 representando un resumen esquematizado del proceso llevado a cabo en la presente sección.

3.1.1. Fuentes de información

Como se ha visto existen multitud de fuentes de información con diferentes tipos de características, que permiten acceder a diferentes partes de los documentos y, en algunos casos, tienen herramientas complementarias que permiten realizar análisis en profundidad.

Dado que tal y como se indica anteriormente se pueden encontrar enlaces en diferentes secciones del cuerpo del documento, es necesario acceder al texto completo de la misma para poder extraerlos todos junto con su localización en el documento. Las fuentes que no ofrecen el texto completo únicamente permiten la visualización de la información bibliográfica (portada, resumen, citas), por lo que se perderían enlaces que pueden ofrecer información muy valiosa. Por este motivo, se debe escoger entre las diferentes fuentes que si ofrezcan toda la patente.

Entre las fuentes de información que sí ofrecen la patente a texto completo y, además, ofrecen otro tipo de herramientas complementarias, algunas de estas herramientas permiten hacer un seguimiento de las citas que existen entre patentes (tanto entrantes como salientes) y desde las patentes a las citas no-patentes. Pero no permiten extraer o analizar todos los enlaces que existen en el documento, lo que refuerza la necesidad del método planteado en la presente tesis. Es por este motivo, que la fuente de información a seleccionar no requerirá de elementos extra, únicamente será necesario que ofrezca el documento.

Otra cosa que se debe tener en cuenta en la selección de fuentes de información es la forma de acceso a la misma. Las diferentes bases de datos pueden ofrecer acceder a la patente de tres formas principalmente:

- → Acceso online
- → Acceso vía API
- → Acceso mediante descarga

Acceso online

Dado que se trata de millones de documentos los que deben ser analizados, la opción online presenta diferentes limitaciones:

- → **Acceso vía URL a cada una de las patentes**: es necesario conocer la dirección de acceso a cada uno de los documentos, dependiendo de la base de datos, esto puede ser complejo de solucionar ya que no todas las páginas mantienen un formato estándar de URL modificando únicamente el número de la patente:

  https://www.basededatos.com/*númerodepatente*

  Por lo que sería necesario realizar tantas consultas a la base de datos como patentes existan, primero para conocer la URL de acceso, después para acceder al documento para la descarga de información, lo que puede derivar, a su vez, en errores de carga.
- → **Posibles errores de carga**: pese a que los sistemas se pueden preparar para evitar problemas de acceso (errores 400 o 500), es posible que estos se den y sea necesario repetir la búsqueda, lo que ralentiza la toma de datos.
- → **Posibles errores de recogida de datos (*scraping*)**: debido a la gran cantidad de datos a recopilar la extracción de datos debe ser realizada de forma automática mediante bots. El uso de estos puede dar errores en el momento de recopilación de información (problemas de cambios en el código HTML desde el que se lee, errores de acceso a la información, errores de lectura, etc.)
- → **Posibles limitaciones de acceso por volumen de búsqueda**: las páginas se protegen de los accesos automáticos de bots bloqueando el acceso para evitar colapsos o saturación de su servicio, por lo que realizar más de tres millones de accesos y consultas puede ser un problema.

Debido a estas limitaciones, se descarta la toma de datos de fuentes de información de tipo online.

Acceso vía API

Con respecto a las opciones vía API, pese a que en los últimos años han surgido cada vez más sistemas que ofrecen acceder a los documentos de patentes completos, no todos los servicios tienen acceso vía API.

Además, no todas las bases de datos ofrecen todos los campos contenidos en una patente. Por ejemplo, en el caso de PatentsView[57] (portal con información sobre datos

[57] www.patentsview.org

de patentes de Estados Unidos [citas, impacto, datos, gráficas, ilustraciones, datos bibliográficos, etc.]) la API de acceso no permite recoger información del texto completo de la patente, aunque sí de otros campos.

En algunas fuentes de datos se debe tener en cuenta que el acceso mediante API a la información puede suponer un coste extra, así como el volumen de consultas, por lo que se deberá tener en cuenta en caso de utilizar esta opción.

Para recabar la información necesaria para el estudio planteado, no se ha localizado una base de datos que asegure contener la totalidad de las patentes de Estados Unidos[58] y permita acceder a todos los campos necesarios para el estudio por lo que se descarta esta vía.

Acceso mediante descarga

Algunas de las bases de datos permiten el acceso y descarga de la información para su posterior análisis en local, permitiendo la descarga en diferentes tipos de formatos (imagen, PDF, XML).

Dado que la Oficina de patentes de Estados Unidos ofrece acceso a sus patentes mediante las tres vías, se selecciona el método de descarga directa para la recopilación de los datos.

3.1.1.1. Análisis de los datos ofrecidos por la USPTO

La Oficina de Patentes de Estados Unidos[59] ofrece una gran cantidad de información relacionada con las patentes, marcas y políticas de protección intelectual, así como recursos, publicaciones y herramientas.

A modo de resumen, en la Tabla 11 se listan las diferentes opciones que se ofrecen en el portal para acceder a la información relacionada con los documentos de patentes.

Dado que la opción escogida para realizar el estudio es la descarga masiva y en bloque de datos, se opta por la opción de descarga mediante la página *Bulk Download*. En esta página se puede descargar bloques de información sobre patentes concedidas, solicitudes de patentes, información adicional sobre patentes (dueños, clasificación, tasas, etc.), marcas y conjuntos de datos de investigación. En la Tabla 12 se encuentra una breve explicación de los datos alojados relativos a patentes concedidas en la página. Todos estos conjuntos de datos se encuentran actualizados de forma semanal (cada martes se realiza la publicación).

[58] La mayoría de bases de datos globales (aquellas que beben de varias Oficinas) como Lens o Google Patentes, no pueden asegurar contener todas las patentes publicadas por las Oficinas desde las que recogen los documentos.

[59] https://www.uspto.gov/

Tabla 11: Portales de datos para la visualización y descarga de patentes en la UPSTO.

Nombre Portal	URL	Tipo de información
Open Data Portal (ODP)	https://developer.uspto.gov/	Portal de datos abiertos de la USPTO, en el que se facilita el acceso a APIs, visualizaciones y datos.
ODP Datasets	https://developer.uspto.gov/data	Portal de descarga directa en diferentes formatos (XML, SGML, APS, imagen). La actualización puede ser diaria, semanal o mensual.
ODP APIs	https://developer.uspto.gov/api-catalog	Página que recopila todas las APIs ofrecidas por la USPTO, junto con su descripción y documentación
ODP Visualizaciones	https://developer.uspto.gov/visualizations	Portal de visualización de datos relativos a patentes y publicaciones. Existen conjuntos de citas, inventores, marcas, empresas, datos geográficos, etc.
Acceso online individual	http://patft.uspto.gov/	Portal de acceso a las bases de datos de texto completo PatFT y AppFT, así como la documentación y estado de ambas
PatFT	http://patft.uspto.gov/netahtml/PTO/search-bool.html	Patent Full Text Database, contiene las concesiones publicadas desde 1976 (HTML o imagen). Se actualiza semanalmente. No contiene todas las patentes[60].
AppFT	http://appft.uspto.gov/netahtml/PTO/search-bool.html	Contiene las aplicaciones realizadas a la Oficina desde 2001. Se actualiza semanalmente.
Bulk Download	https://bulkdata.uspto.gov/	Página de descarga en bloque de patentes (concedidas y aplicaciones), información adicional y marcas.

[60] http://patft.uspto.gov/netahtml/PTO/help/contents.htm

Tabla 12: Resumen datos ofrecidos para la descarga en bloque de patentes concedidas de la USPTO

Tipo descarga	Contenido	Fechas
Boletín Oficial de Patentes	Información bibliográfica, reivindicación representativa y un dibujo de todas las concesiones de la semana, así como avisos oficiales	2 julio 2002 – Presente
PDF múltiples páginas	Imágenes de las concesiones en formato PDF	31 julio 1790 – Presente
TIFF única página	Imágenes de la primera página de las patentes concedidas en formato TIFF	31 julio 1790 – Presente
Texto Completo con imágenes TIFF	Texto completo de la patente en formato XML junto con imágenes de los dibujos, esquemas, etc. en formato TIFF	Enero 2001 – Presente
Texto Completo sin imágenes	Texto completo de todas las patentes sin dibujos, esquemas, etc. en formato XML	Enero 1976 – Presente
Datos Bibliográficos	Primera página de cada patente concedida (sin imágenes ni gráficos) en formato XML/APS/SGML	Enero 1976 – Presente

3.1.2. Proceso de recogida de datos

Como se ha descrito en la sección anterior, la opción escogida para el acceso a la información es mediante descarga en bloque de datos. Por ello, se utilizarán los datos contenidos en el bloque "*Texto Completo sin imágenes*" recogido en la Tabla 12.

Los datos se encuentran publicados en agrupaciones anuales, al acceder al año del que se desean extraer los ficheros, se accede a un listado con las publicaciones semanales. Debe tenerse presente que, pese a que los ficheros de descarga pesan entre 3 y 6 GBs por año alojados en el disco duro; al descomprimirse, pueden llegar a pesar hasta 50 GBs/año, lo que puede suponer un problema en el momento de trabajar con ellos.

Junto con los datos agrupados anualmente, se puede encontrar un fichero .DTD. Este tipo de ficheros contiene la definición de la estructura de un documento XML, describiendo sus elementos, atributos, entidades, anotaciones, etc. Indicando el sistema de notación, aparición y jerarquización; sirviendo, además, de sistema de validación[61]. Además, desde la Oficina se publica un documento de tipo Word que contiene la

[61] Una vez generado un documento XML, se puede comparar el código contra un fichero DTD y evitar problemas de esquematización

documentación de la versión, con la explicación y detalle de todos los elementos contenidos en el XML.

Debido a la evolución propia de cualquier sistema, los documentos se encuentran publicados en diferentes versiones del esquema, por lo tanto, se deberá tener en cuenta en el momento de realizar la extracción de datos. En la Tabla 13 se encuentra reflejada la versión de los documentos, junto con el peso comprimido y la cantidad de ficheros recogidos para los años del estudio.

Dentro de cada fichero comprimido de tipo ZIP se encuentra un documento XML que contiene la agrupación de todas las patentes publicadas durante esa semana en formato XML.

Tabla 13: Recopilación de ficheros, pesos y versiones relativos a los documentos de patentes a utilizar

Año	Nº Ficheros	Versión	Peso descarga
2008	54	4.2	2,49 Gb
2009	53	4.2	2,72 GB
2010	53	4.2	3,65 GB
2011	53	4.2	3,79 GB
2012	53	4.2	4,30 GB
2013	53	4.3/4.4	4,85 GB
2014	53	4.4	5,73 GB
2015	52	4.5	5,41 GB
2016	52	4.5	5,50 GB
2017	52	4.5	6 GB
2018	52	4.5	5,75 GB

3.1.3. Sistema de extracción

Un vez recogidos y extraídos los ficheros XML, de el análisis del contenido para localizar el mejor sistema de recogida de los enlaces contenidos en éstos.

Existen dos puntos críticos a abordar por el sistema de extracción y que serán descritos en la presente sección:

- → La lectura desde los ficheros XML
- → La recogida de los enlaces contenidos

Debido a la naturaleza específica de esta investigación, así como los problemas que se describen a continuación, se desarrolla un programa ad hoc capaz de solventar todas las dificultades localizadas. Este programa se encuentra desarrollado en Java y supone uno de los aportes más importantes de la presente tesis.

3.1.3.1. Ficheros XML

En el apartado anterior se indica que la información ofrecida por la USPTO se encuentra en ficheros de tipo XML. XML[62] es un lenguaje de marcado jerarquizado que favorece la publicación y distribución a gran escala de información; y que permite la personalización de las etiquetas para la descripción y organización de datos logrando que estas se adapten perfectamente a las necesidades de la información a contener y compartir.

Los documentos redactados usando el lenguaje XML se caracterizan por seguir unas normas básicas adaptables a las necesidades del marcado que sea requerido. Todos los documentos XML se encuentran formados por dos secciones: una cabecera y el cuerpo.

En la cabecera o prólogo del documento se encuentra la descripción de este:

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE us-patent-grant SYSTEM "us-patent-grant-v45-2014-04-03.dtd">
```

Estas dos etiquetas indican que se trata de un documento XML y el tipo de esquema (us-patent-grant) y versión que sigue (us-patent-grant-v45-2014-04-03.dtd). En la Figura 22 se puede encontrar un ejemplo del cuerpo de un documento XML, perteneciendo éste a uno de los documentos de patente recogidos para el análisis.

```
<us-patent-grant lang="EN" date-publ="20180102" date-produced="20171218" country="US" id="us-patent-grant" status="PRODUCTION" file="USD0806350-20180102.XML" dtd-version="v4.5 2014-04-03">
  - <us-bibliographic-data-grant>
    - <publication-reference>
      - <document-id>
          <country>US</country>
          <doc-number>D0806350</doc-number>
          <kind>S1</kind>
          <date>20180102</date>
        </document-id>
      </publication-reference>
    - <application-reference appl-type="design">
      - <document-id>
          <country>US</country>
          <doc-number>35500953</doc-number>
          <date>20151119</date>
        </document-id>
      </application-reference>
      <us-application-series-code>35</us-application-series-code>
    - <priority-claims>
      - <priority-claim kind="regional" sequence="01">
          <country>EM</country>
          <doc-number>002705756-0001</doc-number>
          <date>20150522</date>
        </priority-claim>
```

Figura 22: Ejemplo cuerpo documento XML patente perteneciente a la USPTO

El cuerpo de un documento XML se encuentra definido por las etiquetas que lo conforman. Las etiquetas se encuentran delimitadas por los símbolos <> y la información contenida en ellas debe encontrarse a su vez delimitada por la etiqueta de cerrado </>.

[62] https://www.w3.org/XML/

<doc-number>D0806350</doc-number>

Etiqueta Información Etiqueta Fin

Las etiquetas pueden encontrarse anidadas. Además, existen otros elementos de marcado, también contenidos dentro de <>, y que se encuentran predefinidos, es decir, no pueden ser utilizados como etiquetas de marcado ya que existen en el propio lenguaje (i.e.: <b> –> negrita, <p> –> párrafo).

Para la lectura de un fichero XML no es necesario ningún tipo de programa especializado, aunque para la gestión y uso de la información contenida sí se requiere conocer el esquema. Para poder manipular la información es muy útil el fichero de documentación que ofrece la USPTO. De éste se puede extraer el esquema base que siguen los ficheros, en el Anexo II se puede encontrar un resumen esquematizado extraído de la documentación ofrecido por la USPTO realizado para facilitar la tarea de extracción.

Para evitar que el sistema tenga un alto coste temporal en la búsqueda de enlaces, se localizan las etiquetas que deberán ser recorridas por el programa que extraerá los enlaces. Las etiquetas donde puede encontrarse un hiperenlace son:

- <othercit>
- <abstract>
- <description>
- <claims>

Además, para poder realizar análisis en mayor profundidad, se recogerán de los documentos de patentes las etiquetas relativas a la clasificación del documento:

- <classification-national>
- <classification-locarno>
- <classification-ipcr>

Estas etiquetas corresponden a tres categorías diferentes:

- Nacional: sistema de categorización propietario del sistema de patentes estadounidense
- CPC: Cooperative Patent Classification, sistema desarrollado conjuntamente por la OEP y la USPTO, Estados Unidos firma su uso en 2010, pero no es hasta 2013 cuando realiza el primer año de transición para categorizar sus patentes utilizándolo, empezando a ser totalmente efectivo desde 2014
- ICP: International Patent Classification, sistema internacional creado por la WIPO en 1971, y utilizado por más de 100 países.

Debido a que del sistema CPC no existen datos para todos los años y el nacional se encuentra acotado y limitado por el país, se decide realizar el análisis de los datos utilizando el sistema ICP, de este modo se podrá realizar comparativas con otros países en el futuro, aunque se tengan los datos para todas las clasificaciones recogidos.

El sistema ICP se encuentra dividido en 8 áreas:

- A: Necesidades humanas
- B: Técnicas Industriales; Transportes
- C: Química; Metalurgia
- D: Textiles; Papel
- E: Construcciones Fijas
- F: Mecánica; Iluminación; Calefacción; Armamento; Voladuras
- G: Física
- H: Electricidad

De este modo, el programa accederá al fichero y realizará una lectura vertical, recorriéndolo y localizando las etiquetas mencionadas anteriormente, para acceder a ellas y buscar dentro del texto las URLs a extraer de forma automática y autónoma.

Durante la preparación del programa se han encontrado las siguientes particularidades:

- Los ficheros ofrecidos por la USPTO son la recopilación de todas las patentes publicadas durante una semana, por lo tanto, el esquema de marcado no se sigue correctamente (un documento XML contiene una única cabecera, al encontrarse todos los documentos de forma continua ésta se repite en bucle generando un problema de lectura).
- No todos los ficheros mantienen el mismo esquema, existen etiquetas obligatorias, opcionales o que pueden repetirse, por lo que el programa debe ser capaz de recorrer el fichero en búsqueda de etiquetas que pueden no estar o encontrarse repetidas.
- Existen símbolos incluidos en el texto que pueden generar problemas de lectura
- Pueden existir errores de transcripción

Para solventarlos el programa tiene las siguientes funcionalidades:

- Permite recorrer un fichero XML, detectar si existe más de una cabecera y dividir el fichero guardando en el disco duro tantos ficheros XML únicos, como cabeceras existan[63]. Es decir, el programa recorre el fichero XML y cada vez que detecta una patente la guarda en un fichero XML separado. De este modo se obtiene la totalidad de documentos de patente en ficheros individuales.
- Se realiza un análisis textual (*parsing)* a todos los documentos para que los símbolos contenidos aparezcan correctamente.
- Se utiliza un sistema de fórmulas de expresión (*RegEx [Regular Expression]*) para extraer las URLs (explicado en detalle a continuación)

[63] Es posible realizar la extracción de URLs sin necesidad de dividir el fichero realizando otro tipo de programación, pero el coste computacional es excesivamente alto, con el método actual el programa extrae las URLs de un año en menos de 10 horas, sin dividir el fichero puede tardar hasta dos días. En este caso se prima la velocidad a la capacidad de almacenamiento debido a que la cantidad de documentos a analizar es de millones y alargaría la extracción durante días.

Debido a los problemas de almacenamiento que genera la descompresión de todos los ficheros anuales a la vez, se trabaja año a año en la descompresión, división y lectura de los ficheros XML. La Tabla 14 muestra la cantidad de ficheros que se generan tras realizar la división a XML individual.

Tabla 14: Recuento de ficheros individuales extraídos al dividir los XML originales
Fuente: elaboración propia

Año	Cantidad de ficheros XML únicos
2008	189.558
2009	196.687
2010	250.550
2011	254.066
2012	283.407
2013	310.279
2014	334.110
2015	334.128
2016	341.954
2017	360.191
2018	348.786
	3.203.716 Ficheros totales

Dentro de los ficheros divididos, hay que tener en cuenta, que para su publicación la USPTO incluye, para aquellas patentes con datos relativos a secuencias genéticas o ciertas estructuras químicas, un fichero XML extra para su correcta descripción. Por lo tanto, de la cantidad de ficheros XML indicada, no todos corresponden con el cuerpo completo del documento de la patente. Los datos relativos a las cifras exactas del número de patentes se encuentran más adelante en la Tabla 17.

3.1.3.2. Recogida de enlaces contenidos

Realizar la recogida de enlaces supone otro punto crítico debido a diversos factores. El primero de ellos es que, así como en HTML (también lenguaje de marcado) existe una etiqueta (<a>) para indicar que el contenido incluido en su referencia es un hiperenlace (*href*), en XML no existe una etiqueta predefinida como las indicadas para delimitar párrafos o formato de letras. Aunque el W3C recomienda utilizar XLink, que funciona como *href* de HTML para etiquetar y delimitar enlaces dentro de documentos XML, los XML descargados no lo implementan; esto supone un problema a la hora de localizar enlaces dentro del cuerpo del documento.

Para solventar esto se puede utilizar un sistema de búsqueda de cadenas de texto que permita localizar todo aquello que se asemeje a una URL, algo que en sí mismo implica diversos puntos críticos como se explica a continuación. En aras de solventar esto se debe entender cómo se genera una URL, las partes que la conforman y mediante qué fórmulas pueden ser buscadas en textos.

3.1.3.2.1. Formación de URLs

La sintaxis de una URL se encuentra formada por tres partes[64]:

- → Protocolo: 'http', 'https', 'ftp', 'telnet', 'gopher', 'data', 'file', etc.
- → Dominio: 'www.nombreweb.com' o 'es.nombreweb.com'. Formado por:
  - ⇢ Subdominio (opcional): subconjunto de ficheros anexos a un dominio: 'www'[65], 'es' (idioma), 'grupoinvestigacion', etc.
  - ⇢ Nombre del dominio: nombre del espacio web
  - ⇢ *Top Level Domain* (TLD): indica el propósito del sitio web: país, categoría u organización.
- → Ruta de acceso (*path*): indicado mediante '/' tras el dominio, dirige a un archivo concreto
  - ⇢ Consulta: delimitado con '?', aportar información sobre una búsqueda para localización de información en el servidor, i.e.: '?pagina=7'
  - ⇢ Indicador: delimitado con '#' permite navegar dentro de un fichero hasta la posición indicada (sección, imagen, titulo, etc.): '#formaURL'

En ocasiones puede encontrarse un puerto de acceso (i.e.: 8080) entre el dominio y la ruta de acceso, lo que daría una URL completa de ejemplo similar a la siguiente:

https://www.nombreweb.com:8080/path?consulta#indicador

En la Tabla 15 se describe brevemente los elementos para la formación de una URL.

Tabla 15: Descripción y formación de las diferentes partes que conforman una URL
Fuente: elaboración propia

Sección	Opcional	Formación
Protocolo	Si (en caso de ser HTTP, los navegadores asumen que la búsqueda usa dicho protocolo)	Estándar fijo ('http', 'ftp', etc.) seguido de '://'
Dominio	No	Máximo 65 caracteres combinando: [a-z]: cualquier letra [0-9]: cualquier número [-]: guion (no puede ir en primera o última posición)
Ruta	Si. Pese a que la página principal de un portal web es un documento de tipo HTML los navegadores no necesitan la dirección del fichero exacto para acceder a la portada	Máximo 2.048 caracteres combinando: [a-z]: cualquier letra [0-9]: cualquier número Los símbolos: '-', '.', '_', '~', '!', '$', '&', '(', ')', '*', '+', ',', ';', '@' y '%' Los caracteres ':', '=', '&', '/' están reservados

[64] https://www.w3.org/Addressing/URL/url-spec.html

[65] Siguiendo las indicaciones del W3C se trata www como un subdominio pero es también un protocolo.

3.1.3.2.2. Uso de fórmulas para la localización de URLs

Debido a que en los ficheros los enlaces no se encuentran delimitados mediante etiquetas, éstos deben ser buscados en el texto por comparación de cadenas de texto. Para realizar esta comparación se utilizan Expresiones Regulares (RegEx), mediante las que se indica a la máquina la secuencia de caracteres y patrones que debe contener la cadena de texto a buscar.

Las Expresiones Regulares son utilizadas por multitud de lenguajes de programación para localizar textos en motores de búsqueda, procesadores de textos, análisis de textos, etc.

Para formar una expresión regular únicamente es necesario conocer la sintaxis y los símbolos que describen los elementos a buscar. En la Tabla 16 se encuentra una selección de los elementos más comunes utilizados para la definición de RegEx.

Los caracteres especiales indicados (‘.’, ‘\’, ‘*’, etc.) pueden ser buscados en el texto como elementos de la cadena utilizando una barra inversa ‘\’ para escaparlos e indicar que no hacen referencia a los elementos reservados para operaciones: ‘*’ –> busca el símbolo ‘*’ en el texto.

Tabla 16: Recopilación de los elementos sintácticos más utilizados para la formación de RegEx
Fuente: elaboración propia

Carácter	Descripción	Carácter	Descripción
.	Cualquier carácter	**$**	Final de línea
\w \d \s	Palabra, dígito, espacio en blanco	**(abc)**	Grupo de captura, puede ser alguno de los caracteres recogidos
[abc]	Coincide con a, b o c	**(?:abc)**	Grupo de No captura
[^]	No coincide	**ab|cd**	Concuerda con ab o cd
[a-z][A-Z]	Alguna letra minúscula o mayúscula	**ab+c**	Localiza el elemento precedente una o más veces: ‘abc’, ‘abbc’, ‘abbbc’
[0-9]	Algún número	**ab*c**	Localiza el elemento precedente cero o más veces: ‘ac’, ‘abc’, ‘abbc’
^	Comienzo de línea	**ab?c**	Localiza el elemento precedente cero o una vez: ‘ac’ o ‘abc’

Para localizar las URLs contenidas en los ficheros XML se realiza una primera extracción realizando una búsqueda de todas las URLs contenidas que comiencen utilizando los protocolos HTTP, HTTPS o FTP, indicando al programa desarrollado la fórmula RegEx:

```
(http|ftp|https)://([\w_-]+(?:(?:\.[\w_-]+)+))([\w.,@?^=%&:/~+#-]*[\w@?^=%&/~+#-])?
```

Una vez ejecutado para los años 2008 – 2010, se extraen por año entre 7.000 y 10.000 URLs. Tras una comprobación manual, se detecta que esas no son todas las URLs contenidas en los ficheros, existen URLs que no incluyen uno de los protocolos seleccionados o empiezan por www.

Por ello, se plantea la búsqueda mediante TLDs; en lugar de localizar el protocolo y extraer todo aquello que se encuentre detrás hasta un espacio en blanco, se busca dentro del texto la fórmula '.TLD' y se recoge aquello que coincida con una URL antes y después. Para ello, se descarga el listado de aquellos TLD que tienen más de un 0,1% de uso según el portal W3Techs[66], y se prepara una fórmula que permita recoger las cadenas de texto que precedan y sigan a un TLD del listado.

Esto genera diversos errores:

- → Existen URLs rotas por separaciones de espacios o dobles puntos (i.e.: url..com/ url com)
- → Debido a errores en la redacción del XML existen palabras que son detectadas como TLD (i.e.: end-of-**sentence.The** next sentence)
- → Debido a que en las patentes se encuentran recogidas fórmulas (químicas, ADN, etc.) que son redactadas mediante cadenas de letras y puntos, se detectan falsos TLDs (i.e.: bbEEbbaaAA.be)

Por ello, se decide realizar dos extracciones de forma que se minimice lo máximo posible la pérdida de datos. Para cada extracción se utiliza un RegEx diferente.

La primera fórmula busca todas las URLs que contengan únicamente ciertos TLD, se seleccionan en total 92 TLDs (listados en la fórmula) que representan el 97,7% de dominios en internet. Esto se logra mediante la concatenación de las siguientes partes que conforman el RegEx:

`[a-zA-Z0-9][a-zA-Z0-9\.-]` –> Cualquier combinación de letras y números que contenga letras en mayúsculas y minúsculas, números y/o el símbolo guion '-'.

```
*\.(ae|ai|ar|au|az|bd|be|bg|ca|cf|ch|cn|co(m)?|cz|dk|ee|es|eu|fr|ge|gr|hk|hr|hu|i
d|ie|il|io|ir|jp|kr|kz|lk|lt|lv|ma|me|mx|my|ng|nl|no|nz|ph|pk|pl|pt|ro|rs|ru|sa|sg
|si(te)?|sk|su|tk|tr|tv|tw|ua|uk|us|uz|vn|za|info|live|net|online|org|shop|store|xyz
|biz|pro|edu|gov)
```

–> Seguida de un símbolo punto '.' y uno de los posibles TLDs seleccionados.

`\b(:\d+)?(\/[-a-zA-Z0-9@:%_—-\+\.~#\?&//=\$,;ºª\*\\]+)?"` –> Que puede estar o no seguido de un símbolo barra '/' seguido de cualquier combinación alfanumérica y los símbolos: '@', ':', '%'. '_', '-', '–', '+', '.', '~', '#', '&', '/', '=', '$', ',', 'ª', 'º', '*'

De este modo se obtiene el grueso de enlaces a analizar.

[66] https://w3techs.com/technologies/overview/top_level_domain

Mediante la segunda fórmula se recoge todas aquellas URLs que sí empiezan por protocolo HTTP, HTTPS o FTP, y que tienen una estructura de URL:

`(http|ftp|https)://` –> Busca los protocolos en el texto seguidos de los barras '/'

`([\w+?\.\w+])` –> Que puede estar seguidos o no del protocolo www.

`+([a-zA-Z0-9\~\!\@\#\$\%\^\&\*\(\)_\-\=\+\\\/\?\.\:\;\'\,]+)` –> Y que presenta cualquier combinación alfanumérica (con o sin guion) que precede a un punto '.' y que puede estar o no seguida de una barra '/' y otra combinación alfanumérica con/sin símbolos.

Para que no existan problemas de duplicados entre las URLs extraídas buscando el TLD y aquellas que se buscan mediante protocolo, el programa permite utilizar los dos RegEx simultáneamente, indicando que el primero debe excluirse. De este modo, la segunda fórmula extrae todas aquellas URLs que comienzan por protocolo y que no contienen uno de los TLDs extraídos mediante la primera fórmula.

Realizando el proceso utilizando las dos fórmulas, se pasa a recoger entre 89.000 y 173.000 URLs anuales para los años 2008 – 2010, lo que supone un incremento superior al 1.200% con respecto a la opción de recogida inicial mediante el uso único del protocolo.

3.1.4. Sistema de limpieza, preparación y almacenamiento

Una extraídas todas las URLs, el sistema genera un fichero CSV que contiene fila a fila la siguiente estructura:

Fichero	Nº Patente	Año	Zona Aparición	Clasificación	URL

- → **Fichero**: nombre del fichero desde el que se extrae. Se utiliza para permitir la trazabilidad de extracción, y así saber exactamente dónde se debe buscar la URL para entender el contexto en caso de error.
- → **Nº Patente**: número de la patente
- → **Año**: año de concesión
- → **Zona aparición**: dónde se ha localizado la URL (otras citas, descripción, reivindicaciones)
- → **Clasificación**: clasificación a la que pertenece la patente
- → **URL**

Para la preparación de los datos se utiliza en primera instancia Excel, de forma que permita controlar manualmente los resultados, una vez conocidas las limitaciones y modificaciones que se deben realizar se pasa a utilizar R para la automatización y uso a gran escala. En la Figura 24 se puede encontrar el resumen esquematizado de éste proceso.

Debido a errores en los documentos XML, la recolección de datos, pese a que se opta por utilizar un sistema doble de RegEx para evitar ruido en los datos, al a ser masiva y

contener la mayor cantidad de URLs posible también contiene errores debido a la forma en que se redactan los ficheros XML. El siguiente listado contiene aquellos que deben ser tenidos en cuenta a la hora de limpiar los datos:

- Existen URLs rotas por espacios: 'mit edu'
- Existen errores en el protocolo inicial: 'http:///' o 'wwww.'
- Errores de transcripción:
  - Cambio de letras: 'mit.edu' –> 'mil.cdu'
  - Cambio de símbolos: '-' por '.' o viceversa
- Por tipo de redacción: se transforman símbolos por letras:
  - Ctep(dot)cáncer(dot)gov(slash)
- Errores de salto de línea: tras un punto se recoge la primera palabra del siguiente párrafo

Además, pese a que se realiza un *parsing* en los documentos para a correcta lectura de los símbolos y letras, en la redacción de los documentos se encuentran URLs con el símbolo de guion largo o *em dash* '–'[67] usado como subíndice, que al ser recogido y, posteriormente, utilizado los navegadores no traducen correctamente.

Debido a estos problemas, el alcance del estudio se realizará a nivel de dominio y subdominio, descartando el análisis del *path* completo ya que se detectan muchos errores por URLs rotas tras la barra del directorio '/'. Se realizará un análisis preliminar de URLs que contengan ficheros (PDF, XLS, DOC, etc.) teniendo en cuenta la limitación encontrada.

Para la limpieza de los datos se procede a preparar las URLs y realizar una clasificación inicial y comprobar la viabilidad y naturaleza de las URLs. Para eso se analizan los años 2008, 2009 y 2010 manualmente. Se genera una nueva columna en la que se guarda la URL sin protocolos, subdominio www. y sin *path*. En Excel esto se realiza con la fórmula:

```
=SUSTITUIR(REEMPLAZAR(REEMPLAZAR(N2; 1; SI.ERROR(ENCONTRAR("//"; N2)+1; 0);
"")&"/"; ENCONTRAR("/";REEMPLAZAR(N2; 1; SI.ERROR(ENCONTRAR("//"; N2)+1; 0);
"")&"/"); LARGO(N2); ""); "www."; "")
```

En R se realiza utilizando la librería *urltools*, usando el comando

```
suffix_extract(domain(URLs))
```

Donde URLs es el conjunto de URLs almacenadas.

De este modo se extraen las URLs iniciales sobre las que realizará el análisis tras su limpieza, la Tabla 17 muestra la cantidad de enlaces extraídos inicialmente con cada método.

[67] https://www.codetable.net/hex/2014

Tabla 17: Total en bruto de las URLs extraídas mediante las dos fórmulas utilizadas. Fuente: elaboración propia

	Nº Patentes	Bruto RegEx1	Bruto RegEx2
2008	185.260	87.856	1.745
2009	192.052	113.307	2.250
2010	244.599	169.027	3.358
2011	248.101	184.244	3.738
2012	277.285	215.477	4.562
2013	303.642	268.682	6.106
2014	327.014	301.394	6.761
2015	326.969	291.437	7.042
2016	334.674	318.628	7.558
2017	352.547	367.762	8.440
2018	341.104	363.500	7.760
Total	**3.133.247**	**2.681.314**	**59.320**

Una vez almacenadas las URLs únicamente con su dominio, se realiza una limpieza manual para comprobar el estado. Como se puede ver en la Tabla 18 el análisis muestra que el porcentaje de error localizado en el fichero extraído para 2008 mediante la fórmula RegEx1 es de 5,82%.

Tabla 18: Datos y porcentajes de tipos de URLs recogidas en fichero 2008 para RegEx1. Fuente: elaboración propia

Estado	Cantidad de URLs	Porcentaje
Correctas	84.563	94,18%
Error	1352	1,50%
Typos	450	2,65%
Ejemplos	1494	1,66%
Total	**89.793**	**100%**

En el caso de las URLs extraídas mediante RegEx2 el análisis ofrece los datos en porcentajes mostrados en la Tabla 19.

Tabla 19: Porcentaje de errores clasificados con extracción mediante RegEx2. Fuente: elaboración propia

RegEx2	Bien	Bien - Error	Ejemplo	Error	IP	Rota
2008	44,58	13,75	12,38	1,72	11,58	15,99
2009	42,22	14,13	18,58	1,07	9,33	14,67
2010	43,66	14,15	13,85	1,07	11,94	15,34
2011	46,15	13,83	11,16	1,71	9,26	17,90
2012	49,32	15,41	7,98	2,70	9,10	15,50
2013	49,66	14,56	14,20	2,64	5,98	12,97
2014	52,77	13,40	9,14	2,68	6,58	15,43
2015	51,16	11,46	11,76	3,29	5,23	17,10
2016	58,00	11,09	7,25	2,65	5,05	15,96
2017	57,33	10,62	9,28	2,27	4,45	16,04
2018	58,61	12,28	9,27	0,94	4,06	14,85

Leyenda: Bien –> URLs sin error; Bien-Error –> URLs con error modificable; Ejemplo –> URLs usadas para ejemplificar (i.e.: ejemplo.com); Error –> URLs con error no entendibles; IP –> dirección IP; Rota –> URLs que no se encuentran completas

Debido a que se comprueba que los errores de transcripción disminuyen con los años, siendo el máximo 5% en 2008 para RegEx1, se decide utilizar los URLs recogidos mediante RegEx1 sin necesidad de realizar una limpieza manual de datos. En el caso de RegEx2 sí que se realiza una limpieza y preparación manual de los datos, ya que el volumen de URLs recogidas lo permite. Para ello, las URLs clasificadas como Bien – Error son reparadas manualmente, siendo agregadas a la lista final clasificadas como bien.

Una vez RegEx2 se encuentra preparado, los datos se incorporan a las URLs obtenidas anualmente mediante RegEx1.

Para almacenar los datos obtenidos se utiliza una base de datos sencilla en MySQL, así como ficheros de respaldo en XLS y R.

3.2. Análisis de enlaces incluidos en contenidos Web a Patentes

Del mismo modo que se localizan los enlaces contenidos en los documentos de patentes, se debe comprobar la visibilidad e impacto de las éstas en la web, es decir, cuantas páginas y cuáles son las que contienen citas a documentos de patentes.

De este modo se localizan aquellas que resultan de mayor interés en el medio y pueden contener enlaces relevantes, lo que indirectamente ofrece un indicador de los mismos.

La Figura 24, recogida en el §3.3, representa el modelo descrito a lo largo de la presente sección.

3.2.1. Fuentes de información

Como se explica en el apartado §2.2.1.3.2 existen diversos métodos para localizar enlaces en la web, desde métodos públicos a métodos mediante herramientas privadas.

Inicialmente para el presente estudio se plantea realizar una búsqueda mediante *Hit Count Estimate*[68] (Font-Julian et al., 2018) contra Google, para esto es necesario conocer la URL que se desea contabilizar en el motor de búsqueda.

En el apartado de análisis de la USPTO (§3.1.1.1) se menciona que la Oficina permite el acceso individual online a los documentos de las patentes. El problema que plantea este acceso es la propia URL de análisis. Al realizar una búsqueda en la base de datos de la Oficina no mantiene una URL estandarizada para cada patente, es decir, cada consulta genera una URL diferente y no sigue un esquema estándar de creación. A continuación, se puede ver una consulta realizada a la base de datos sobre la patente US10.722.783

http://patft.uspto.gov/netacgi/nph-Parser?Sect1=PTO1&Sect2=HITOFF&d=PALL&p=1&u=%2Fnetahtml%2FPTO%2Fsrchnum.htm&r=1&f=G&l=50&s1=10,722,783

[68] Tras realizar una búsqueda, el motor de búsqueda ofrece un número con la cantidad de resultados que concuerdan, la recogida de este dato permite realizar análisis de visibilidad e impacto.

En la URL se pueden observar, el dominio, diferentes métodos de extracción necesario para la consulta, y en la parte final el número de la patente. Pero si se realiza una modificación incluyendo en la parte final otro número de patente, por ejemplo, utilizando la patente US9.888.919, la página no muestra el resultado correctamente. Por lo tanto, no es posible utilizar la USPTO como fuente de información individual de cada patente.

Existen otras fuentes de datos que cuentan con las patentes a texto completo, como Google Patents o Lens.org que sí mantienen una URL estandarizada con un sistema propietario generado por sus herramientas y bases de datos, utilizando el número de patente como identificador para ofrecer la patente; y al permitir acceso a los datos y cuerpo de la patente son utilizados generalmente para referenciarlas en internet. Como muestran los datos ofrecidos por Google Trends en la Figura 23, Google Patents se muestra como la plataforma más popular, por lo que se decide centrar las búsquedas de enlaces referenciando a patentes utilizando sus URLs. El formato de las URLs en Google Patents se encuentra formado por una parte fija y una variable:

https://patents.google.com/patent/US5184830A

Parte fija Parte variable

Por lo tanto, basta cambiar el número de patente tras la última barra para tener las URLs de las patentes recopiladas.

Figura 23: Comparación de búsquedas en Google entre Portal de Patentes USPTO, Google Patentes y Lens.org en Google Trends durante 2020
Fuente: Google Trends

El método de búsqueda de Hit Count Estimate realiza una búsqueda en el motor de búsqueda Google con la siguiente consulta estándar:

"patent.google.com/patent/" -site:google.com

De este modo, los resultados ofrecen todas las páginas web en cuyo cuerpo aparezca el texto "patent.google.com/patent/" y cuyo dominio no pertenezca a "google.com". Este sistema tiene un problema crítico inicial, Google bloquea las búsquedas masivas, por lo

que tratar de realizar esta búsqueda para más de 3 millones de documentos puede resultar complejo ya que el motor bloquearía el robot cada cierto número de búsquedas.

Este problema se puede solucionar alargando el tiempo entre consultas, utilizando VPNs y realizando las búsquedas utilizando diferentes IPs. De este modo se trata de sortear las medidas de seguridad del motor de búsqueda para tratar de recopilar todos los resultados.

Uno de los problemas a los que puede afrontar la búsqueda de resultados mediante *Hit Count Estimate* es que los resultados ofrecidos sean cero debido a que no existen páginas referenciando o utilizando la URL de búsqueda (o porque el motor de búsqueda no ha registrado el enlace creado) y no se puedan realizar los estudios estadísticos necesarios para la creación de indicadores o la localización de recursos de calidad.

Dado que las mediciones a realizar son sobre una cantidad de patentes muy concreta se realiza una prueba inicial para comprobar la viabilidad de utilizar la selección de patentes del Bloque Patent Outlink para realizar un análisis circular. Para ello, se descarga de la página Lens.org las 100 patentes del periodo seleccionadas más enlazadas (por otras patentes), es decir las de mayor impacto; se extraen sus números de identificación, se generan las URLs con el formato de Google Patents indicado anteriormente. Tras esto, se realiza una búsqueda manual de cada una de las URLs generadas y se recopilan los resultados. El resultado para todas es 0.

Por lo tanto, para el presente estudio y, teniendo en cuenta que el principal foco de la investigación es la identificación de recursos de calidad, se plantea un método alternativo para la recopilación de información relacionada con el impacto y visualización de patentes en la web realizado utilizando Majestic.

3.2.1.1. Análisis de los datos ofrecidos por Majestic

Majestic, como se explica en el §2.2.1.3.2, es una herramienta de tipo buscador que recorre la web mediante bots automáticos, inspeccionando y almacenando el contenido enlazado. De este modo, se genera una base de datos de enlaces de tipo comercial, que permite a la herramienta ofrecer comparaciones e información sobre las paginas webs.

La herramienta puede ser utilizada online y mediante descarga de información (vía directa o API). Para la recopilación de datos necesarios para la presente tesis se utiliza la vía de descarga directa ya que se cuenta con el acceso Pro, lo que permite la descarga de 20 millones de unidades de crédito, necesarios para realizar las descargas (un enlace = un crédito).

La Tabla 20 recoge el glosario de lo términos a utilizar durante el transcurso de la presente tesis.

Tabla 20: Descripción de los indicadores propios de la herramienta Majestic
Fuente: elaboración propia

Término	Descripción
Trust Flow	Indicador de 0 a 100 relativo a la calidad de los enlaces que recibe un portal web
Citation Flow	Indicador de 0 a 100 relativo a la cantidad y valor de los enlaces recibidos por un portal web
External Backlinks	Enlace de referencia recibido por un portal web desde un dominio diferente.
Referring Domains	Dominio o portal web que enlaza al sitio analizado.
Referring IPs	Dirección IP desde la que se enlaza al sitio analizado
Referring Subnets	Subredes IP desde las que se enlaza al sitio analizado
Texto ancla	Texto con el que se describe el enlace recibido desde el exterior
Fresh	Datos rastreados más recientes (momento de recogida)
Historic	Datos históricos de todos los enlaces rastreados generados.

3.2.2. Proceso de recogida y extracción de datos

El día 12 de abril de 2020, entre las 16:10 y las 23:07 horas, se realiza la descarga de información desde la herramienta. Desde la misma se obtiene en bruto la siguiente información:

- → Informe general del dominio Patents.google.com: en el que se incluye información condensada relativa a los indicadores descritos en la Tabla 20
- → Datos brutos para el análisis masivo relativos a:
  - ⇢ External Backlinks Fresh
  - ⇢ Listado Subdomains Fresh
  - ⇢ Texto ancla
  - ⇢ Países de referencia
  - ⇢ Perfiles de enlazado

El motivo por el que se recopila la información de enlaces externos de tipo reciente en vez del histórico es debido a la limitación de unidades de análisis (20 millones de unidades). En el momento de realizar la recopilación de información, el volumen de datos relativo a enlaces históricos era de 23 millones, dado que se consume una unidad por enlace descargado, se excedía la capacidad de la cuenta.

3.2.3. Proceso de limpieza, preparación y almacenamiento

Para el proceso de limpieza y preparación de los datos la herramienta utilizada es Open Refine. Esto es debido a que Excel no puede abrir los 3,03 millones de filas (el máximo es 1,04 millones) y con R no se generaba correctamente la apertura del fichero también por un problema de memoria y almacenamiento.

Open Refine[69] es una herramienta acceso abierto, que permite trabajar con datos todo tipo de datos. Sus aplicaciones principales permiten:

- → Limpieza básica de datos
- → Separación o fusión de campos
- → Exportación del fichero preparado en diferentes formatos

Mediante Open Refine se abre el fichero, inicialmente la herramienta trabaja sobre un conjunto pequeño de datos sobre los que se puede realizar modificaciones, en el momento de exportar se hace sobre el conjunto completo de datos, lo que puede requerir más memoria del dispositivo (para trabajar el fichero de 3 Gbs de datos, se utilizan 42 Gbs de memoria de los 50 que tiene el equipo).

Una vez abierto, además de 3 millones de filas, el fichero contiene 72 columnas, de estas se seleccionan aquellas que ofrecen los datos relativos a la investigación. Además, permite la búsqueda y selección de subconjunto. Dado que Majestic ofrece la información relativa al dominio "Patents.google.com" y el análisis se realiza sobre las patentes americanas enlazadas, se realiza una búsqueda de todas aquellas URLs que se encuentren en la columna "Target ULR" y que contengan:

"patents.google.com/patent/US"

De este modo, se exportarán únicamente aquellas direcciones que pertenezcan al enlace a una patente de la oficina USPTO. El número de enlaces a patentes total es de 2.297.366.

Antes de exportar el fichero, se debe modificar los campos que contentan URLs, añadiendo un entrecomillado, ya que existen URLs que contienen símbolos que en el momento de exportar generan problemas.

Una vez finalizado el proceso de exportación al formato que interese (en este caso se utiliza CSV y TSV para evitar problemas de apertura en otros programas) se pasa a utilizar R para la lectura de información (Excel todavía no es viable ya que el fichero contiene más de 2 millones de registros). El fichero CSV se utilizará, además, para cumplimentar una base de datos en formato MySQL para almacenar la información y facilitar el acceso en el futuro.

[69] https://openrefine.org/

3.3. Method's Summary*

Pese a que el proceso descrito se encuentra dividido en dos grandes bloques de análisis, puede ser llevado a cabo como un proceso completo para la localización de recursos de información. La Figura 24 muestra un resumen esquematizado de todo el proceso descrito en el presente capitulo, junto con el volumen de datos de cada etapa.

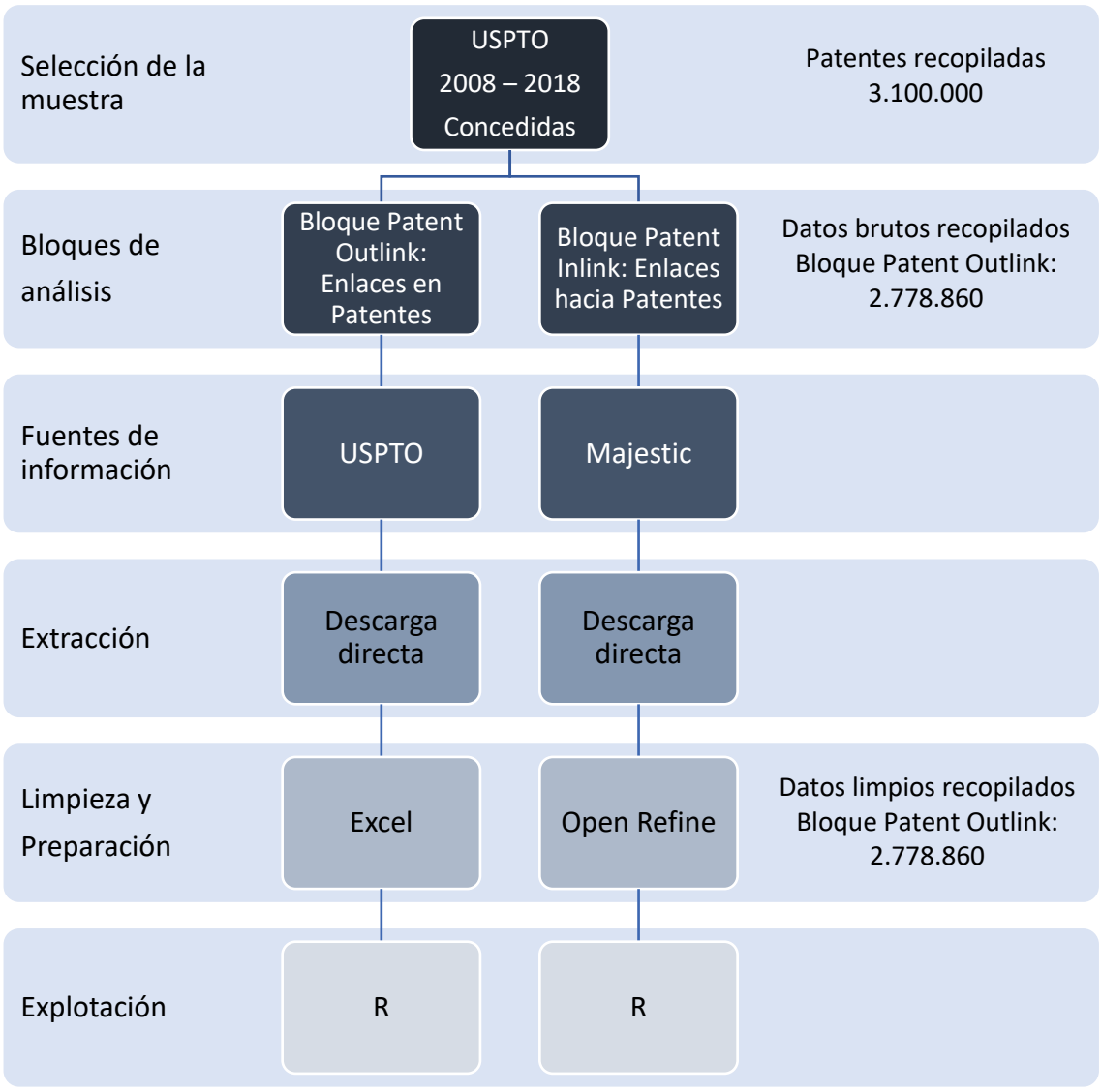


Figura 24: Esquematización del método utilizado para la recopilación, preparación y explotación de los resultados
Fuente: elaboración propia

Capítulo 4
Resultados

El presente capítulo recoge los resultados extraídos utilizando el método explicado en el Capítulo 3.

El Capítulo se divide, tal como se ha explicado anteriormente, en dos grandes bloques:

- → Bloque Patent Outlink: Análisis de enlaces de patentes a recursos web
- → Bloque Patent Inlink: Análisis de enlaces de recursos web a patentes

El Bloque Patent Outlink se caracteriza por analizar el corpus de patentes concedidas por la Oficina de Patentes y Marcas de Estados Unidos (USPTO) durante los años 2008 a 2018. Con esto se obtienen 3.133.247 patentes para su análisis que, tras aplicar el método detallado en el capítulo anterior, contienen 2.745.973 enlaces web. Sobre estos enlaces, se procede a realizar un análisis estadístico del primer nivel (dominio, i.e.: url.com) de los enlaces.

El Bloque Patent Inlink se encuentra formado por todas las páginas web recopiladas utilizando la herramienta Majestic, que contienen un enlace web a una página de Google Patents referenciando a una patente concedida por la USPTO. Se encuentran recogidos 2.297.366 enlaces que serán analizados estadísticamente.

4.1. Bloque Patent Outlink: Análisis de enlaces de patentes a recursos web

4.1.1. Número y evolución de las patentes concedidas

Durante los años 2008 a 2018, en Estados Unidos se conceden en promedio 284.840 patentes anualmente. La Figura 25 muestra gráficamente la evolución del número de patentes concedidas durante los años indicados por la USPTO. Como se puede observar, la tendencia es al alza pese a existir pequeños descensos, tal y como sucede con la tendencia global vista en el §2.1.5.

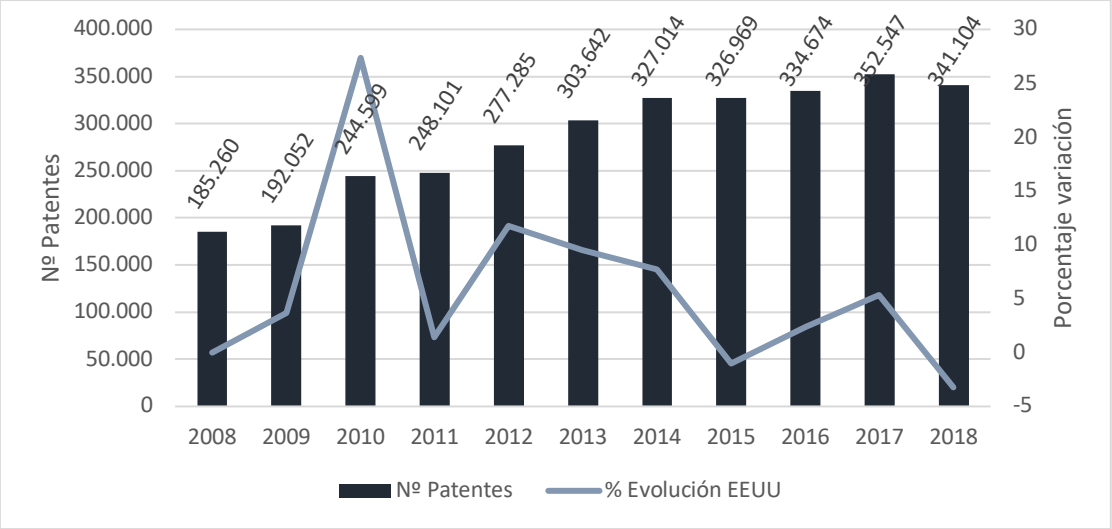


Figura 25: Gráfica de la evolución de concesión de patentes en EE. UU. entre 2008 y 2018

Como se puede observar, el porcentaje de variación anual muestra una tendencia de evolución al alza en la concesión de patentes en Estados Unidos. Este incremento es más acusado durante los primeros años, observándose cierta estabilización en los últimos 4 años, reflejado en los porcentajes de variación anual más bajos.

Esta estabilización es visible, también, aunque en menor medida, en los datos globales, ya que anualmente desde 2013 el porcentaje de patentes concedidas a nivel mundial, aunque continúa incrementando, lo hace en un porcentaje menor.

Esta estabilización se da tanto en la USPTO como otras de las Oficinas del mundo (i.e.: China o Japón) y no implica una saturación inventiva, ya que las solicitudes de patentes si se encuentran en aumento durante el mismo periodo, la diferencia apunta a una búsqueda de mayor calidad en los sistemas de concesión, una gran cantidad de solicitudes que deben revisarse, entre otros según indica la OMPI[70].

[70] https://www.wipo.int/edocs/pubdocs/en/wipo_pub_941_2019.pdf

4.1.2. Análisis descriptivo de los enlaces recogidos

Tal y como se indica en el capítulo anterior, la cantidad total de enlaces recogidos en bruto (sin haber sido limpiados y arreglados) utilizando ambas fórmulas es de **2.745.973**. La Tabla 21 muestra todos los enlaces brutos recogidos anualmente, así como los resultados totales.

Del mismo modo que existe una tendencia al alza en la cantidad de patentes concedidas, año tras año aumenta la cantidad de enlaces que los inventores utilizan para contextualizar y aportar valor en sus textos.

Tabla 21: Enlaces recogidos anualmente para las patentes analizadas según fórmula y total
Fuente: elaboración propia

	Nº Patentes	Bruto RegEx1	Bruto RegEx2	Total Enlaces	Enlaces normalizados
2008	185.260	87.856	1.745	89.601	0,48
2009	192.052	113.307	2.250	115.557	0,60
2010	244.599	169.027	3.358	172.385	0,70
2011	248.101	184.244	3.738	187.982	0,76
2012	277.285	215.477	4.562	220.039	0,79
2013	303.642	268.682	6.106	274.788	0,90
2014	327.014	301.394	6.761	308.155	0,94
2015	326.969	291.437	7.042	298.479	0,91
2016	334.674	318.628	7.558	326.186	0,97
2017	352.547	367.762	8.440	376.202	1,07
2018	341.104	363.500	7.760	371.260	1,09
	3.133.247	**2.681.314**	**59.320**	**2.740.634**	0,87

La cantidad de enlaces incluida en documentos de patente pasa de ser 89.601 a 371.260, lo que supone un incremento del 414% en 10 años, cuando el incremento en patentes es únicamente del 84% en el mismo periodo.

En 2008 existe un 48,37% de enlaces con respecto a la cantidad de patentes contenedora, esto significa que, de media, se puede encontrar menos de un enlace cada dos patentes. Con el paso del tiempo esta tendencia cambia, incrementando en 10 puntos porcentuales con respecto al año anterior hasta 2013, donde se alcanza el 90% de enlaces con respecto a patentes. Es en 2017 cuando se supera el 100% (exactamente se alcanza un 106,71%), lo que implica que, de media, todas las patentes contienen un enlace.

La Figura 26 muestra la evolución comparativa del número de patentes frente al número de enlaces contenidos. Se puede observar claramente la tendencia al alza de ambos valores, llegando a superar los enlaces a las patentes en los años 2017 y 2018. Se ha incluido el porcentaje de variación (representado mediante una línea naranja) para optimizar la visualización de los datos de evolución.

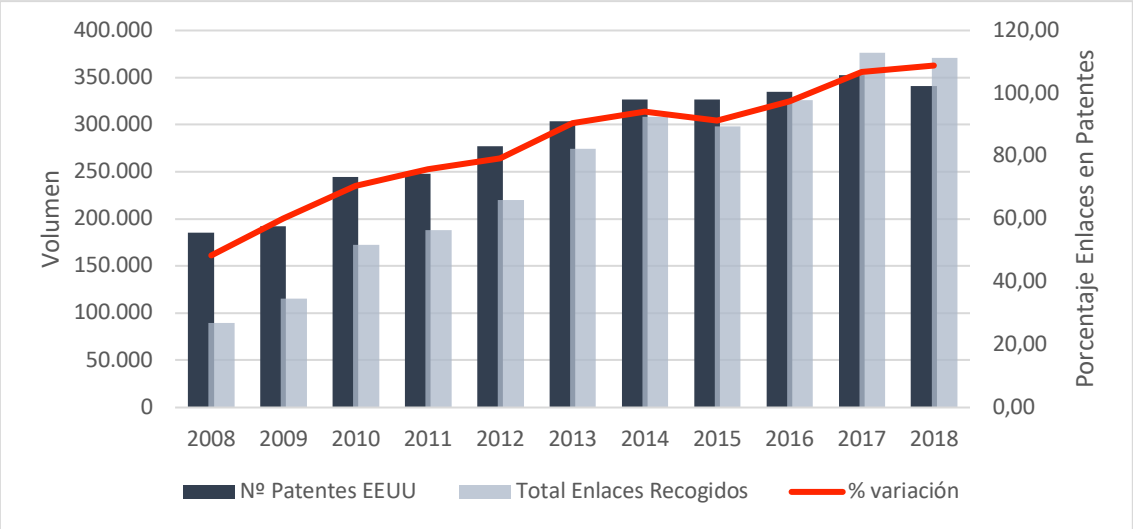


Figura 26: Cantidad de enlaces total recogidos anualmente junto con el número de patentes desde los que se extraen
Fuente: elaboración propia

La Figura 27 permite apreciar como, tal y como se indica con anterioridad, pese a que el porcentaje de enlaces recopilados con respecto al total de patentes es alto incluso desde el primer año analizado (48,37%), la cantidad de patentes que contienen al menos un enlace es, en realidad, baja ya que se mantiene por debajo del 20% durante todos los años del análisis, siendo el mínimo 24.624 (2009) y el máximo 71.469 (2017). Aunque se encuentra en continuo aumento, teniendo en cuenta que el porcentaje de patentes es, a su vez, mayor cada año.

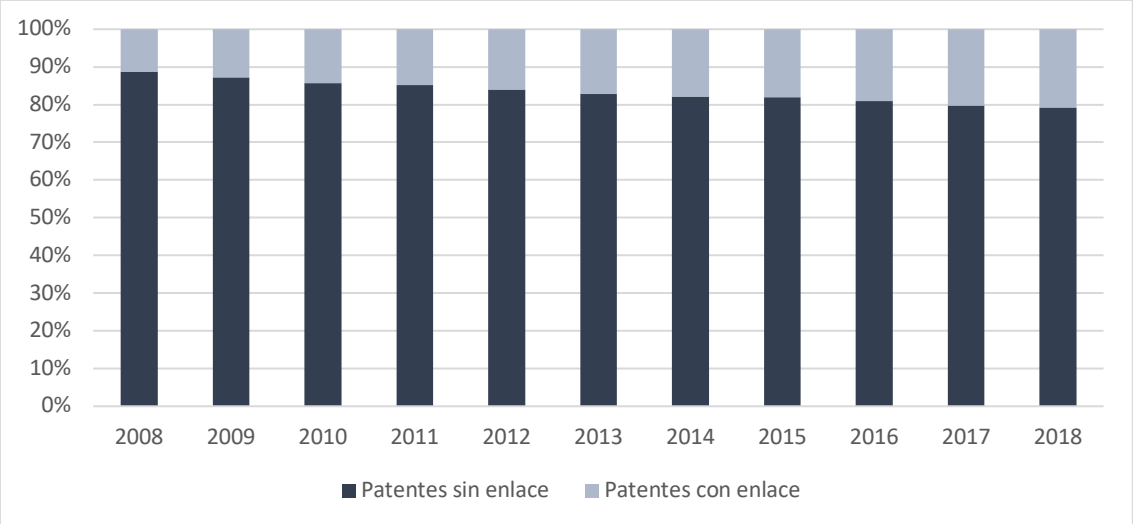


Figura 27: Representación gráfica mediante porcentaje de patentes con y sin enlaces
Fuente: elaboración propia

La Tabla 22 recopila un análisis descriptivo de los datos. En ella se puede observar cómo respecto a la cantidad mínima de enlaces recogidos por patente en todos los años estudiados es 1, existiendo 190.879 patentes que contienen un único enlace. En cambio, la cantidad máxima varía entre 234 (2008) y 1.554 (2015). Exceptuando 2015 y 2016 (donde existen 1.254 enlaces en una patente) en todos los años la cantidad máxima de enlaces recogidos en una misma patente se encuentra por debajo de 1.000. Respecto a

los máximos, resulta anecdótico (debido a la cantidad) que en 2012 existen dos patentes con 500 enlaces. La Figura 28 muestra la agrupación de patentes por cantidad de enlaces para todos los dominios; como se puede observar, el grupo que más enlaces genera es aquel en el que las patentes contienen entre 1 y 10 enlaces (227.562 en total), seguido del grupo entre 11 y 100 (27.597). Como se puede observar, únicamente existen 9 dominios que reciben entre 10.001 y 20.000 enlaces y únicamente 6 con más de 20.000 enlaces.

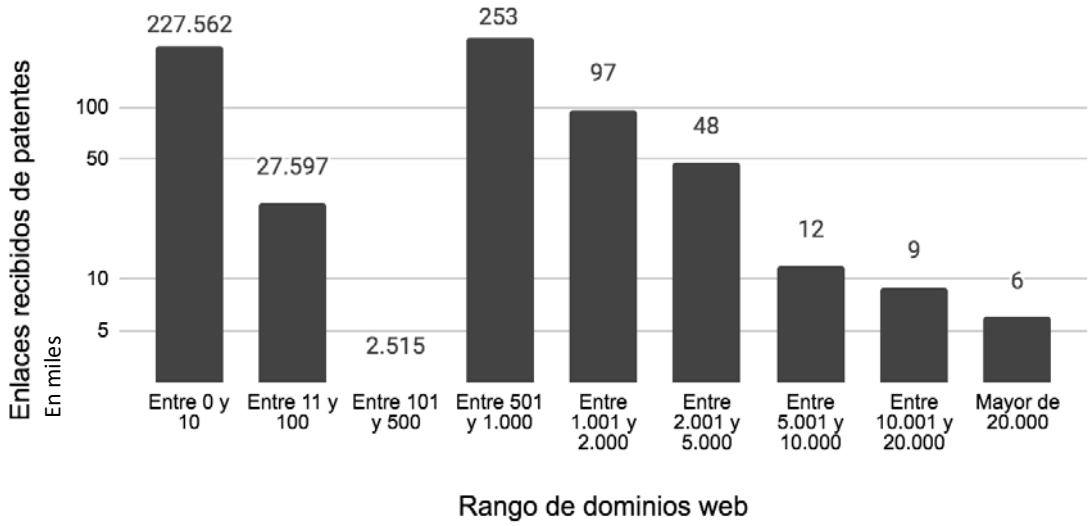


Figura 28: Cantidad de enlaces salientes de patentes hacia dominios web
Fuente: elaboración propia

La media de enlaces por patente se encuentra entre 4,26 de mínimo en 2008, y 5,251 de máxima (en 2013). Pese a que el máximo se da en 2013, los años posteriores se encuentran todos en el rango 5,027 – 5,233, por lo que la diferencia es muy pequeña y se puede decir que todos se encuentran ligeramente por encima de los 5 puntos. Por lo tanto, de media las patentes contienen entre 4 y 5 enlaces para los años estudiados.

Atendiendo al análisis de la mediana, en cambio, todos los años analizados ofrecen como resultado dos enlaces por patente. Respecto a los datos por cuartiles, se puede observar como para todos los años el primer cuartil se encuentra 1 enlace por patente, y para el tercero, exceptuando 2008 y 2009, todos obtienen 5 enlaces por patente.

El análisis de la varianza y la desviación típica muestran que la dispersión de los datos con respecto a la media varía entre 7,689 de mínimo y 13,663 de máximo, lo que implica que la cantidad de enlaces contenidos en patentes se encuentra bastante dispersa con respecto a la media, encontrándose una cantidad significativa diferente de enlaces por patente.

Tabla 22: Datos descriptivos sobre la cantidad limpia de enlaces recogidos por año. Fuente: elaboración propia

Año	Total Patentes	Patentes c/enlace	%Pat. c/enlace	Total Enlaces	Mín	Máx	Frec. mín	Frec. máx	1° Cuartil	Mediana	3° Cuartil	Media	Varianza (n-1)	Desviación típica (n-1)
2008	185.260	20.837	15,57%	88.858	**1**	234	8.336 29,91%	1	1	2	4	4,264	59,123	7,689
2009	192.052	24.624	12,82%	114.569	**1**	722	9.460 38,42%	1	1	2	4	4,653	98,892	9,944
2010	244.599	34.824	14,24%	170.958	**1**	501	13.234 38,00%	1	1	2	5	4,909	117,943	10,86
2011	248.101	36.900	14,87%	186.480	**1**	717	13.839 37,50%	1	1	2	5	5,054	125,737	11,213
2012	277.285	44.212	15,94%	218.437	**1**	500	16.342 36,96%	2	1	2	5	4,941	125,12	11,186
2013	303.642	51.913	17,10%	272.597	**1**	713	19.170 36,93%	1	1	2	5	5,251	161,92	12,725
2014	327.014	58.451	17,87%	305.854	**1**	850	21.226 36,31%	1	1	2	5	5,233	169,66	13,025
2015	326.969	58.846	18,00%	295.829	**1**	1.554	21.685 36,85%	1	1	2	5	5,027	165,315	12,857
2016	334.674	63.947	19,11%	323.802	**1**	1.254	23.108 36,14%	1	1	2	5	5,064	186,685	13,663
2017	352.547	71.469	20,27%	373.435	**1**	394	25.491 35,67%	1	1	2	5	5,225	144,171	12,007
2018	341.104	70.541	20,68%	368.886	**1**	403	24.988 35,42%	1	1	2	5	5,229	146,701	12,112

4.1.3. Número de enlaces únicos

Como se ha visto anteriormente, la cantidad de enlaces recogidos para cada uno de los años aumenta; del mismo modo, como se puede ver en la Tabla 23, la cantidad de enlaces únicos también lo hace, pasando de existir 25.060 enlaces únicos en 2008 a 67.689 enlaces únicos en 2018.

Este incremento del 270% existente entre los dos años, debe analizarse teniendo en cuenta que existe un incremento del 415% de enlaces recogidos. Esto implica que se añaden enlaces de fuentes ya conocidas, siendo el 28,20% en 2008 frente al 18,35% en 2018.

Tabla 23: Relación enlaces recogidos por año, totales y únicos
Fuente: elaboración propia

Año	Enlaces totales	Enlaces únicos	Porcentaje enlaces únicos
2008	88.858	25.060	28,20%
2009	114.569	29.834	26,04%
2010	170.958	39.659	23,20%
2011	186.480	43.739	23,46%
2012	218.437	49.951	22,87%
2013	272.597	57.000	20,91%
2014	305.854	61.787	20,20%
2015	295.829	60.900	20,59%
2016	323.802	64.446	19,90%
2017	373.435	69.913	18,72%
2018	368.886	67.689	18,35%
Total	**2.719.705**	**569.978**	

En la Figura 29 se puede ver la evolución de la cantidad de enlaces únicos frente a los enlaces totales por año recogidos.

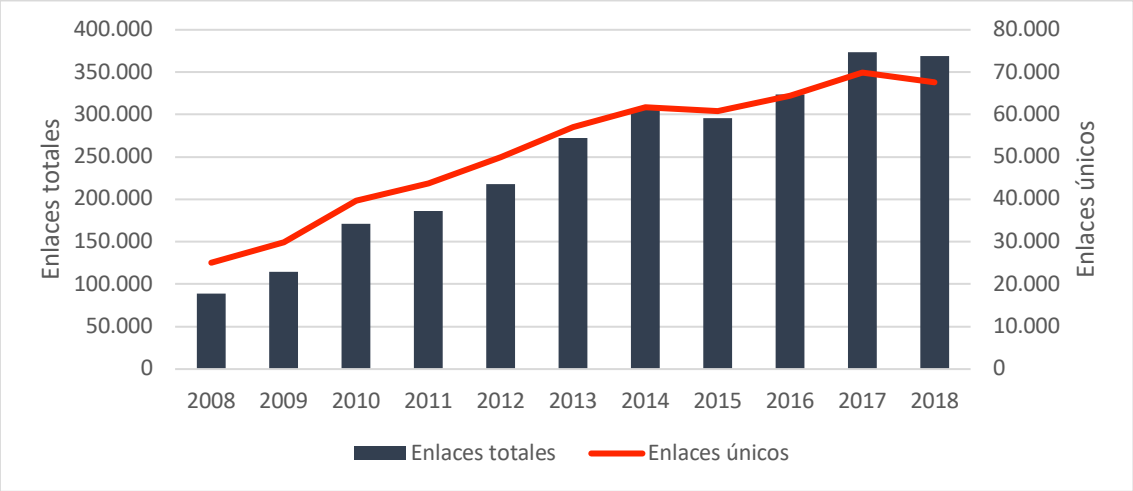


Figura 29: Enlaces totales frentes a enlaces únicos recogidos por año (2008-2018)
Fuente: elaboración propia

4.1.4. Análisis TLD

En primer lugar, se realiza el análisis de los Top Level Domains (TLD) relativos a los enlaces recopilados.

Tras realizar un proceso de limpieza de datos, se obtiene un total de 256.721 dominios únicos recogidos en 2.707.636 URLs.

Respecto a los TLD que conforman esos dominios, se pasa a realizar dos tipos de análisis complementarios. El primero tiene por objeto un análisis de los TLD de primer nivel (i.e.: .com, .es, .org, …), el segundo se realiza teniendo en cuenta los TLD de segundo nivel o SLD (i.e.: .co.uk, .ac.jp, .edu.au, …). La diferencia existente entre los dos tipos de análisis radica en la cantidad de TLDs a revisar. La cantidad de TLDs recogidos para el primer nivel es de 231, en cambio, teniendo en cuenta el segundo nivel se pasa a tener 2.759.

Con respecto al análisis de TLDs de primer nivel, la Tabla 24 muestra ordenados en orden descendente todos los TLDs recogidos. Se puede observar como destaca claramente el TLD .com siendo utilizado en 1.659.025 enlaces recogidos, lo que supone el 61% del total. Teniendo en cuenta que, tal y como se indica en el §2.2.1.1, .com es el TLD más utilizado en toda la web (52% de las webs lo utilizan) el resultado no es inesperado.

Los resultados muestran al TLD .org como el segundo más enlazado desde patentes (492.664; 18,13%), seguido por .edu (138.052; 5,08%) y .gov (106.186; 3,90%). Debido a la distribución desigual de la información, se puede apreciar el rápido descenso en los porcentajes de uso, con un tercer resultado que no alcanza el 4%. Estos tres TLDs destacan por pertenecer a páginas centradas en contenido de Organizaciones no gubernamentales, Educación y Gobiernos.

En la siguiente franja de análisis, aquella que recibe resultados por encima de 10.000, se pueden encontrar seis TLD .net: (2.90%), .uk (1,63%), .jp (0,91%), .ca (0,51%), .de (0,45%), y .au (0,42%). En este caso, los TLD representan –a excepción de .net– países; contando, además, con un alto grado de similitud con la representación en el porcentaje total utilizado en internet.

Tras estos TLD, se encuentra la franja con resultados por encima de 1.000. En ella se puede encontrar un total de 39 TLDs, con una representación entre el 0,03% y el 0,34%. 36 de los dominios son de tipo ccTLD (country code TLD), que representan a países o zonas territoriales. Los 3 TLDs no pertenecientes a países son .info (información), .int (organizaciones, oficinas y programas internacionales) y .biz (empresas), con una representación online superior a la recogida en las patentes.

Tabla 24: Cantidad de TLDs de primer nivel recogida
Fuente: elaboración propia

TLD	Nº Enlaces	TLD	Nº Enlaces	TLD	Nº Enlaces	TLD	Nº Enlaces	TLD	Nº Enlaces
com	1.659.025	sg	1.167	im	91	cr	6	vg	1
org	492.664	hr	1.071	az	91	sh	5	vc	1
edu	138.052	me	1.060	cx	85	ps	5	tt	1
gov	106.186	hu	952	online	80	lv	5	systems	1
net	78.838	tr	756	lv	78	ibm	5	sy	1
uk	44.351	xyz	698	is	77	global	5	style	1
jp	24.818	my	657	live	58	tl	4	studio	1
ca	14.068	ee	519	nu	54	ong	4	ss	1
de	12.443	mx	514	ge	53	mc	4	so	1
au	11.612	ua	499	bs	46	life	4	science	1
fr	9.319	rs	494	am	43	apple	4	sca	1
ch	7.991	store	444	ms	40	yahoo	3	reviews	1
nl	6.859	si	425	md	34	top	3	review	1
us	6.376	pk	407	kz	34	sm	3	press	1
ng	5.344	cc	405	technology	33	page	3	pink	1
info	5.287	ar	395	lu	29	ny	3	parts	1
int	5.197	br	382	mp	28	link	3	one	1
eu	5.074	ro	376	ve	26	il	3	om	1
id	4.570	sk	348	asia	21	dev	3	nf	1
be	3.910	sa	348	shop	17	dell	3	na	1
io	3.621	ae	344	by	17	ci	3	mu	1
co	3.585	ir	340	yu	16	bank	3	mn	1
no	3.486	ly	328	ni	16	tz	2	mk	1
cn	3.132	th	317	mom	15	tn	2	mit	1
dk	3.032	ai	315	et	15	tc	2	microsoft	1
kr	2.743	su	311	li	12	red	2	media	1
pl	2.739	pro	286	cy	12	re	2	je	1
ru	2.520	bg	262	as	11	pe	2	ink	1
tw	2.466	fm	257	tech	10	nyc	2	ht	1
il	2.269	site	255	ne	10	ninja	2	health	1
nz	2.211	ws	225	education	10	news	2	guru	1
se	2.210	ph	212	cm	10	museum	2	gi	1
biz	2.148	lt	199	ag	10	mg	2	fo	1
it	1.951	mobi	184	today	9	kg	2	fk	1
es	1.834	vn	161	test	9	help	2	fishing	1
tv	1.822	name	158	ba	9	guide	2	docs	1
cz	1.569	gl	156	space	8	gs	2	do	1
mil	1.565	lk	151	sc	8	final	2	digital	1
gr	1.554	to	139	la	8	energy	2	cu	1
hk	1.477	cl	131	dz	8	eh	2	clothing	1
za	1465	tk	129	ac	8	cern	2	cg	1
pt	1430	cf	119	st	7	bt	2	cfd	1
in	1418	uz	102	mo	7	blue	2	cba	1
at	1386	ma	101	google	7	al	2	cat	1
fi	1361	bd	99	gg	7	zone	1	bike	1
ie	1318	watch	94	bz	7	xin	1	alibaba	1

Tras los TLDs destacados anteriormente y que representan el 99,42% de todos los TLDs recogidos, se encuentra una larga cola de TLDs que contiene un total de 15.454 enlaces recogidos (0,58%).

Para mostrar gráficamente la diferencia que existe se incluye la Figura 30, en la que representan los TLDs con más de 1.000 enlaces recogidos. Ha sido necesario representar el eje Y mediante escala logarítmica debido a la gran diferencia existente entre el uso de los TLDs.

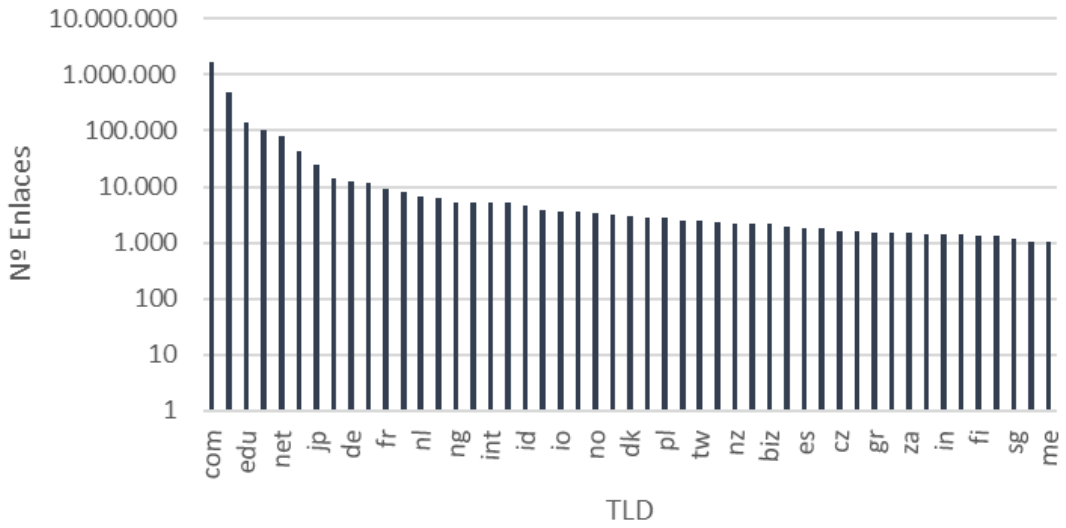


Figura 30: Representación TLDs primer nivel con más de 1.000 enlaces recogidos
Fuente: elaboración propia

Para poder comprobar la agrupación de enlaces por TLD, se incluye la Figura 31, en la que se realiza un sumatorio por cantidad de enlaces recopilados de los TLDs. En este caso, de nuevo es necesario representar el eje Y en modo logarítmico.

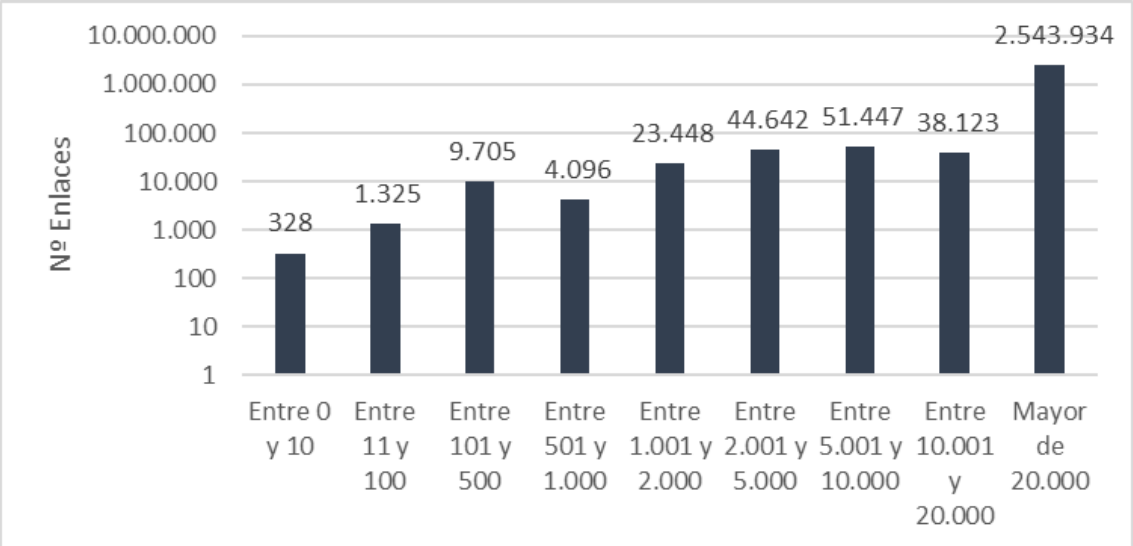


Figura 31: Recopilación por cantidad de TLD
Fuente: elaboración propia

Análisis TLDs de segundo nivel (SLDs)

Debido a que un SLD es un ccTLD junto con un segundo nivel para indicar el tipo de entidad a la que pertenece el dominio (ie.e: Cambridge.ac.uk – Dominio: Cambridge; SLD: ac, academia; TLD: uk, Reino Unido). En el apartado anterior se indica que existen 2.759 SLDs, para su análisis se ha desgranado cada uno de ellos y agrupado para conocer los países que utilizan esta nomenclatura. De este modo, se obtienen 213 ccTLDs y 2.546 de segundo nivel.

En el listado de TLDs de primer nivel se mantienen los resultados similares a lo analizado en el apartado anterior, teniendo en cuenta que las cifras para aquellos TLD que utilicen un segundo nivel quedarán desglosados, por ejemplo es el caso de .uk que desaparece de los primeros puestos ya que pasa a dividirse por cada una de las opciones que utiliza (i.e.: .ac.uk, .co.uk, etc.).

Respecto al segundo nivel, el listado por países indica que únicamente se utilizan SLDs pertenecientes a 12 países. La Tabla 25 muestra estos países y la cantidad total de enlaces recogidos para cada TLD.

Tabla 25: Representación de países en TLDs de segundo nivel con total de enlaces
Fuente: elaboración propia

TLD	País	Total Enlaces
uk	Reino Unido	44.351
jp	Japón	24.818
au	Australia	11.612
tw	Taiwán	2.466
il	Israel	2.269
kr	Ucrania	2.743
nz	Nueva Zelanda	2.211
za	Zambia	1.465
cn	China	3.132
sg	Singapur	1.167
us	Estados Unidos	6.376
at	Austria	644
be	Bélgica	519
my	Malasia	657
ar	Argentina	395
vn	Vietnam	161
mx	Méjico	126
ph	Filipinas	212
br	Brasil	56
ge	Georgia	53
cr	Costa Rica	6

En cambio, de los 2.546 SLDs recogidos existen 2.314 únicos. De entre todos los SLDs recogidos, el 80% aparecen menos de 10 veces (1.100 aparecen en una única ocasión y 955 menos de 10 veces), esto muestra una gran variedad y diversificación en los segundos niveles (teniendo presente los posibles errores de toma de datos que puedan existir estimados en un 5%).

La Tabla 26 muestra los enlaces recogidos para los 10 primeros SLDs indicando el tipo de dominio al que pertenecen.

Tabla 26: 10 primeros resultados SLDs con total enlaces recogidos
Fuente: elaboración propia

SLD	Tipo	Total enlaces
co	Comercial	36.458
ac	Academia	26.477
com	Comercial	9.046
edu	Educación	6.035
go	Gobierno	4.648
org	Organización NG	3.279
or	Organización	1.630
gov	Gobierno	1.347
ne	Network	603
dni	Institutos Nacionales Distribuidos (EE. UU.)	433
net	Network	385

Los 10 SLDs mostrados en la tabla representan el 86% de los SLDs recogidos. Debido a las diferentes nomenclaturas existentes por TLD, existe duplicidad en el tipo de contenido. Por ejemplo, .co y com representan empresas comerciales, pero el primero se utiliza habitualmente en países como Reino Unido, mientras que el segundo es más utilizado en países asiáticos.

Debido a esta duplicidad, los tipos de SLDs se reducen a Comercial, Educación, Gobierno, Organizaciones y Redes. Este tipo de páginas webs se espera que sean de contenido técnico, empresarial o legislativo, por lo que el análisis de los SLDs en este caso, y los TLDs como muestra el apartado anterior, indica que las patentes enlazan a contenido de alto nivel.

4.1.5. Análisis por dominios

El total de enlaces recogidos es de 2.707.696, aglutinados en 256.724 dominios únicos (asumiendo un porcentaje de error inferior al 5% en la limpieza). Para realizar un primer análisis por dominio se ha tenido en cuenta la cantidad de subdominios que existen en los enlaces recogidos (i.e.: Wikipedia.org o es.wikipedia.org).

Diferenciando los primeros cuatro niveles de subdominios, la Tabla 27 muestra los resultados totales para cada uno de ellos, teniendo en cuenta los dominios únicos y la cantidad de enlaces que representan y preparando un análisis descriptivo de los datos.

Tabla 27: Datos descriptivos para los enlaces recogidos y dominios únicos por nivel de subdominio
Fuente: elaboración propia

	Nivel 1	Nivel 2	Nivel 3	Nivel 4
Dominios Únicos	184.598	56.711	14.058	1.334
Nº Enlaces	1.858.111	703.795	137.134	8.596
Mínimo	1	1	1	1
Máximo	33.846	86.422	30.775	1.725
Frec. del mínimo	80.295	26.494	6.710	745
Frec. del máximo	1	1	1	1
1° Cuartil	1	1	1	1
Mediana	2	2	2	1
3° Cuartil	4	4	4	3
Media	10,06	12,41	9,75	6,44
Varianza (n-1)	24.314,27	245.162,23	75.383,51	3.523,02
Desviación típica (n-1)	155,930	495,139	274,561	59,355

Se puede observar como el Nivel 1 recoge el grueso de los enlaces con 1.858.598, recopilados en 184.598 dominios únicos. Esto supone un 69% del total de enlaces recogidos y el 72% de los dominios únicos totales. El Nivel 2 acumula el 26% del total de enlaces y el 22% del total de dominios únicos. En cambio, el Nivel 3 y 4 representan el 5% y menos del 1% respectivamente.

Es normal esta diferencia de porcentajes, ya que las URLs con múltiples subdominios son menos frecuentes que aquellas que no tienen o tienen uno. Resulta remarcable como para el Nivel 2 y el Nivel 3, la cantidad de dominios recogidos con un único enlace es prácticamente igual o mayor al Nivel 1, 86.422 y 30.775 respectivamente.

Atendiendo a la cantidad de resultados por dominio, para todos los niveles se obtiene un porcentaje alto (entre 43% y 56%) de dominios únicos con un único enlace. Esto implica que existe una larga cola de dominios con pocos enlaces. Se puede observar como para todos los niveles, excepto el cuarto con un uno, la mediana se sitúa en dos enlaces por dominio, fortaleciendo el dato anterior.

4.1.5.1. Análisis de dominios totales

Para comprobar la cantidad de dominios totales, se realiza la búsqueda de los dominios en los subsiguientes niveles. La Tabla 28 muestra los resultados combinados de la existencia o ausencia en cada uno de los niveles de los anteriores. En las celdas de color verde se puede observar aquellos dominios que se localizan entre el cómputo total de los niveles siguientes. En cambio, las celdas de color naranja indican aquellos dominios únicos que no se encuentran en el resto de los niveles.

Tabla 28: Resultados relativos a los dominios existentes en nivel superior
Fuente: elaboración propia

	Nivel 1	Nivel 2	Nivel 3	Nivel 4
Nivel 1	-	10.588	1.493	290
Nivel 2	174.010	-	2.164	223
Nivel 3	183.105	54.547	-	189
Nivel 4	184.308	56.488	13.869	-

Leyenda: - Color verde: dominio existe en nivel superior. – Color naranja: dominio no se encuentra en el nivel superior

Como se puede observar, para el Nivel 1 existen 174.010 dominios que no tienen representación en el Nivel 2 y 10.588 dominios que sí se encuentran en el Nivel 2. Del mismo modo, para el Nivel 2, 54.547 dominios no se encuentran en el Nivel 3, pero 2.164 si que tienen representación.

La cantidad de dominios únicos teniendo en cuenta los subdominios en el total de dominios recopilados no es elevada, ya que como máximo suponen el 5,73% contenidos en el Nivel 2. En cambio, atendiendo a la cantidad de enlaces que acumulan, suponen el 94,68%. El total para todos los niveles es de 14.947 de dominios localizados en el resto de los niveles. La cantidad de enlaces recopilados para dichos dominios es de 2.664.331 (98,40% del total de enlaces recopilados en el análisis).

La Tabla 29 recoge los primero 18 dominios (aquellos con más de 10.000 enlaces) junto con los enlaces recopilados para cada uno de los niveles, así como el sumatorio total. Se puede observar la discrepancia que existe para algunos de los dominios en cuanto a resultados obtenidos entre los niveles.

Por ejemplo, en el caso de archive.org se recopilan 97.150 enlaces entre todos los niveles, en cambio el primer nivel únicamente tiene una representación del 10%.

Se ha resaltado mediante un fondo naranja la celda con el valor máximo de la fila, para que se pueda apreciar claramente en qué nivel se encuentra el mayor número de enlaces para cada uno de los dominios. Encontrándose 7 valores máximos en el primer nivel, 8 en el segundo y 2 en el tercer nivel.

Tabla 29: Primeros 18 resultados ordenados según total de enlaces recopilados
Fuente: elaboración propia

	Numero de enlaces				
Dominio	**Nivel 1**	**Nivel 2**	**Nivel 3**	**Nivel 4**	**Total**
archive.org	9.692	87.177	246	35	97.150
Wikipedia.org	5.306	76.700	179	6	82.191
nih.gov	447	5.397	35.886	3.102	44.832
microsoft.com	10.312	28.199	772	20	39.303
amazon.com	33.846	1.036	143	1	35.026
youtube.com	23.317	73	7	24	23.421
google.com	10.781	12.032	15	16	22.844
ieee.org	642	21.881	69	0	22.592
gsmarena.com	20.435	42	0	0	20.477
w3.org	18.500	475	0	0	18.975
ietf.org	10.901	7.928	125	0	18.954
ip.com	17.569	258	0	0	17.827
ibm.com	5.309	8.757	3.569	96	17.731
ClinicalTrials.gov	15.435	1	0	0	15.436
psu.edu	155	688	13.257	13	14.113
acm.org	1.396	12.712	1	0	14.109
3gpp.org	9.709	487	6	7	10.209
yahoo.com	3.644	5.815	558	61	10.078

4.1.5.2. Categorización de enlaces recopilados

Utilizando el listado generado tras realizar la suma por dominio de todos los niveles. Se realiza una categorización doble para todos los dominios que obtienen más de 1.000 enlaces. De este modo se analizan 201 dominios, que aglutinan 956.530 enlaces recopilados (35,32% del total).

La primera clasificación se realiza en función del tipo de organismo al que pertenece el portal, pudiendo ser: Empresa, Servicio, Organización, Gobierno, Media, Universidad. La Tabla 30 muestra los enlaces recogidos para cada una de las categorías generas, así como la cantidad de dominios contenidos en las mismas.

Tabla 30: Primera categoría para los dominios recogidos de Nivel 1
Fuente: elaboración propia

Tipo de organismo	Cantidad de Dominios	Número de Enlaces
Empresa	47	260.890
Servicio	48	219.347
Organización	20	209.548
Universidad	35	92.095
Gobierno	12	87.901
Media	39	86.749

Aparecen dos grupos, el primero que representa el grueso de los enlaces recogidos (72% del total), contiene los dominios de tipo Empresa, Servicio y Organización.

El segundo grupo se encuentra formado por los dominios clasificados como Universidad, Gobierno y Media.

La segunda clasificación se centra en el tipo de información que ofrece el portal, pudiendo ser: Finanzas, Automoción, Gobierno, Noticias, Educación, Tecnología, Motor de búsqueda, Salud y Medicina, Social Media y Compras. La Tabla 31 muestra los resultados obtenidos.

Tabla 31: Categorización por tipo de contenido con total de dominios y número de enlaces recogidos
Fuente: elaboración propia

Tipo de contenidos	Cantidad de Dominios	Número de enlaces	Porcentaje
Tecnología	70	410.519	43%
Educación	73	282.108	30%
Salud y Medicina	14	83.156	9%
Compras	5	46.799	5%
Motores de Búsqueda	6	42.263	4%
Social Media	5	32.072	3%
Noticias	15	29.019	3%
Gobierno	9	19.274	2%
Sitios Personales	1	6.453	1%
Finanzas	2	2.923	0%
Automoción	1	1.944	0%

Destaca el contenido de tipo Tecnológico, representado el 43% de los enlaces recopilados, presentes en 70 de los 201 dominios (34,82%), junto con Educación (29% del total de enlaces).

Los siguientes dominios se encuentran divididos entre contenido de tipo Salud y Medicina, Compras, Motores de búsqueda, Medios Sociales, Noticias y Gobiernos, representando el 26% de los enlaces recopilados. Por último, se encuentran páginas con contenido de tipo Página Personal, Finanzas y Automoción, que suman 4 dominios con un total de 11.320 enlaces (2%).

Como anécdota, se indican los dominios recopilados para Sitios Personales (Blogspot.com) y Automoción (netcarshow.com).

4.1.6. Número de enlaces por sección

Como se indica en el apartado de metodología, cada enlace extraído desde los ficheros XML de patentes se guarda junto con información de la patente y la sección en la que aparece, pudiendo ser ésta:

- → <othercit>: Otras citas
- → <abstract>: Resumen
- → <description>: Descripción
- → <claims>: Reivindicaciones

La Tabla 32 recoge los enlaces extraídos para cada una de las etiquetas por año, así como una columna con los totales ya vistos anteriormente. Se puede observar como la mayor parte de los enlaces se recogen en la sección de citas, incrementando anualmente y superando para todos los años el 78,67% (siendo el mínimo en 2008), alcanzando el máximo en 2018 con un 85,14%.

Tabla 32: Enlaces recogidos anualmente en cada una de las secciones de un documento de patente
Fuente: elaboración propia

	Othercit	Description	Claims	Abstract	Total
2008	69.908	18.915	32	3	88.858
2009	87.956	26.594	14	5	114.569
2010	134.918	36.018	19	3	170.958
2011	147.228	39.212	30	10	186.480
2012	174.934	43.476	5	22	218.437
2013	222.390	50.159	35	13	272.597
2014	248.676	57.143	13	22	305.854
2015	241.838	53.963	18	10	295.829
2016	266.230	57.481	90	1	323.802
2017	315.339	58.069	25	2	373.435
2018	314.053	54.798	32	3	368.886

Descripción es la segunda sección en la que más enlaces se incluyen. A diferencia de los enlaces contenidos en 'Otras Citas', éstos no incrementan anualmente. Durante los cuatro primeros años suponen más de un 20% (21,29%, 23,21%, 21,07%, 21,03% respectivamente), en cambio de 2012 en adelante los porcentajes decrecen entre 0,28% y 2,20% anualmente hasta alcanzar un 14,85% en 2018.

A diferencia de Otras Citas y Descripción, las secciones de Reivindicaciones y Resumen no recogen más de unas decenas de resultados por año, en el mejor de los casos. En el caso de Reivindicaciones los resultados no superan los 90 enlaces –recogidos en 2016–, el mínimo se da en 2012 con 5 enlaces recogidos. Lo que implica que no superan el 0,03% de resultados obtenidos para ningún año, siendo el promedio 28 enlaces por año.

En el caso de los enlaces recogidos en el Resumen, éstos no superan lo 22 enlaces (2012 y 2014) y existiendo menos entre 1 y 5 enlaces para 6 de los años. Entre todos ellos, únicamente suponen un 0,04% de los enlaces totales recogidos.

La Figura 32 muestra la evolución anual, en ella se puede observar cómo el incremento de los enlaces recogidos por sección es constante, pero con ligeras variaciones. No se aprecian los enlaces recogidos para Resumen ni Reivindicaciones, ya que el porcentaje es excesivamente pequeño.

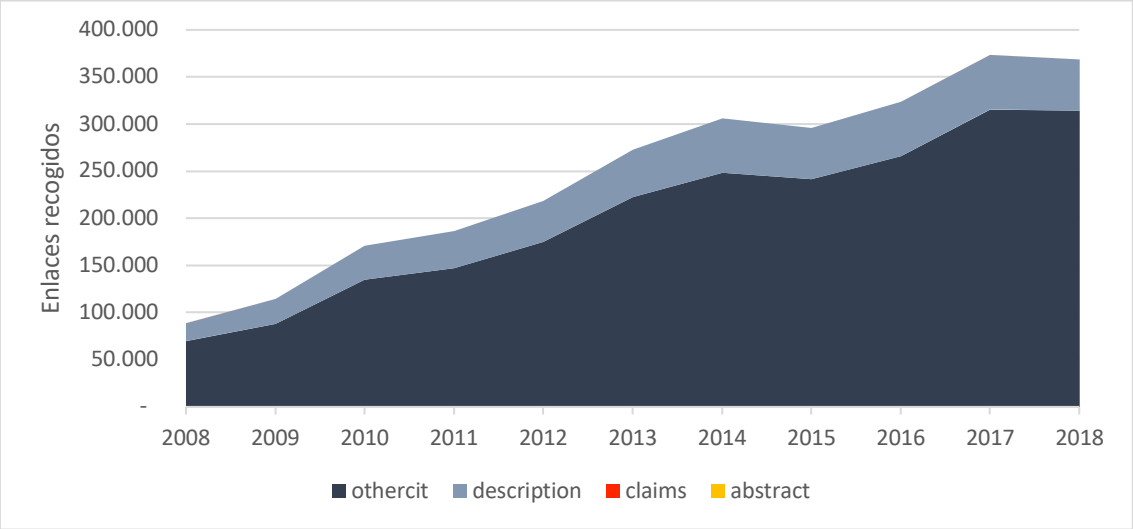


Figura 32: Evolución enlaces recogidos anualmente por sección
Fuente: elaboración propia

Realizando una comparación sobre porcentajes para poder evaluar correctamente sobre el total anual, la Figura 33 muestra que los enlaces en la sección de Otras Citas aumentan de forma continua anualmente.

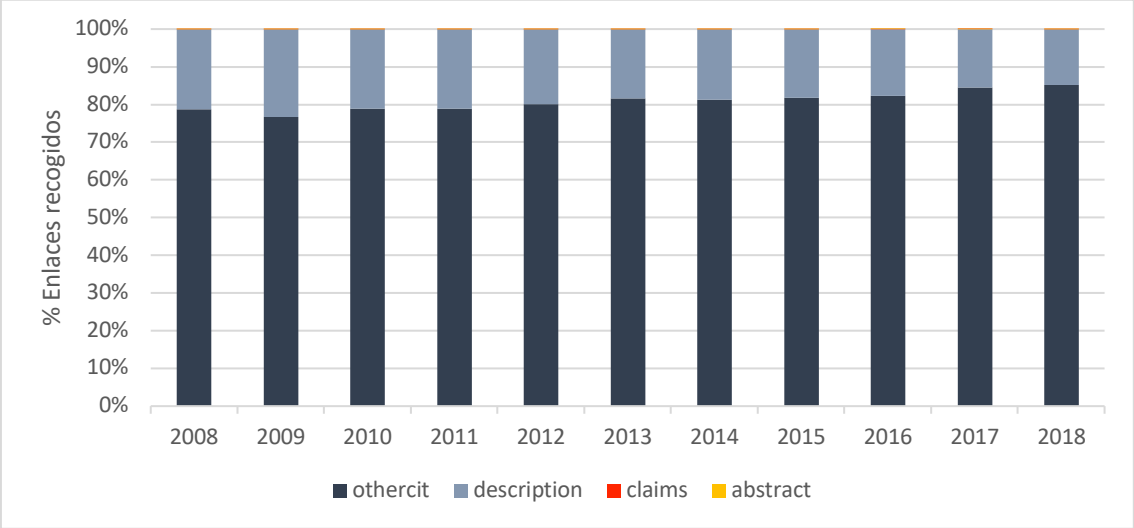


Figura 33: Porcentaje de enlaces recogidos anualmente por etiqueta
Fuente: elaboración propia

4.1.7. Número de enlaces por categoría

La Tabla 33 muestra para cada año, la cantidad de patentes con enlaces recogidos para las áreas que forman el ICPR.

Tabla 33: Áreas de las patentes con enlaces recogidas
Fuente: elaboración propia

	A	B	C	D	E	F	G	H	Total
2008	2.284	1.179	1.757	38	241	419	8.599	3.623	**18.140**
2009	2.954	1.267	2.031	59	270	492	10.655	4.421	**22.149**
2010	4.953	1.910	2.741	80	482	672	15.372	5.991	**32.201**
2011	5.656	2.035	3.206	96	479	748	15.758	6.583	**34.561**
2012	7.390	2.403	3.250	119	610	857	18.798	8.087	**41.514**
2013	9.003	2.742	3.829	130	678	1.181	21.726	9.625	**48.914**
2014	10.172	3.046	4.464	138	823	1.217	23.838	11.415	**55.113**
2015	9.845	3.176	5.065	127	806	1.456	22.158	12.010	**54.643**
2016	10.232	3.342	5.343	126	811	1.649	22.727	12.866	**57.096**
2017	11.428	3.885	5.761	147	1.000	1.926	23.935	14.308	**62.390**
2018	11.045	4.046	5.904	193	951	2.000	23.035	14.184	**61.358**
Total	**84.962**	**29.031**	**43.351**	**1.253**	**7.151**	**12.617**	**206.601**	**103.113**	

Leyenda: A: Necesidades humanas; B: Técnicas industriales, Transportes; Química, Metalurgia; D: Textiles, Papel; E: Construcciones fijas; F: Mecánica, Iluminación, Calefacción, Armamento, Voladuras; G: Física; H: Electricidad

Existe una ligera discrepancia con la cantidad total de patentes únicas recogidas, esto es debido a que en el ICPR no se contempla la clasificación de diseños, en cambio la USPTO los considera patentes, por este motivo las cifras totales anuales son menores que las cantidades vistas anteriormente.

Como se puede observar, existen dos diferencias significativas, una relativa a la evolución anual y otra relativa a las áreas, manifestadas de forma desigual.

Atendiendo a la evolución anual, se puede observar como los resultados se mantienen estables, aunque existe un incremento anual. Para poder analizar correctamente los resultados se realiza la Tabla 34, en la que se muestra mediante porcentajes los enlaces recogidos en función del total para cada una de las áreas. De este modo se puede apreciar cómo, si bien existe alguna diferencia anual dentro de la misma área, ésta es mínima, ya que los porcentajes son en su mayoría idénticos para todos los años.

En cambio, respecto a la diferencia entre áreas, se puede diferenciar tres grupos. El primero se encuentra formado por las áreas B, C, D, E y F, ya que su porcentaje de aparición es menor de 10%. Particularmente, las áreas D, E y F son las que menos enlaces recogen, suponiendo un promedio de resultados 0%, 1% y 2% respectivamente.

El segundo grupo se encuentra formado por las áreas A (Necesidades Humanas) y H (Electricidad), que representan un 17% y 21% respectivamente de los enlaces recogidos.

Tabla 34: Porcentajes de patentes por área según categorización de patentes ICPR
Fuente: elaboración propia

	A	B	C	D	E	F	G	H
2008	13%	6%	10%	0%	1%	2%	47%	20%
2009	13%	6%	9%	0%	1%	2%	48%	20%
2010	15%	6%	9%	0%	1%	2%	48%	19%
2011	16%	6%	9%	0%	1%	2%	46%	19%
2012	18%	6%	8%	0%	1%	2%	45%	19%
2013	18%	6%	8%	0%	1%	2%	44%	20%
2014	18%	6%	8%	0%	1%	2%	43%	21%
2015	18%	6%	9%	0%	1%	3%	41%	22%
2016	18%	6%	9%	0%	1%	3%	40%	23%
2017	18%	6%	9%	0%	2%	3%	38%	23%
2018	18%	7%	10%	0%	2%	3%	38%	23%
Total	17%	6%	9%	0%	1%	3%	42%	21%

Por último, el área G (Física) es en la que más enlaces se encuentran recogidos, suponiendo un promedio para todos los años del 42% (con un mínimo del 40% y un máximo de 48%). Debido a esto, se realiza un análisis preliminar del primer nivel[71] dentro del área, para poder comprobar la distribución de enlaces dentro de la misma.

[71] Como se indica en el §2.1.3, existen más de 70.000 subsecciones, generadas mediante códigos por nivel dentro de la clasificación. Un estudio por niveles completo y anual requeriría de un análisis que queda fuera del contexto de la presente tesis

La Tabla 35 muestra los resultados obtenidos al disgregar por cada una de las subsecciones de primer nivel que conforman el área G para cada uno de los años que conforman el análisis.

De puede observar cómo dentro de una misma área existen, a su vez, discrepancias en la distribución de enlaces. Aparecen de nuevo tres tipos de grupos (pocos enlaces, nivel intermedio y muchos enlaces recopilados).

En el grupo con pocos enlaces se puede encontrar 4 de las 14 subsecciones. Siendo los subgrupos G12 (Detalles instrumentales) y G16 (Tecnología de información y comunicación) en los que prácticamente no se recogen enlaces.

En el grupo intermedio se puede encontrar el grueso de las subsecciones, ya que cuenta con 8 de ellas. En general todas tiene entre 10 y 1000 enlaces. Excepto el grupo G01 (Medidas, Comprobaciones) dónde se puede encontrar hasta 2.900 enlaces (2017).

Por último, en el grupo con mayor cantidad de enlaces se encuentra una única subsección, G06 (Computación, Cálculos) dónde el porcentaje de enlaces contenidos es muy superior al resto (75% del total).

En cambio, realizando en análisis de la evolución anual, se comprueba que el comportamiento es similar al que se refleja en el nivel superior (Tabla 34).

Tabla 35: Cantidad de patentes por desglose de primer nivel del área G (Física)
Fuente: elaboración propia

	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	Total
G01	1.142	378	1.591	1.703	1.787	2.058	2.270	2.293	2.578	2.907	2.765	**21.472**
G02	318	80	290	275	279	348	427	466	561	678	645	**4.367**
G03	146	76	205	223	249	225	230	192	215	218	221	**2.200**
G04	21	16	30	28	32	36	35	20	29	44	44	**335**
G05	148	52	222	237	294	341	373	380	395	479	548	**3.469**
G06	5.982	227	11.760	11.920	14.585	16.905	18.482	16.855	16.802	17.152	16.403	**147.073**
G07	27	18	61	65	69	106	175	168	208	302	295	**1.494**
G08	244	35	342	340	374	402	423	432	475	607	663	**4.337**
G09	216	48	349	358	427	536	646	596	550	551	545	**4.822**
G10	163	56	254	279	297	373	353	361	436	556	482	**3.610**
G11	166	61	227	298	351	345	373	342	399	348	322	**3.232**
G12	2	1	2	1	5	1	1	1	0	0	0	**14**
G16	0	17	1	0	0	0	0	0	0	0	40	**58**
G21	17	0	35	29	42	46	50	52	79	93	62	**505**
Total	**8.592**	**1.065**	**15.369**	**15.756**	**18.791**	**21.722**	**23.838**	**22.158**	**22.727**	**23.935**	**23.035**	**196.988**

Leyenda: G01: Medidas, Comprobaciones; G02: Óptica; G03:Fotografía, Cinematografía, Técnicas analógicas con ondas distintas a las ópticas, Electrografía, Holografía; G04: Horología; G05: Controladores, Reguladores; G06: Computación, Cálculos; G07: Dispositivos de control; G08: Señalización; G09: Educación, Criptografía, Muestras, Publicidad, Sellos; G10: Instrumentos musicales, Acústica; G11: Almacenamiento de información; G12: Detalles de instrumentos; G16: Tecnología de información y comunicación; G21: Física nuclear, Ingeniería Nuclear; G99: Otros.

4.1.8. Tipo de fichero contenido en enlaces

Por último, se ha buscado en todos los enlaces limpios completos recogidos la existencia de enlace directo a ficheros.

Relativo al total de enlaces recogidos por año mostrado en apartados anteriores, la cantidad de enlaces que dirigen directamente a un fichero evoluciona incrementando anualmente, pasando de representar el 6,39% en 2008, hasta alcanzar el 9,33% en 2018.

La Tabla 36 muestra para cada año, los enlaces que se encontraban dirigidos a ficheros, los diferentes tipos de ficheros recopilados.

Tabla 36: Total de enlaces dirigidos a ficheros por año
Fuente: elaboración propia

	PDF	XLS	DOC	PPT	EPUB	TXT	RTF	PNG	JPG	JPEG	Total
2008	3.811	4	72	42	-	1.116	6	8	618	1	**5.678**
2009	5.489	8	111	65	-	1.176	5	9	496	7	**7.366**
2010	9.939	9	336	165	1	1.653	5	40	611	9	**12.768**
2011	12.032	5	418	181	2	2.091	12	65	752	8	**15.566**
2012	16.985	7	516	275	-	1.648	5	73	888	14	**20.411**
2013	22.765	8	624	258	-	1.635	10	104	903	9	**26.316**
2014	26.151	16	742	281	-	1.698	10	142	1.113	7	**30.160**
2015	26.089	18	713	234	2	1.601	15	206	1.089	9	**29.976**
2016	27.748	7	676	254	-	1.113	1	220	1.286	17	**31.322**
2017	31.056	12	852	246	1	904	2	232	1.371	13	**34.689**
2018	31.058	16	759	223	-	1.007	2	218	1.102	16	**34.401**
Total	**213.123**	**110**	**5.819**	**2.224**	**6**	**15.642**	**73**	**1.317**	**10.229**	**110**	**248.653**

En la Figura 34 se puede apreciar como el tipo de fichero más utilizado es el PDF (86%), seguido de ficheros de texto (DOC –2%– y TXT –6%–) y de imágenes de tipo JPG (4%). Los formatos con menos resultados son las hojas de cálculo (XLS), fichero de texto de tipo RTF, presentaciones de PowerPoint (PPT) y los formatos de imagen PNG y JPEG. Resulta anecdótico los resultados obtenidos para los libros electrónicos de formato EPUB, ya que tan solo se dan 6 resultados para todos los años.

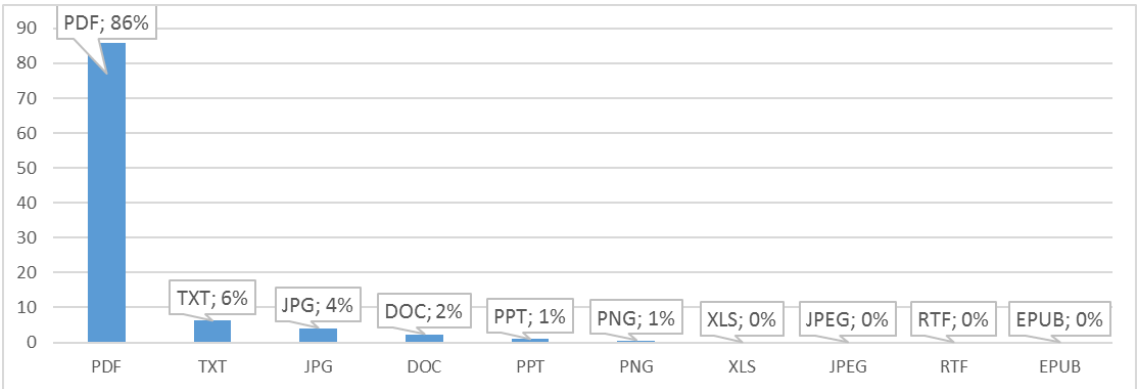


Figura 34: Distribución mediante porcentajes de los resultados totales por tipo de fichero enlazado
Fuente: elaboración propia

4.2. Bloque Patent Inlink: Análisis de enlaces de recursos web a Patentes

El Bloque Patent Inlink realiza un análisis de enlaces dirigidos a los documentos de patentes desde cualquier documento web. De este modo se puede comprobar la visibilidad e impacto de los documentos en el entorno web.

4.2.1. Análisis descriptivo de enlaces web a patentes

Dentro de los datos recogidos, existen 2.297.366 enlaces recogidos dirigidos a 990.663 URLs de patentes, realizados desde 17.001 dominios únicos.

La Tabla 37 muestra un análisis descriptivo de los datos recopilados por dominio. Además de los datos indicados anteriormente relativos a los totales de enlaces y dominios, se encuentran los relativos a los máximos y mínimos. En este caso existe un dominio (datamp.org) cuyo máximo es de 744.679 enlaces (32% del total). Respecto al mínimo, existen 5.916 dominios web (35% del total) con un único enlace a patente.

Tabla 37: Análisis descriptivo de los dominios con enlaces a patentes
Fuente: elaboración propia

	Datos descriptivos
Nº Dominios	17.001
Total enlaces	2.297.366
Mínimo	1
Máximo	744.679
Frec. del mínimo	5.916
Frec. del máximo	1
1° Cuartil	1
Mediana	2
3° Cuartil	7
Media	135,13
Varianza (n-1)	50.267.212,39
Desviación típica (n-1)	7.089,94

Se puede apreciar la diferencia entre la mediana y el 3º cuartil, encontrándose en este último 7 enlaces. Atendiendo al valor promedio, se obtienen 135,13 enlaces a patentes por dominio.

El análisis de la varianza y la desviación típica muestran que la dispersión de los datos con respecto a la media ofrece 50.267.212,39 y 7.089,94 respectivamente. Esto nos indica que existe una variabilidad en los datos muy elevada, es decir, unos pocos dominios web generan la mayor cantidad de enlaces a patentes.

4.2.2. Enlaces desde recursos web hacia patentes del Bloque Patent Outlink

Debido a que para poder realizar un estudio masivo se han descargado todos los datos existentes en Majestic para el subdominio "patents.google.com" entre los datos relativos a enlaces a patentes (aquellos con una URL de tipo "patents.google.com/US*número de patente*") existen patentes que no pertenecen al periodo analizado en el Bloque Patent Outlink. Por ello, se realiza una comprobación comparativa para conocer cuántas de las URLs del Bloque Patent Inlink contienen patentes del Bloque Patent Outlink (patentes concedidas en Estados Unidos entre 2008 y 2018).

Se recogen 990.663 patentes únicas entre las URLs del Bloque Patent Inlink, aunque tan solo se localizan enlaces para 8.279 patentes del Bloque Patent Outlink, todas ellas pertenecientes a 2018. Se trata de un número pequeño en comparación con el total de patentes del Bloque Patent Outlink (0,26%); que acumulan 30.574 enlaces, un 1,33% del total de enlaces recopilados en el Bloque Patent Inlink.

4.2.3. Análisis por Categoría

Majestic ofrece una categorización propia para describir el contenido de los dominios. Entre todos los dominios recogidos existen 18 posibles categorías: Entretenimiento, Tecnología, Deportes, Artes, Negocios, Ciencia, Sociedad, Salud, Referencias, Regional, Juegos, Compras, Noticias, Hogar, Internacional, Adulto.

Esta categorización no siempre se recoge correctamente por la herramienta, por lo que existen 4.760 dominios que no se encuentran categorizados.

La Tabla 38 muestra la cantidad de enlaces existentes para cada una de las categorías.

Analizando el contenido, algunas de las URLs contenidas en la categoría "Entretenimiento" son:

- → Google en diversos países
- → Dataamp.org, un directorio de patentes americano de herramientas y maquinaria
- → Explainthatstuff.com, una colección de artículos a modo de enciclopedia de diferentes tipos de contenido (ciencia, tecnología, etc.)

Si bien es cierto que Majestic ofrece una categorización más facetada y en este análisis únicamente se centra en la primera de las categorías para poder aglutinar el grueso de éstas y tener una descripción general, se debe tener en cuenta este dato en el momento de equiparar los resultados a los obtenidos en el Bloque Patent Outlink.

Tabla 38: Categorización de los dominios según Majestic
Fuente: elaboración propia

Ref Domain Topical Trust Flow	Enlaces
Desconocido	804.992
Entretenimiento	775.256
Tecnología	226.501
Artes	109.910
Sociedad	85.932
Deportes	81.654
Ciencia	65.732
Negocio	62.778
Salud	32.049
Referencias	27.721
Regional	7.213
Juegos	6.019
Compras	5.270
Noticias	4.330
Hogar	1.450
Internacional	306
Adulto	253

4.2.3.1. Categorización de enlaces recopilados

Para poder comparar ambos bloques con el mismo sistema de análisis, para el Bloque Patent Inlink se selecciona la misma cantidad de dominios seleccionada en el anterior bloque (201), pese a que únicamente 97 dominios superan los 1.000 enlaces recopilados (característica para limitación del Bloque Patent Outlink). El dominio con menos enlaces de los analizados cuenta con 389.

La Tabla 39 muestra la categorización realizada mediante el primer sistema de categorización (Empresa, Servicio, Organización, Gobierno, Media, Universidad). En esta ocasión, no se encuentra ningún dominio bajo la categoría Gobierno. Destacan los dominios de Empresa (55% del total) y Servicios (39%) por ocupar los dos primeros puestos, alcanzando a suponer 166 de los 201 dominios analizados.

Tabla 39: Primera categoría para los dominios recogidos
Fuente: elaboración propia

Tipo de organismo	Cantidad de Dominios	Número de enlaces
Servicio	94	1.160.634
Empresa	72	826.499
Media	13	96.434
Organización	7	8.433
Universidad	3	1.895

Existe otra diferencia con respecto a la Tabla 30, Media supera tanto a Organización como Universidad y recoge casi 100.000 enlaces, suponiendo el 5% del total de enlaces del Bloque. En cambio, Organización y Universidad no llegan a representar un 1%.

Atendiendo a la clasificación por contenido de los dominios, la Tabla 40 recoge los resultados obtenidos para los resultados del Bloque Patent Inlink.

Tabla 40: Categorización por tipo de contenido con total de dominios y número de enlaces recogidos
Fuente: elaboración propia

Tipo de contenidos	Cantidad de dominios	Número de enlaces
Negocios	41	896.078
Sitios Personales	20	533.486
Tecnología	26	300.618
Gobierno	5	122.363
Deportes	3	81.244
Adultos	9	43.353
Entretenimiento	24	35.554
Educación	12	19.607
Compras	13	18.917
Salud y Medicina	13	17.605
Motores de búsqueda	12	13.524
Noticias	6	5.426
Automoción	2	1.494
Finanzas	1	581

Se puede observar como el primer resultado obtenido es para Negocios (o Empresas) que obtiene 896.078 enlaces (42,88%), seguido de Sitios Personales (25,53%) y Tecnología (14,38%).

Es remarcable que Gobierno se encuentra representado únicamente por 5 dominios, pero supone un total del 5,86% de los enlaces.

Aparecen las secciones Deportes, Adultos y Entretenimiento que recopilan más enlaces que aquellos dominios relativos a Educación.

4.2.4. Análisis por TLD

Dentro de los 17.001 dominios únicos que se encuentran enlazando a patentes, únicamente 1.026 (6%) contienen un SLD (i.e.: ac.uk), siendo 382 únicos y aglutinando 38.638 enlaces en total.

Para poder analizar los TLDs se muestra en la Tabla 41 los primeros 50 que representan el 99% de los enlaces.

Tabla 41: Recuento de enlaces por TLD para dominios enlazando a patentes
Fuente: elaboración propia

TLD	Enlaces								
com	638.766	**kz**	5.988	**center**	2.204	**es**	1.076	**tw**	718
io	506.831	**nl**	5.019	**ch**	2.061	**live**	1.030	**za**	712
today	67.040	**pe**	4.868	**be**	1.783	**vu**	999	**ovh**	634
net	64.789	**info**	3.256	**hu**	1.564	**cz**	998	**gi**	631
ca	15.094	**me**	2.789	**ge**	1.544	**fi**	984	**pt**	595
co	12.623	**eus**	2.713	**au**	1.436	**social**	894	**na**	590
cn	11.964	**ai**	2.671	**fr**	1.377	**gm**	864	**sg**	579
us	11.550	**ru**	2.619	**it**	1.272	**buzz**	830	**pl**	544
de	8.197	**edu**	2.297	**in**	1.148	**se**	779	**icu**	538
uk	6.739	**no**	2.243	**br**	1.094	**zm**	762	**lv**	526

Se puede observar como los primeros puestos son ligeramente diferentes a los mostrados en el §4.2.4 para los TLDs recogidos dentro de las patentes. En este caso, si bien .com mantiene la primera posición con el 27,80% de los enlaces totales, se incorpora a la lista .io y .today como TLDs que enlazan a patentes. .io, con un total del 22,06% de enlaces, es el TLD representativo del territorio Británico del Océano Indico, pero se utiliza extraoficialmente por empresas tecnológicas.

Tanto .today como .net recopilan el 3% de los enlaces por TLD. Los siguientes en la lista son .ca, .co, .cn y .us que pese a tener entre 11 y 16 mil enlaces no superan el 1%. El resto de TLDs no alcanzan una representación del 1%, siendo en total 327 TLDs, de los cuales 132 reciben menos de 10 enlaces.

En relación con los SLD mencionados anteriormente, la Tabla 42 muestra los resultados para los territorios que utilizan el formato doble a nivel de dominio. En los primeros puestos se puede encontrar a China y Estados Unidos con ~11.000 resultados cada uno (31 y 30% respectivamente). Seguidos por Reino Unido con un 17% de los enlaces. Tras esto existe un salto en cantidad de enlaces, ya que tanto Georgia como Australia recopilan un 4%.

Tabla 42: Países que utilizan SLDs y número de enlaces
Fuente: elaboración propia

ccTLD	País	Nº Enlaces
cn	China	11.964
us	Estados Unidos	11.550
uk	Reino Unido	6.739
ge	Georgia	1.544
au	Australia	1.436
br	Brasil	1.094
tw	Taiwán	718
za	Zambia	712
sg	Singapur	579
ar	Argentina	449
il	Israel	438
mx	Méjico	391
jp	Japón	325
kr	Ucrania	230
zw	Zimbabue	138
vn	Vietnam	125
nz	Nueva Zelanda	120
ph	Filipinas	36
my	Malasia	32
at	Austria	9
be	Bélgica	5
cr	Costa Rica	4

Ateniendo al segundo nivel de SLDs, la Tabla 43 muestra los primeros resultados por tipo de SLD, siendo la mayoría de tipo Comercial, quedando aquellos relacionados con Educación y Gobierno con menor peso.

Tabla 43: Resultados obtenidos por tipo de SLD
Fuente: elaboración propia

SLD	Tipo	Nº Enlaces
com	Comercial	6.152
co	Comercial	5.946
org	Organización NG	1.347
net	Network	1.129
edu	Educación	294
ac	Académico	190
gov	Gobierno	35

4.2.5. Análisis por Idioma

Majestic ofrece la categorización por idioma de los sitios recogidos. Este dato no se encuentra siempre disponible, pero es un indicador que permite realizar análisis sencillos sobre el lenguaje del contenido. En este caso, debido a que las patentes son todas de Estados Unidos, resulta interesante comprobar mediante qué lenguaje se enlaza hacia las mismas pese a que el sistema no es perfecto y no siempre localiza el idioma.

De los 17.001 dominios recopilados, existen 264 de los cuales no se localiza información relacionada con el idioma del su contenido, esto supone 766.979 enlaces (33,38% del total) sin categorizar.

Existen 62 idiomas, la Tabla 44 contiene los primeros 10 idiomas, aquellos que contienen más de 1.000 enlaces. Se puede apreciar como el idioma con mayor número de enlaces (64,76% del total) es el inglés. El resto de los idiomas de la tabla, pese a que recogen más de 1.000 enlaces (aunque menos 10.000), así como aquellos que se encuentran fuera de ella, no llegan a representar más del 1,86% del total de enlaces.

Tabla 44: Idioma con número de enlaces para los dominios recopilados
Fuente: elaboración propia

Idioma	Nº Enlaces	Porcentaje
Inglés	1.487.762	64,76%
Desconocido	766.979	33,38%
Alemán	9.893	0,43%
Español	7.256	0,32%
Ruso	4.163	0,18%
Francés	3.343	0,15%
Vasco	2.713	0,12%
Húngaro	1.843	0,08%
Chino	1.555	0,07%
Finlandés	1.342	0,06%
Holandés	1.117	0,05%
Italiano	1.072	0,05%

Comparándolo con los resultados del apartado anterior, a tenor de lo visto en la Tabla 42, se puede observar como por ejemplo en el caso de China, se recogen 11.964 enlaces, pero de acuerdo con el idioma de la página, únicamente se reconocen 1.555.

Como anécdota, la totalidad de los enlaces en vasco se encuentran bajo el mismo dominio laboratorium.eus, la página web del museo científico Laboratorium de Vergara (Guipúzcoa).

4.2.6. Análisis descriptivo por patentes

Como se ha indicado anteriormente, existen 990.663 patentes únicas enlazadas desde los 17.001 dominios. La Tabla 45 muestra el análisis descriptivo de los datos de enlaces a patentes recopilados.

Tabla 45: Datos descriptivos de los enlaces dirigidos desde recursos web a patentes
Fuente: elaboración propia

	Valores
No. de observaciones	990.663
Suma	2.297.366
Mínimo	1
Máximo	80.630
Frec. del mínimo	830.782
Frec. del máximo	1
1° Cuartil	1
Mediana	1
3° Cuartil	1
Media	2
Varianza (n-1)	20.477,57
Desviación típica (n-1)	143,1
Asimetría (Pearson)	435,9
Media geométrica	1,24

Tal y como se muestra, la frecuencia del valor mínimo es 839.782, que representan el 84% del total de enlaces. Atendiendo a los datos de los cuartiles y la mediana aparece en todos ellos como resultado 1, corroborando el dato anterior. La varianza muestra que existe una gran dispersión en el número de enlaces, ya que su valor es de 20.477,57. Así como la desviación típica que se encuentra en 143.

La Tabla 46 muestra aquellas patentes con un porcentaje de número de enlaces superior al 1%, siendo en total 12 patentes. Como se puede observar, existen dos patentes con 80.630 (US8554769: *Identifying gibberish content in resources, 2009*) y 80.590 (US8955129: Method and system for detecting fake accounts in online social networks, 2013) enlaces cada una, suponiendo cada una un 7% del total de los enlaces recopilados.

Estas 12 patentes, aglutinan el 35% del total de enlaces, siendo su sumatorio 399.446.

Tabla 46: Pantentes enlazas con un porcentaje superior al 1%
Fuente: elaboración propia

Patente Enlazada	Nº Enlaces	Porcentaje	Patente Enlazada	Nº Enlaces	Porcentaje
US8554769	80.630	7	US8638908	33.645	3
US8955129	80.590	7	US20140248834	27.497	2
US20140161250	33.645	3	US7424516	14.063	1
US20140177813	33.645	3	US6630507B1	11.843	1
US8068604	33.645	3	US10055034B2	8.299	1
US8553852	33.645	3	US10061457B2	8.299	1

4.2.7. Indicadores de sitios que enlazan (Majestic Style)

En el presente apartado se realiza un análisis de los indicadores que ofrece la herramienta Majestic comentados en el §3.2.1.1. La Tabla 47 muestra un análisis descriptivo de los datos relativos a ambos indicadores.

Tabla 47: Análisis descriptivo para los datos recopilados relativos a Citation Flow y Trust Flow
Fuente: elaboración propia

	Trust Flow	Citation Flow
Mínimo	0	0
Máximo	97	96
Frec. del mínimo	4.759	1.750
Frec. del máximo	1	1
1° Cuartil	0	11
Mediana	9	22
3° Cuartil	22	36
Media	15,148	24,218
Varianza (n-1)	330,079	298,04
Desviación típica (n-1)	18,168	17,264
Correlación (Spearman)	0,849	

Como se puede observar, se encuentra enlaces prácticamente para todos los posibles valores del indicador ya que existen 97 y 96 valores únicos. En el caso del TF, existe un mayor número de páginas cuyo resultado es 1 (35% del total); además, tanto la media como la mediana son valores bajos, 15 y 9 respectivamente, lo que implica que la calidad de los enlaces, en general, no es muy alta.

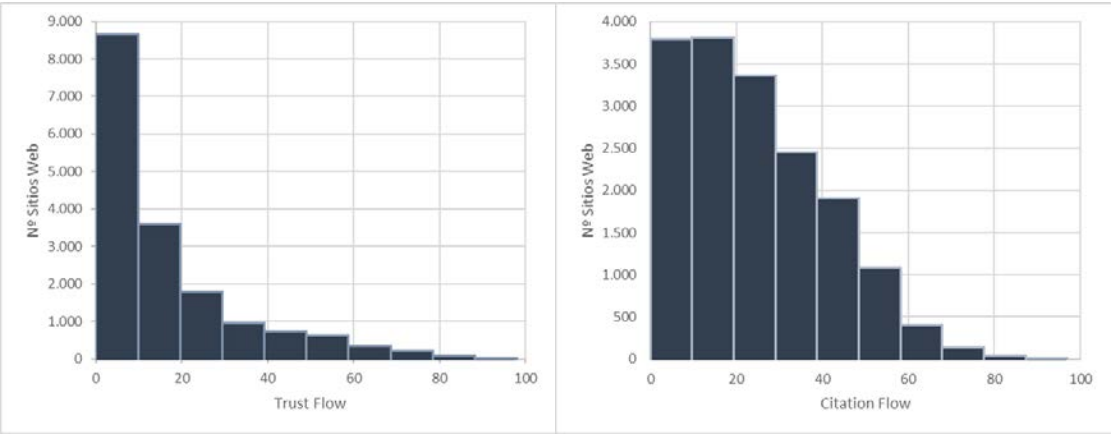


Figura 35: Histograma Trust Flow y Citation Flow
Fuente: elaboración propia

El primer cuartil recoge el 54% de los enlaces recopilados y el segundo cuartil obtiene el 45%, por lo que este dato no deja de señalar que la mayoría de los enlaces tienen un TF menor de 50 puntos. En relación con el CF, se puede observar como en este caso los valores son algo más altos, ya que tanto la frecuencia del valor mínimo (1.750) como la

mediana o los datos contenidos en el tercer cuartil son mayores (22 y 36 respectivamente).

Para ambos indicadores se observa como tanto la varianza como desviación típica es similar. Así mismo, su correlación es significativa ya que obtienen un 0,849. La Figura 36 muestra los gráficos de dispersión y correlación generados, comparando los dos indicadores. Se puede observar cómo existe una correlación alta con pocos valores atípicos, siendo estos valores atípicos, en su mayoría, superiores a 50 para CF.

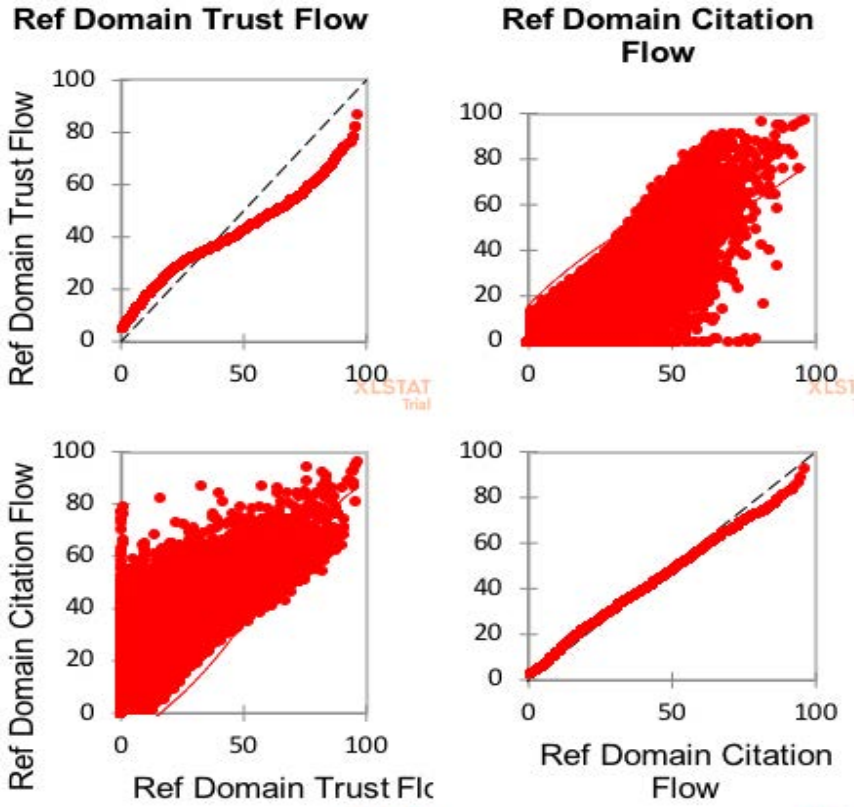


Figura 36: Gráficos de dispersión de Citation Flow y Trust Flow
Fuente: elaboración propia

La Tabla 48 muestra los resultados obtenidos para cada uno de los valores de CF que obtiene más de 1% de enlaces recopilados. Como se puede observar, el 40% está formado por enlaces que tienen un CF de 47 y 50; en cambio, aquellos con un valor CF menor a 10 (9 y 0) obtienen en total el 31%.

De igual modo que sucede con los resultados de TF, pese a que los valores recopilados para CF son mayores en general, el grueso de los enlaces se encuentra recopilados en los dos primeros cuartiles.

Tabla 48: Enlaces recopilados por valor de Citation Flow con una representación superior al 1%
Fuente: elaboración propia

Citation Flow	Nº Enlaces	Porcentaje
47	752.599	33%
9	591.399	26%
50	170.519	7%
0	108.077	5%
15	89.770	4%
23	77.320	3%
31	34.543	2%
6	30.793	1%
7	28.503	1%
8	28.326	1%
10	28.189	1%
12	25.163	1%
28	23.040	1%
34	21.186	1%
29	19.535	1%
32	16.860	1%
21	14.198	1%
33	13.526	1%
48	12.704	1%

La Tabla 49 muestra aquellos dominios que obtienen una puntuación para TF superior a 90 puntos por lo que se pueden considerar de alto nivel, junto con los enlaces recopilados y el valor que obtienen en CF. Los dominios destacan por representar a Universidades y empresas tecnológicas mayoritariamente.

Tabla 49: Dominios con TF superior a 90, junto con el número de enlaces recopilados y el valor de CF
Fuente: elaboración propia

Dominio	Nº Enlaces	Trust Flow	Citation Flow
blogspot.com	99	90	86
si.edu	27	90	68
harvard.edu	24	91	74
loc.gov	16	91	68
github.com	11	96	95
bbc.co.uk	8	95	93
yahoo.com	7	95	88
ietf.org	6	96	81
mit.edu	6	91	71
mozilla.org	3	93	89
stanford.edu	3	91	71
cam.ac.uk	2	90	64
flickr.com	2	95	87
statcounter.com	2	94	92
linkedin.com	1	97	96

A continuación, se incorporan las Figura 37 y Figura 38 para la representación gráfica de ambos indicadores web, de forma que se pueda visualizar de un modo sencillo la segregación y diferencia de tamaños para cada uno de ellos.

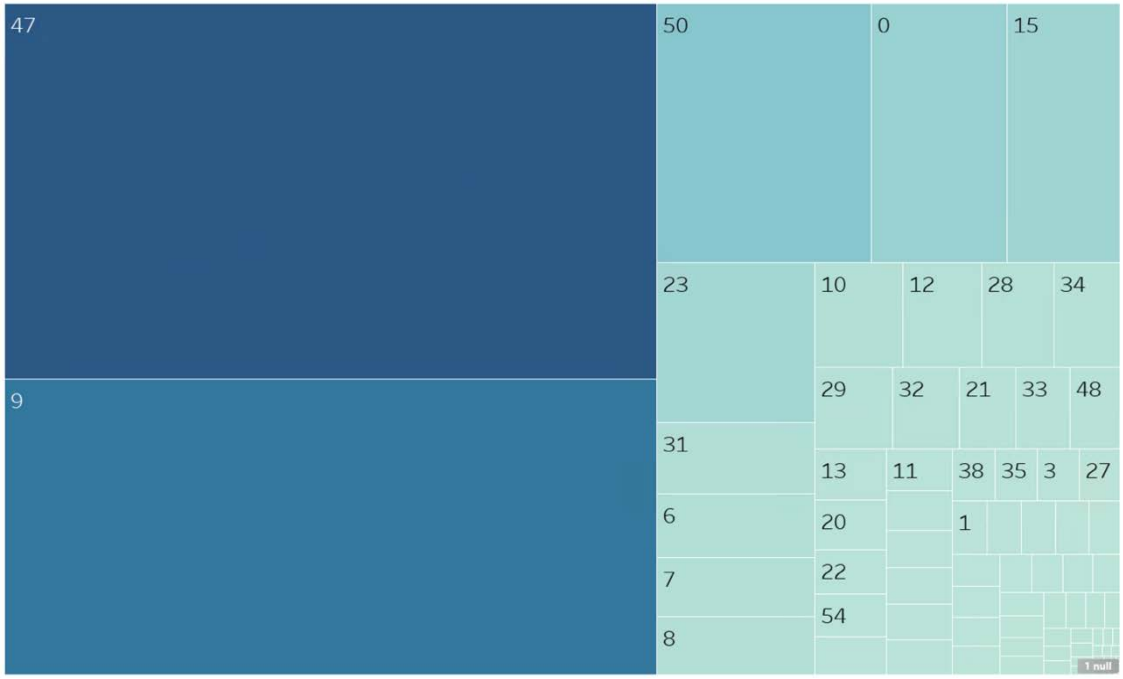


Figura 37: Representación gráfica de los datos relativos a Trust Flow
Fuente: datos propios; visualización: Tableau

En ambas figuras se puede observar cómo existen dos valores de los indicadores con una gran cantidad de resultados. En el caso de Citation Flow la segregación es algo menor que en Trust Flow entre los valores con mayor cantidad de enlaces.

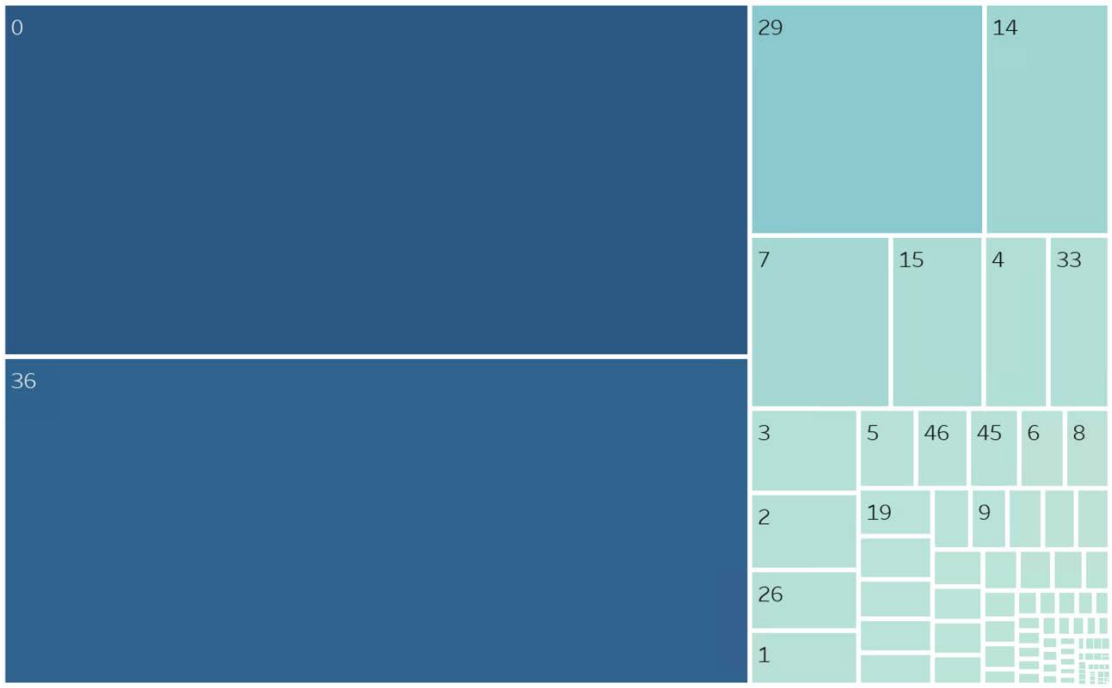


Figura 38: Representación gráfica de los datos relativos a Citation Flow
Fuente: datos propios; visualización: Tableau

Capítulo 5
Discusión

Tras la propuesta de un modelo para la recopilación, extracción, preparación y almacenamiento de los datos en el Capítulo 3, así como su análisis en el Capítulo 4, el presente Capítulo se centra en realizar una discusión de ambos.

Siguiendo la estructura de la tesis, la información se encuentra dispuesta en dos apartados; el primero recoge la discusión relativa a la validez del modelo propuesto, teniendo en cuenta los dos bloques de información, el sistema de captura y manejo de los datos, así como las oportunidades de mejora existentes.

El segundo apartado analiza los datos recogidos y la validez de los resultados y su análisis, adentrándose en la aplicabilidad del conocimiento extraído en el análisis.

Se pretende dar una visión crítica de todo el proceso desarrollado, desde el planteamiento inicial hasta la consecución de los resultados mediante su análisis descriptivo y estadístico.

5.1. Discusión de los resultados

Como se ha podido observar en el capítulo de resultados, las cifras globales tanto de patentes como de enlaces son muy altas (más de 3 millones de patentes, y más de 5 millones de enlaces entre los dos bloques), existiendo una alta prevalencia en los datos de enlaces. Esto ha permitido realizar un análisis relativo a la evolución temporal capaz de arrojar suficiente información para comprobar la utilidad del método planteado y su proyección futura, así como la viabilidad del uso de la técnica de Patent Link Analysis.

A continuación, se desarrolla una discusión de los resultados obtenidos para cada uno de los bloques.

5.1.1. Análisis de Patent Outlink

A lo largo de este bloque se ha trabajado con los datos de 3.133.247 patentes y 2.740.634 enlaces recogidos, producidos por la USPTO desde 2008 a 2018. Debido a la gran cantidad de enlaces, el número de links totales es significativo para poder realizar el análisis.

El análisis de ambos objetos muestra cómo mantienen una evolución de incremento anual durante el periodo examinado. En el caso de las patentes, esta evolución muestra signos de desaceleración; en cambio para los enlaces, los últimos años incrementa su uso con una mayor rapidez, lo que sugiere que su uso es cada vez más relevante.

En ese sentido, se observa un incremento en la cantidad de patentes que contienen enlaces, pasando de suponer un 15,57% en 2008, a un 20,68% en 2018. Existe una anomalía para los años 2009 – 2011, donde el porcentaje de patentes con enlaces es menor, pero se podría asociar a uno de los efectos de la crisis de 2008, ya que pese a incrementar la concesión de patentes, el porcentaje de subida es menor que en años posteriores. En todo caso, debería comprobarse con la incorporación en el análisis de años anteriores, asumiendo los problemas técnicos y comparativos que esto supondría (posible peor calidad de los datos, errores de recolección, etc.).

En relación con los enlaces contenidos, la media y le mediana son 2 y 4-5 respectivamente; existiendo, además, una distribución altamente desigual (existen patentes con muchísimos enlaces, pero son una minoría). Además, si bien es cierto que para todos los años el valor mínimo de enlaces contenidos es 1, su frecuencia de aparición disminuye con el paso de los años (38,42% en 2008, 35,42% en 2018), esto implica que, pese a no ser un recurso ampliamente utilizado en actualidad, cada vez existen más patentes con una mayor cantidad de enlaces.

Sucede así mismo una tendencia a la baja con los enlaces únicos, aquellas URLs que aparecen una única vez a lo largo del estudio. En la evolución anual se puede observar cómo el porcentaje disminuye (pasando de 28% inicial a un 18% final), aumentando la variedad de dominios que se incluyen y abriendo el abanico informativo. Esta variedad de fuentes de información nos permite, además de localizar los recursos web más relevantes debido a su mención dentro de un documento de patente, encontrar

aquellos que reciben una cantidad alta de menciones convirtiéndolos en los recursos y fuentes de información más importantes.

Atendiendo a los dominios recopilados y realizando un análisis por nivel, encontramos que el grueso de enlaces recogidos (69%) pertenece al primer nivel. En el segundo y tercer nivel, pero a encontrarse menos dominios se han localizado algunos con un número mayor de enlaces recopilados. Debido a que los dominios con múltiples niveles se relacionan con sistemas organizacionales de información específicos (grupos o centros de investigación dentro de una institución mayor, servicios, productos concretos, etc.) estos ofrecen una mayor cantidad de información (relativa al recursos) que a su vez es considerada un recurso de calidad. Además, dado que normalmente en la redacción de un texto se busca la simplificación, el que se encuentre detallado más de un nivel implica una mayor intencionalidad informativa por parte del inventor.

La diferenciación por nivel permite realizar una agrupación posterior, sumando aquellos dominios que se encuentran en niveles superiores a los resultados del primer nivel, de modo que se pueda comprobar la presencia real de los dominios en el computo total. De este modo se puede observar que dominios como archive.org (97.150 enlaces totales) o wikipedia.org (82.191 enlaces totales), que en primer nivel únicamente reciben 9.692 y 5.306 respectivamente, en segundo nivel 87.177 y 76.100 enlaces respectivamente. Por tanto, esto deberá ser tenido en cuenta en futuros análisis, de modo que los resultados reflejen correctamente la presencia y uso de los dominios. Asimismo, sería interesante realizar un análisis multinivel para los dominios que cumplan este requisito, de modo que se pueda estudiar en profundidad el alcance y tipología de los dominios.

Estos resultados refuerzan los obtenidos por (Orduña-Malea et al., 2016), quienes identifican Archive.org, Wikipedia.org y Youtube.com (aunque en los resultados de esta tesis con una menor presencia) como fuentes receptoras de enlaces desde patentes con una muestra diferente tanto en años como en cobertura.

En línea con los resultados obtenidos hasta este momento que apuntan a que los enlaces contenidos en patentes se utilizan cada vez más para documentar e informar bien los documentos, se encuentran los resultados obtenidos en relación con las secciones en las que se incrustan. De las cuatro secciones analizadas (‘Otras citas’, Resumen, Descripción, Reivindicaciones) la que más acumula es ‘Otras citas’ (85% en 2018), seguida de Descripción (15% en 2018). Si bien es cierto que se podría comparar con la cantidad de citas que no contienen enlaces para ver la proporción existente, el resultado obtenido para la sección ‘Otras Citas’ es relevante. Otro análisis que permitiría obtener más información sería el estudio en detalle de las citas contenidas en el propio texto (Descripción, Resumen, Reivindicaciones) junto con el texto colindante para descubrir la relación existente con las palabras clave utilizadas y el peso o importancia del enlace en el texto.

Además, pese a que los datos para los enlaces recopilados en las secciones textuales del documento son muy reducidos (particularmente para Resumen y Reivindicaciones)

debido a la importancia de estas secciones sería interesante ampliar el estudio de los enlaces contenidos.

Observando la categorización realizada de los 201 dominios con más enlaces recopilados, se aprecia la alta presencia en número de enlaces tanto de empresas, servicios y organizaciones. Es un resultado que indica una predilección, por parte de las personas inventoras, a la hora de incluir recursos online procedentes del sector privado.

Resulta remarcable que tanto las Universidades como los Gobiernos reciben una cantidad significativa de enlaces. Obteniendo los gobiernos casi la misma cantidad de enlaces (-5.000 enlaces) que Universidades, pero con la mitad de los dominios (35 frente a 12 respectivamente).

Respecto a la temática, en la primera posición se encuentra Tecnología (410.000 enlaces), seguida por Educación (280.000 enlaces), datos que concuerdan con lo visto en los párrafos anteriores, pero que destaca teniendo en cuenta que la representación de universidades comentada anteriormente. Aunque es cierto que ambas categorías logran esos resultados –y su diferencia– con 70 y 73 dominios, lo que denota el escalón existente entre una y otra. En el futuro se podría realizar una tercera categorización (o categorización temática doble) que mostrase la tipología de una doble naturaleza y comprobar si existe un mayor o menor solape entre tipos y categorías de contenidos.

Estos resultados se pueden alinear, a su vez, con los obtenidos al analizar las áreas de las patentes que contienen los enlaces. La mayoría de las patentes (42%) pertenecen al área de física, especialmente al área G06 de Computación y Cálculos, por lo que existe una alta relación con la temática de los dominios vista anteriormente. Futuros trabajos tratarán de analizar en detalle el resto de áreas de conocimiento para conocer posibles diferencias por disciplina. Además, se debería comprobar los dominios más utilizados por áreas para comprobar si existen diferencias significativas entre fuentes de información destino.

Al analizar la presencia por TLD se puede observar como gran parte de los resultados corresponden a .com (.co en los SLDs, 52%). Existen dos motivos principales para que esto suceda: .com se crea para denominar un alojamiento comercial y, como muestran los resultados de la categorización de las páginas, las Empresas son las que más enlaces reciben. Pero, además, desde los años 1990 no existen restricciones para utilizar este TLD, por lo que, aunque no contenga contenido de tipo comercial, puede ser utilizado por los usuarios. Es por ello que .com es el TLD con mayor presencia en internet como se indica en el §2.2.1.1, por lo que debe ser tenido en cuenta al contextualizar los datos.

Aunque exista un salto en la cantidad de datos recopilados, destaca que los siguientes puestos se encuentren ocupados por los TLD .edu y .gov. Éstos están dirigidos a páginas centradas en contenido de Organizaciones no gubernamentales, Educación y Gobiernos. Comparando estos resultados con los totales utilizados en internet como en el caso de .com, resulta notable que tanto .edu como .gov únicamente constituyen el 0,1% o menos de todo internet, en cambio en las patentes tienen una alta representación. Los datos obtenidos revelan una fuerte relación existente entre las patentes y los recursos web de instituciones oficiales, como gobiernos, universidades y organizaciones.

Por último, aunque no se encuentra relacionado estrictamente con la información descrita hasta este punto, el análisis de los tipos de ficheros contenidos muestra cómo su uso es cada vez mayor. En futuro trabajos se analizarán los recursos a nivel de fichero con el fin de identificar y clasificar los documentos enlazados (siempre que el enlace llegue a ese nivel de especificidad), con el fin de conocer la naturaleza de estos documentos (artículos, mapas, gráficos, software, datos brutos, presentaciones, etc.).

5.1.2. Análisis de Patent Inlink

Para el Bloque Patent Inlink se utilizan los 2.297.366 enlaces recopilados utilizando la herramienta Majestic. En este caso, existe una menor cantidad de dominios (17.001), lo que indica la existencia de pocas páginas web que utilizan las patentes como recurso informacional. El número de recursos online que incluyen enlaces a patentes es reducido, lo que implica que las patentes no son ampliamente enlazadas o utilizadas como recursos de información en las páginas web, considerando la restricción de haber analizado enlaces a Google Patents. Entre las páginas recopiladas destaca datamp.org con un total de 744.679 enlaces a documentos de patentes. En este portal se recopila y documenta desde Google Patents información sobre patentes relacionadas con maquinaria y herramientas.

Pese a que la media indica que se realizan 135 enlaces, la varianza (50.267.212) y desviación típica (7.089), muestran una elevada dispersión de los datos, con unos pocos dominios web generando la mayor cantidad de enlaces a patentes. Por ello, sería interesante realizar un análisis en detalle de las fuentes que realizan todos estos enlaces para conocer más sobre los motivos que llevan al enlazado.

Los enlaces recopilados apuntan a 990.663 patentes, por lo tanto, éstas reciben de media más de dos enlaces. La desviación y varianza en este caso muestran una menor dispersión, quedando más repartidos los enlaces realizados. Existen dos patentes en particular que reciben una gran cantidad de enlaces (7% del total cada una), aunque tras ellas el número de enlaces baja rápidamente, existiendo únicamente 12 que obtienen más del 1% de enlaces recopilados.

En relación con los enlaces entrantes a patentes del Bloque Patent Outlink, debido a la falta de datos ya que únicamente se encuentra 8.279 el análisis es inconcluso. Para realizar un análisis completamente circular en la presente tesis se debería haber realizado una búsqueda de enlaces para todos los documentos del Bloque Patent Outlink, pero debido a la falta de recursos económicos para acceder a una cuenta superior se opta por analizar todos los enlaces generados. Esta opción, en cualquier caso, permite comprobar la validez del método y la existencia de una masa crítica de documentos enlazantes, que quizás utilizando el sistema de análisis únicamente a documentos del Bloque Patent Outlink no hubiera podido ser observada. Por esto, se considera aconsejable realizar un análisis centrado en las patentes del Bloque Patent Outlink.

El análisis por categoría muestra que la gran mayoría de los enlaces a patentes se realizan desde Servicios (55%) o Empresas (40%), algo que coincide con los resultados del Bloque Patent Outlink. Aunque en este caso los resultados de enlaces desde Universidades y Organizaciones (representan el 1% de los enlaces) quedan más alejados de los obtenidos para el Bloque Patent Outlink (19%), esto puede ser debido a que las Universidades albergan contenido de interés para informar los documentos de patentes, y por el contrario las universidades no otorgan difusión a las patentes, esto será estudiado en futuros trabajos.

El análisis por TLD muestra de nuevo la alineación de los resultados con la categorización de los dominios, donde .com es el TLD con mayor número de enlaces a patentes (27,8%). Además, ocupando en segundo lugar .io (22%) que aparece en puesto 21 el Bloque Patent Outlink (0,13%), dominio utilizado por empresas tecnológicas. Las Universidades y Gobiernos en cambio no encuentran una reciprocidad con los TLDs del Bloque Patent Outlink –igual que sucede con la categorización– ya que únicamente se obtienen para .edu 2.297 enlaces (comparados con 138.052 del Bloque Patent Outlink). Incluso en el análisis de segundo nivel (SLDs) la representación de Educación y Académica es mucho menor (Bloque Patent Outlink 6.035 enlaces, Bloque Patent Inlink 484 enlaces).

Majestic proporciona datos relativos a los idiomas en los que se han redactado los contenidos de las páginas que generan los enlaces. Gracias a esta funcionalidad, se ha podido constatar que el inglés (64,76%) es el más utilizado, resultado esperado debido tanto a que las propias patentes como el país inicial de análisis es angloparlante. Seguido de alemán, español y ruso. Sería interesante realizar un análisis de las páginas de estos idiomas para categorizarlos y comprobar si existen diferencias por países (por ejemplo, si desde Alemania o España enlazan más universidades que desde Estados Unidos).

Se debe tener en cuenta que existe una gran cantidad de dominios cuyo idioma ha sido catalogado como “Desconocido”, así como la existencia de discrepancias en algunos de los idiomas; por ejemplo, según TLD de China se recogen 11.964 enlaces, pero Majestic únicamente cataloga 1.555. Esto no quiere decir que no sea un dato correcto, es posible que, pese a ser el TLD de China el contenido se encuentre en inglés, pero debería estudiarse en detalle.

Por último, la herramienta ofrece dos indicadores que permiten conocer más aspectos relativos a los dominios que enlazan a patentes: Citation Flow y Trust Flow. Los resultados ofrecidos por estos indicadores muestran una alta correlación (0,849; Spearman; alpha < 0,05), así como que los sitios web que enlazan a patentes no obtienen datos elevados de visibilidad y autoridad web.

Algunos de los dominios con el TF más alto pertenecen a la categoría Universidad o Gobierno (Smithsonian, Harvard o Library of Congress), conteniendo Empresas y Servicios, ambos altamente ligados con la Tecnología (Github, Mozilla o Linkedin), datos que se corresponden con lo visto anteriormente.

Se ha recopilado mucha más información que podría ayudar a saber si los recursos web que enlazan a patentes reciben una bonificación o ayuda para posicionar mejor en los

buscadores (al emitir enlaces a sitios confiables y de calidad, como son las patentes). Este aspecto se estudiará en detalle en futuros trabajos de investigación.

5.2. Validez del modelo

Como se ha podido observar, el campo de las patentes es de una alta complejidad, no únicamente por las diferencias que existen entre los propios sistemas de protección, como en la forma de ofrecer y consumir los datos relativos a los documentos de patentes.

Existen multitud de bases de datos, tanto de las propias Oficinas como de instituciones no gubernamentales o empresas privadas, y cada una de ellas ofrece los datos y el acceso a la información de manera diferente. Además, la información contextualizada de los propios documentos o su uso en internet resulta complejo de estudiar debido a la falta de herramientas que permitan rastrear el contenido completo de la Web.

La necesidad de plantear un modelo que permita realizar el estudio y análisis de modo complementario de ambos bloques queda patente cuando se proyecta la necesidad de conocer más, tanto sobre los enlaces contenidos en los documentos de patentes (y su complejo acceso a la información), como los enlaces dirigidos a documentos de patentes (en un universo web del que se desconoce la localización el 95% de los documentos).

Al aplicar el modelo, siendo este doble y complementario, se puede realizar un análisis diferenciado para localizar las patentes más enlazadas, así como recursos de calidad que las enlacen; y los propios recursos de calidad enlazados desde los documentos de patente.

5.2.1. Validez del modelo para el Bloque Patent Outlink

Realizando un análisis crítico del Bloque Patent Outlink (enlaces salientes), utilizar la información contenida en el propio documento original de la patente permite asegurar que no se pierden datos ni quedan porciones de análisis a realizar tras muros de pago o errores de acceso, como podría suceder utilizando fuentes de datos propietarias. Aunque supone un problema técnico, ya que como se ha visto, los propios documentos de patentes (al menos los ofrecidos por la USPTO) no cuentan con un formato correcto para la lectura y extracción de enlaces de forma automática, lo que añade la necesidad de realizar una limpieza importante en los datos de modo semiautomático.

Dado que la W3C propone y promueve el uso de etiquetas identificadoras <XLink> en XML (equivalente a la etiqueta <a> en HTML), bastaría con utilizar este sistema para evitar problemas de extracción en los datos y así como la necesidad de uso de los dos RegEx propuestos en el modelo ya que con una única ejecución se podrían detectar y extraer todos los enlaces.

Para tomas de datos mucho más pequeñas en las que la posibilidad de error de acceso sea asumible, se podría utilizar un sistema basado en recogida desde portal web. Fuentes de datos como Google Patents o Lens.org al ofrecer el texto completo de los documentos, para algunos de ellos generan una lectura de los enlaces, transformándolos a formato HTML; por lo tanto, realizando una búsqueda de las etiquetas de enlaces <a> se podría acceder a las URLs sin necesidad del uso de dos RegEx. Aunque, incluso en estas fuentes, la transformación a enlace no se realiza para todos los documentos (ni para todos los enlaces contenidos en el documento), por lo que la posibilidad de necesitar el uso de al menos un RegEx debería ser tenido en cuenta.

Debido a los errores que se pasa a comentar a continuación, no se recomienda el uso de otras fuentes de datos como documentos cerrados (i.e.: pdf) ya que la lectura del contenido puede incrementar considerablemente el número de errores debido a la necesidad de realizar pasos intermedios para el acceso a la información que contienen.

Una vez recogidas las URLs se debe realizar un proceso de limpieza semiautomático ya que existen errores de transcripción. Estos errores se dan en el momento de generar el documento en la propia Oficina; pueden deberse a errores tipográficos realizados en el momento de redactar el documento (i.e.: inserción de dos puntos: url..com), o errores generados en el proceso de transformación del documento a formato digital (i.e.: w se digitaliza como vv).

Independientemente del motivo por el que se generen estos errores deben ser tenidos en cuenta antes de analizar las URLs ya que, en el momento de contabilizar unificando por valor de URL, no se agregarían correctamente al no tener los mismos caracteres.

Para la presente tesis se realizó un estudio manual de los primeros años (2008-2010) y se asumió un error menor al 5% para los restantes, pero este valor puede ser diferente en otras Oficinas, así como en otros años, ya que con el tiempo las solicitudes se realizan cada vez más de forma telemática (existen Oficinas que reducen las tasas para que se realicen los trámites online) por lo que se eliminarían los errores de transcripción. Asimismo, se entiende que el porcentaje de error podría aumentar para años anteriores debido a la falta de digitalización en las Oficinas.

Del mismo modo, en la presente tesis no se ha realizado un estudio a nivel de subdominio debido a que se pretende comprobar la viabilidad del modelo que requería de un gran volumen de URLs a analizar, por lo que la existencia de errores tipográficos y de transcripción comentados anteriormente dificultaban al análisis y aumentaban el porcentaje de error. Una vez comprobada la validez del modelo y sentadas las bases del Patent Link Analysis, se propone realizar en trabajos futuros un estudio en detalle teniendo en cuenta las URLs completas.

Una ventaja del análisis mediante la descarga de ficheros propuesto es la de evitar la estacionalidad o desaparición de los datos. Al tratarse de documentos estables, los estudios pueden repetirse o realizarse en cualquier momento ya que los datos son invariables.

5.2.2. Validez del modelo para el Bloque Patent Inlink

Así como el Bloque Patent Outlink existen diversas posibilidades para acceder a los datos, cada una de ellas con ventajas e inconvenientes. En el caso del Bloque Patent Inlink (enlaces entrantes) las opciones son más limitadas ya que todas ellas pasan por el uso de herramientas de terceros, la creación de una herramienta desarrollada ad-hoc que permita el escaneo de la web completa requiere de una inversión, principalmente temporal pero no se debe descartar económica, importante.

La utilización de herramientas que permitan el rastreo de información en la web, como los motores de búsqueda, limitan el acceso masivo a la información, por lo que, para evitar problemas o bloqueos, así como para asegurar la consecución de resultados independientemente del Bloque Patent Outlink, se opta por utilizar Majestic para la extracción de la información.

Esto supone que se debe pagar por los datos a utilizar, lo que puede limitar ciertos estudios. Además, como se ha explicado en el Capitulo 3, en la presente tesis se utiliza la descarga de información de tipo Fresh Index, ya que permite recoger un gran volumen de enlaces, pero no seleccionar aquellos que se desean estudiar. Por lo que en caso de querer acceder a ciertos enlaces preseleccionados (aquellos relativos a las patentes del Bloque Patent Outlink) o todos los recogidos en la base de datos (Historic Index), se deberá acceder a una cuenta superior. Existiendo de este modo una dependencia en la accesibilidad a recursos de pago que pueden, además, modificar sus condiciones en el tiempo.

En este Bloque, si que existe un problema de estacionalidad de los datos, ya que en la Web los documentos, y por ende los enlaces, son muy volátiles, lo que conlleva que un análisis pueda llegar a ser irrepetible.

Por último, y aplicable a ambos bloques, se debe tener en cuenta que, para estudios con un gran volumen de patentes a analizar, se debe contar con equipamiento informático preparado para ello. Ya que, tanto a nivel de almacenamiento, como a nivel de ejecución, se trabaja con una gran cantidad de gigas de datos. Esto limita a su vez las herramientas a utilizar ya que no todas pueden trabajar con datos masivamente. Por ejemplo, el límite de filas en Excel se encuentra en 1.040.000, en Google Sheets en 5 millones de celdas. Y otras herramientas como R y OpenRefine requieren de modificaciones en la asignación de memoria para poder trabajar.

Capítulo 6
Conclusiones

En el presente capítulo se recogen las conclusiones generales para los objetivos planteados.

Objetivo Principal

Diseñar, aplicar y validar un método orientado a la identificación de recursos de información web de calidad a través de la técnica de análisis de enlaces aplicada a patentes.

- → Pese a ser técnicamente viable, es complejo y no se encuentra exento de dificultades y limitaciones (dependencia de las fuentes, problemas en la extracción de datos, operación de datos masiva, etc.).
- → Los resultados obtenidos se consideran satisfactorios, ya que permiten identificar los recursos enlazados/enlazantes, y se ha comprobado cómo, especialmente en el caso de los recursos enlazados, son de calidad.
- → Es un modelo compuesto que permite analizar enlaces entrantes y salientes (inlinks y outlinks) mediante dos bloques de análisis que se complementan.
- → Ambos bloques de análisis pueden ser utilizados independientemente, pero su uso complementario ofrece una mayor potencia, fiabilidad y precisión.
- → El modelo de análisis de enlaces y patentes es posible, tal y como ha demostrado el diseño, aplicación y análisis descritos en la presente tesis.

Objetivos Específicos:

Objetivos Bloque Patent Outlink

Determinar la viabilidad del uso de patentes para la identificación de recursos online de calidad.

- → Los resultados indican que los enlaces si son utilizados para informar los documentos de patentes y ofreciendo contexto a las explicaciones aportadas (gracias a su utilización en diferentes secciones de los documentos de patente)

- → El número de enlaces por patente es todavía bajo (2 enlaces de media), pero está subiendo (la mediana pasa de 4 a 5 los años 2017 y 2018), y dada la cantidad de patentes (3.133.247), el número total de enlaces (2.297.366) es muy elevado.

- → El modelo propuesto facilita y permite la recopilación de enlaces existente hasta el momento.

- → Existe una amplia cantidad de sitios web únicos enlazados, destacando las empresas públicas y privadas, cuya presencia se incrementa con los años.

- → Existe una gran variedad de TLD, siendo el más utilizado .com, siendo relevantes los dominios relativos a Universidades y Gobierno (.edu.y .gov).

- → El análisis evolutivo muestra que los enlaces son cada vez más utilizados como objetos informacionales dentro de los documentos de patentes. Existe un incremento anual constante tanto en número totales (volumen de enlaces) como en parciales (diversidad de dominios, tipología de documentos y secciones)

- → Existe una diferencia clara de uso de enlaces entre áreas de conocimiento, generándose tres grupos (uso elevado, medio y bajo). Destacando el área de Física, con una gran cantidad de enlaces recopilados en el área de Computación y Cálculos. Quedando patente el cariz tecnológico de los enlaces.

Objetivos Bloque Patent Inlink

Caracterizar el impacto de la patente, como objeto informacional en Internet (*patent as web genre*)

- → Existe una gran cantidad de enlaces dirigidos a documentos de patentes realizados desde dominios altamente relacionados con la Empresa y la Tecnología demostrando el interés que generan.

- → Pese a existir una gran cantidad de enlaces recogidos, la cantidad de dominios que los generan es mucho más baja (17.001). Existiendo, además, una alta concentración en los datos, ya que unos pocos dominios generan una gran parte de los enlaces.

- → Los TLD muestran una alta relación con empresas y tecnología debido al uso de .com y .io como TLDs con mayor cantidad de enlaces. Existiendo relación con la

temática de las áreas que generan los enlaces y los dominios enlazados en el Bloque Patent Outlink.

→ En contraposición con el Bloque Patent Outlink, las Universidades y Gobiernos generan muy pocos enlaces a documentos de patentes

→ La calidad de los enlaces dirigidos a patentes es variada, generalmente baja, con un valor promedio de Citation Flow de 24,21 y Trust Flow de 15,14 (ambos sobre 100). Pero con dominios de alto nivel.

→ El Bloque Patent Inlink no identifica, en general, tan buenos recursos de calidad como el Bloque Patent Outlink, pero si permite detectar un volumen menor con una calidad notablemente mayor.

A modo de conclusión final, la presente tesis ha realizado el análisis de más de tres millones de patentes y más de cinco millones de enlaces entrantes y salientes a documentos de patentes. Este análisis ha permitido por una parte evidenciar una masa crítica significativa de enlaces para ser analizados estadísticamente (lo que abre un campo de estudio y aplicación de la Cibermetría) y, por otra parte, comprobar la viabilidad técnica de este análisis (en tiempo y precisión), un aspecto que no se había realizado hasta la fecha a este nivel de detalle.

Esta técnica, denominada en este trabajo doctoral como Patent Link Analysis, permite avanzar en el conocimiento acerca de la utilización de los hiperenlaces (en este caso en el contexto de una patente, con las implicaciones que ello supone debido a la naturaleza y funciones de estos documentos) así como evidenciar su utilización a la hora de identificar recursos web de calidad académicos.

El Patent Link Analysis puede ser utilizado en la realización de estudios y análisis de visibilidad e impacto web, así como en la generación de sistemas de monitorización basados en enlaces web para diversos ámbitos, tanto académicos (impacto de recursos web), como empresariales (vigilancia tecnológica, redes de información tecnológica, etc.) y sociales (estudio de la emergencia y obsolescencia de tecnologías en el tiempo, interés de los medios y de la Sociedad, etc.).

Por último, se plantean líneas de investigación futuras que complementan y permiten profundizar en los resultados obtenidos, relacionadas con la influencia de las patentes en el posicionamiento web de recursos (a través de los enlaces entrantes y salientes en documentos de patente), así como el diseño de indicadores orientados a la evaluación (tanto del impacto de patentes como del impacto de los recursos web que las enlazan). Así mismo, este estudio podría ser ampliado mediante la incorporación de otros sistemas de concesión de patentes (especialmente aquellos regionales como la EPO o WIPO) y el análisis de plataformas de redes sociales (para estudiar distintas comunidades de atención, conversación e interés en torno a las patentes).

Capítulo 7
Referencias bibliográficas

Adams, J. N. (2019). History of the patent system. In *Research Handbook on Patent Law and Theory: Second Edition* (pp. 2–26). Edward Elgar Publishing Ltd. https://doi.org/10.4337/9781785364129.00009

Agrawal, A. K., & Henderson, R. (2009). *Reprinted Article Putting patents in context: Exploring knowledge transfer from MIT*. *Advances in Strategic Management* (Vol. 26). Elsevier. https://doi.org/10.1108/S0742-3322(2009)0000026033

Aguillo, I. F. (2009). Cibermetría: introducción teórico-práctica.

Alonso Arroyo, A. (2004). *Producción científica de la Universidad Politécnica de Valencia (1973-2001): análisis bibliométrico*. Universidad Politécnica de Valencia.

Altuntas, S., Dereli, T., & Kusiak, A. (2015). Forecasting technology success based on patent data. *Technological Forecasting and Social Change*, *96*, 202–214. https://doi.org/10.1016/j.techfore.2015.03.011

Alvarez Gil, L.; Albarez Gonzalez, M.; Contreras Villavicencio, D. M. . (2016). Índices bibliométricos en base de datos de patentes para la investigación científico-técnica empresarial. In *X Concerencia Internacional de Ciencias Empresariales CICE* (p. 15).

Aon. (2019). 2019 Intangible Assets Financial Statement Impact Comparison Report. *Global Edition*, (April), 50. Retrieved from https://www.aon.com/getmedia/60fbb49a-c7a5-4027-ba98-0553b29dc89f/Ponemon-Report-V24.aspx

Archontopoulos, E. (2004). Prior art search tools on the Internet and legal status of the results: A European Patent Office perspective. *World Patent Information*, *26*(2), 113–121. https://doi.org/10.1016/j.wpi.2003.08.004

Arias, E. (2003). Fuentes de información sobre Patentes.

Aristodemou, L., & Tietze, F. (2018). Citations as a measure of technological impact: A review of forward citation-based measures. *World Patent Information*. Elsevier. https://doi.org/10.1016/j.wpi.2018.05.001

Barberá-Tomás, D., Jiménez-Sáez, F., & Castelló-Molina, I. (2011). Mapping the importance of the real world: The validity of connectivity analysis of patent citations networks. *Research Policy*, *40*(3), 473–486. https://doi.org/10.1016/j.respol.2010.11.002

Battke, B., Schmidt, T. S., Stollenwerk, S., & Hoffmann, V. H. (2016). Internal or external spillovers - Which kind of knowledge is more likely to flow within or across technologies. *Research Policy*, *45*(1), 27–41. https://doi.org/10.1016/j.respol.2015.06.014

Benson, C. L., & Magee, C. L. (2015). Quantitative determination of technological improvement from patent data. *PLoS ONE*, *10*(4), 1–23. https://doi.org/10.1371/journal.pone.0121635

Björneborn, L. (2004). *Small-World Link Structures across an Academic Web Space: A Library and Information Science Approach*. *Library*. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.84.1358&rep=rep1&type=pdf

Bojo Canales, C., Fraga Medín, C., Hernández Villegas, S., Jaén Casquero, M. B., Jiménez Planet, V., Mohedano Macías, L., & Novillo Orti, A. (2004). Internet Visible e Invisible: búsqueda y selección de recursos de información en Ciencias de la Salud. *Biblioteca Nacional de Ciencias de La Salud*. https://doi.org/10.1016/S0213-9111(05)71406-0

Campbell, E. G., Powers, J. B., Blumenthal, D., & Biles, B. (2004). Inside the triple helix: Technology transfer and commercialization in the life sciences. *Health Affairs*, *23*(1), 64–76. https://doi.org/10.1377/hlthaff.23.1.64

Campbell, R. S. (1983). Patent trends as a technological forecasting tool. *World Patent Information*, *5*(3), 137–143. https://doi.org/10.1016/0172-2190(83)90134-5

Casado-Serviño, A., & Sanz-Martínez, J. M. (2013). *200 años de patentes*. Madrid, España.

Chen, H., Zhang, G., Zhu, D., & Lu, J. (2017). Topic-based technological forecasting based on patent data: A case study of Australian patents from 2000 to 2014. *Technological Forecasting and Social Change*, *119*, 39–52. https://doi.org/10.1016/j.techfore.2017.03.009

Chen, Y. S., Shih, C. Y., & Chang, C. H. (2014). Explore the new relationship between patents and market value: A panel smooth transition regression (PSTR) approach. *Scientometrics*, *98*(2), 1145–1159. https://doi.org/10.1007/s11192-013-1110-9

Criscuolo, P., & Verspagen, B. (2008). Does it matter where patent citations come from? Inventor vs. examiner citations in European patents. *Research Policy*, *37*(10), 1892–1908. https://doi.org/10.1016/j.respol.2008.07.011

Czarnitzki, D., Hussinger, K., & Schneider, C. (2012). The nexus between science and industry: Evidence from faculty inventions. *Journal of Technology Transfer*, *37*(5), 755–776. https://doi.org/10.1007/s10961-011-9214-y

de Rassenfosse, G., Dernis, H., & Boedt, G. (2014). An Introduction to the patstat database with example queries. *Australian Economic Review*, *47*(3), 395–408. https://doi.org/10.1111/1467-8462.12073

De Rassenfosse, G., Palangkaraya, A., & Webster, E. (2016). Why do patents facilitate trade in technology? Testing the disclosure and appropriation effects. *Research Policy*, *45*(7), 1326–1336. https://doi.org/10.1016/j.respol.2016.03.017

de Solla Price, D. J. (1976). A General Theory of Bibliometric and Other Cumulative Advantage Processes. *Journal of the American Society for Information Science*, *27*(5–6), 292–306.

EPO, E. P. O. (2019). *Patent Index 2019*.

Etzkowitz, H., & Leydesdorff, L. (1995). The Triple Helix -- University - Industry - Government Relations: a Laboratory for Knowledge Based Economic Development. *EASST Review*, *14*(1), 14–19.

EUIPO. (2019). *2019 INTELLECTUAL PROPERTY SME SCOREBOARD 2019 INTELLECTUAL PROPERTY Commissioned by*. https://doi.org/10.2814/294170

Font-Julian, C. I., Ontalba-Ruipérez, J. A., & Orduña-Malea, E. (2018). Hit count estimate variability for website-specific queries in search engines: The case for rare disease association websites. *Aslib Journal of Information Management*, *70*(2), 192–213. https://doi.org/10.1108/AJIM-10-2017-0226

Frumkin, M. (1945). The origin of patents. *Journal of the Patent Office Society*, *27*(3).

Gifford, D. J. (2004). How Do the Social Benefits and Costs of the Patent System Stack Up in Pharmaceuticals? *J. Intell. Prop. L*, *75*. Retrieved from https://scholarship.law.umn.edu/faculty_articlesathttps://scholarship.law.umn.edu/faculty_articles/348.

Giménez, G. (2018). The impact of the patent system on the social welfare: A critical view. *Intangible Capital*, *14*(2), 253–269. https://doi.org/10.3926/ic.789

Gkoumas, K., & Christou, M. (2020). A triple-helix approach for the assessment of

hyperloop potential in Europe. *Sustainability (Switzerland)*, *12*(19), 1–20. https://doi.org/10.3390/SU12197868

Glänzel, W., Moed, H. F., Schmoch, U., & Thelwall, M. (2019). *Springer Handbook of Science and Technology Indicators*. https://doi.org/10.1007/978-3-030-02511-3

Han, E. J., & Sohn, S. Y. (2015). Patent valuation based on text mining and survival analysis. *Journal of Technology Transfer*, *40*(5), 821–839. https://doi.org/10.1007/s10961-014-9367-6

Han, E. S., & Goleman, D.; boyatzis, R.; Mckee, A. (2008). *Web Search: Multidisciplinary Perspectives*. *Information Science and Knowledge Management*.

Harhoff, D., Narin, F., Scherer, F. M., & Vopel, K. (1999). Citation frequency and the value of patented inventions. *Review of Economics and Statistics*, *81*(3), 511–515. https://doi.org/10.1162/003465399558265

Haustein, S. (2015). Scholarly communication and evaluation : from bibliometrics to altmetrics. *COAR-SPARC Conference 2015*. Retrieved from http://www.slideshare.net/StefanieHaustein/haustein-coar-sparcapril2015

Hegde, D., & Sampat, B. (2009). Examiner citations, applicant citations, and the private value of patents. *Economics Letters*, *105*(3), 287–289. https://doi.org/10.1016/j.econlet.2009.08.019

Iffat, R., & Sami, L. K. (2010). Understanding the deep web. *Library Philosophy and Practice*, *2010*(MAY), 1–5.

International Telecommunication Union. (2020). *Measuring digital development Facts and figures 2020*. *ITUPublications*. Retrieved from https://www.itu.int/en/mediacentre/Documents/MediaRelations/ITU Facts and Figures 2019 - Embargoed 5 November 1200 CET.pdf

Iversen, E. J. (2000). An excursion into the patent-bibliometrics of Norwegian patenting. *Scientometrics*, *49*(1), 63–80. https://doi.org/10.1023/A:1005609224740

Jensen, P. H., Palangkaraya, A., & Webster, E. (2015). Trust and the market for technology. *Research Policy*, *44*(2), 340–356. https://doi.org/10.1016/j.respol.2014.10.001

Kang, K., & Sohn, S. Y. (2016). Evaluating the patenting activities of pharmaceutical research organizations based on new technology indices. *Journal of Informetrics*, *10*(1), 74–81. https://doi.org/10.1016/j.joi.2015.10.006

Kani, M., & Motohashi, K. (2012). Understanding the technology market for patents: New insights from a licensing survey of Japanese firms. *Research Policy*, *41*(1), 226–235. https://doi.org/10.1016/j.respol.2011.08.002

Karvonen, M., & Kässi, T. (2013). Patent citations as a tool for analysing the early stages

of convergence. *Technological Forecasting and Social Change*, *80*(6), 1094–1107. https://doi.org/10.1016/j.techfore.2012.05.006

Karytinos, A., & Ingham, A. (2015). A growing interest for intellectual property in universities. *Pharmaceutical Patent Analyst*, *4*(2), 59–61. https://doi.org/10.4155/ppa.14.57

Kaya Firat, A., Madnick, S., & Lee Woon, W. (2008). Technological forecasting–A review.Massachusetts Institute of Technology, Cambridge, USA. *Working Paper CISL*, *15*(September). Retrieved from https://pdfs.semanticscholar.org/8ea2/bd1792cf794506966ecaacb2e3315de1fc5a.pdf?_ga=2.88571091.913885628.1546016906-495708997.1535723386%0Ahttp://web.mit.edu/smadnick/www/wp/2008-15.pdf

Khasseh, A. A., Soheili, F., Moghaddam, H. S., & Chelak, A. M. (2017). Intellectual structure of knowledge in iMetrics: A co-word analysis. *Information Processing and Management*, *53*(3), 705–720. https://doi.org/10.1016/j.ipm.2017.02.001

Kleinrock, L. (2010). An early history of the Internet. *IEEE Communications Magazine*, (August), 26–36.

Kousha, K., & Thelwall, M. (2017). Patent citation analysis with Google. *Journal of the Association for Information Science and Technology*, *68*(1), 48–61. https://doi.org/10.1002/asi.23608

Kurzweil, R. (2004). The Law of Accelerating Returns. *Alan Turing: Life and Legacy of a Great Thinker*, 381–416. https://doi.org/10.1007/978-3-662-05642-4_16

Langinier, C., & Moschini, G. (2002). The Economics of Patents: An Overview. *CARD Working Papers*, (December), 13–14. Retrieved from http://lib.dr.iastate.edu/card_workingpapers/335

Lee, Sanghoon, & Kim, W. (2017). The knowledge network dynamics in a mobile ecosystem: a patent citation analysis. *Scientometrics*, *111*(2), 717–742. https://doi.org/10.1007/s11192-017-2270-9

Lee, Sungjoo, Yoon, B., & Park, Y. (2009). An approach to discovering new technology opportunities: Keyword-based patent map approach. *Technovation*, *29*, 481–497. https://doi.org/10.1016/j.technovation.2008.10.006

Leiner, B. M., Cerf, V. G., Clark, D. D., Kahn, R. E., Kleinrock, L., Lynch, D. C., ... & Wolff, S. (2009). A brief history of the Internet. *ACM SIGCOMM Computer Communication Review*, *39*(5), 22–31.

Leydesdorff, L.; Milojevic, S. . (2012). *Scientometrics* (No. preprint arXiv:1208.4566).

Leydesdorff, L. (2012). Triple Helix of University-Industry-Government Relations. *Encyclopedia of Creativity, Invention, Innovation and Entrepreneurship*, (February), 2356–2364. https://doi.org/10.1007/978-3-319-15347-6_452

Li, X., Xie, Q., Jiang, J., Zhou, Y., & Huang, L. (2018). Identifying and monitoring the development trends of emerging technologies using patent analysis and Twitter data mining: The case of perovskite solar cell technology. *Technological Forecasting and Social Change*, (June), 0–1. https://doi.org/10.1016/j.techfore.2018.06.004

López Jiménez, D., & Dittmar, E. C. (2019). Triple Helix Model of Innovation: University, Industry and Government Interactions. *EDULEARN19 Proceedings*, *1*(July), 2570–2574. https://doi.org/10.21125/edulearn.2019.0705

Maltseva, D., & Batagelj, V. (2020). *iMetrics: the development of the discipline with many names*. *Scientometrics* (Vol. 125). Springer International Publishing. https://doi.org/10.1007/s11192-020-03604-4

Manglano, B. G.-A., & Zulueta, M. Á. (2008). Estudio comparativo de bases de datos de patentes en internet. *Anales de Documentación*, *10*(0), 145–162.

McDonald, M. K. (2015). *The Social Impact of Intellectual Property Rights: Public health, Education, and Income inequality*. University of Maryland. Retrieved from http://library1.nida.ac.th/termpaper6/sd/2554/19755.pdf

Meyer, M. (2000a). Does science push technology? Patents citing scientific literature. *Research Policy*, *29*(3), 409–434. https://doi.org/10.1016/S0048-7333(99)00040-2

Meyer, M. (2000b). What is Special about Patent Citations. Differences between Scientific and Patent Citations.

Meyer, M., Siniläinen, T., & Utecht, J. T. (2003). Towards hybrid triple helix indicators: A study of university-related patents and a survey of academic inventors. *Scientometrics*, *58*(2), 321–350. https://doi.org/10.1023/A:1026240727851

Michel, J., & Bettels, B. (2001). Patent citation analysis: A closer look at the basic input data from patent search reports. *Scientometrics*, *51*(1), 185–201. https://doi.org/10.1023/A:1010577030871

Milojević, S., & Leydesdorff, L. (2013). Information metrics (iMetrics): A research specialty with a socio-cognitive identity? *Scientometrics*, *95*(1), 141–157. https://doi.org/10.1007/s11192-012-0861-z

Moore, G. E. (1965). Cramming more components onto integrated circuits. *IEEE Solid-State Circuits Society Newsletter*, *38*(8), 114–117. https://doi.org/10.1109/n-ssc.2006.4785860

Mowshowitz, A., & Kawaguchi, A. (2005). Measuring search engine bias. *Information Processing and Management*, *41*(5), 1193–1205. https://doi.org/10.1016/j.ipm.2004.05.005

Nard, C. A., & Morriss, A. P. (2006). Constitutionalizing patents: From Venice to Philadelphia. *Review of Law and Economics*, *2*(2). https://doi.org/10.2202/1555-5879.1054

Narin, F. (1994). Patent bibliometrics. *Scientometrics*, *30*(1), 147–155. https://doi.org/10.1007/BF02017219

Nemet, G. F., & Johnson, E. (2012). Do important inventions benefit from knowledge originating in other technological domains? *Research Policy*, *41*(1), 190–200. https://doi.org/10.1016/j.respol.2011.08.009

Noh, H., Song, Y. K., & Lee, S. (2016). Identifying emerging core technologies for the future: Case study of patents published by leading telecommunication organizations. *Telecommunications Policy*, *40*(10–11), 956–970. https://doi.org/10.1016/j.telpol.2016.04.003

Oehmke, J. F. (2006). *The Social Welfare Implications of Intellectual Property Protection in Regulating Agricultural Biotechnology: Economics and Policy*. Retrieved from http://library1.nida.ac.th/termpaper6/sd/2554/19755.pdf

OMPI. (2016a). Principios Básicos de la Propiedad Industrial. *Organizacion Mundial de La Propiedad Intelectual*.

OMPI. (2016b). *Principios básicos del derecho de autor y los derechos conexos*. *Ompi* (Vol. 909(S)).

Orduña-Malea, E.; Aguillo, I. F. (2014). *Cibermetría: Midiendo el espacio red*.

Orduña-Malea, E.; Alonso-Arroyo, A. (2017). *Cybermetric Techniques to Evaluate Organizations Using Web-based Data*. Elsevier.

Orduña-malea, E. (2012). Fuentes de enlaces web para análisis cibermétricos. *Anuario ThinkEPI*, *6*(1), 276–280.

Orduña-Malea, E., Thelwall, M., & Kousha, K. (2016). Web citations in patents: Evidence of technological impact? *Journal of the Association for Information Science and Technology*, *68*(8), 1967–1974. https://doi.org/10.1002/asi.23821

Paden, R. (2001). The two professions of Hippodamus of Miletus. *Philosophy & Geography*, *4*(1), 25–48. https://doi.org/10.1080/10903770124644

Park, W. G. (2008). International patent protection: 1960-2005. *Research Policy*, *37*(4), 761–766. https://doi.org/10.1016/j.respol.2008.01.006

Prager, F. D. (1950). The early growth and influence of Intelectual Property. *Journal of the Patent Office Society*, *34*(2), 106–140.

Priem, J., Taraborelli, D., Groth, P., & Neylon, C. (2011). altmetrics: a manifesto.

Raturi, M. K., Sahoo, P. K., & Tiwari, A. K. (2012). Delphion: A World Class Patent Database - A Comprehensive Analysis from Patent Information Professional's Perspective. *SSRN Electronic Journal*, 1–11. https://doi.org/10.2139/ssrn.1510630

Rutter, N. K. (1970). Sybaris-legend and reality. *Greece and Rome*, *17*(2). https://doi.org/10.1017/S0017383500017836

Sampat, B. (2018). A survey of Empirical Evidence on Patents and Innovation. *Natural Bureau of Economic Research*.

Sánchez, A., Hortal, P., & Cuesta, D. (2015). *Patent costs and impact on innovation*.

Sarin, S., Haon, C., Belkhouja, M., Mas-Tur, A., Roig-Tierno, N., Sego, T., … Carley, S. (2020). Uncovering the knowledge flows and intellectual structures of research in Technological Forecasting and Social Change: A journey through history. *Technological Forecasting and Social Change*, *160*(October 2019), 120210. https://doi.org/10.1016/j.techfore.2020.120210

Segev, A., & Kantola, J. (2012). Identification of trends from patents using self-organizing maps. *Expert Systems with Applications*, *39*(18), 13235–13242. https://doi.org/10.1016/j.eswa.2012.05.078

Seymour, T., Frantsvog, D., & Kumar, S. (2011). History Of Search Engines. *International Journal of Management & Information Systems (IJMIS)*, *15*(4), 47. https://doi.org/10.19030/ijmis.v15i4.5799

Singh, V., Chakraborty, K., & Vincent, L. (2016). Patent database: Their importance in prior art documentation and patent search. *Journal of Intellectual Property Rights*, *21*(1), 42–56.

Sterzi, V. (2013). Patent quality and ownership: An analysis of UK faculty patenting. *Research Policy*, *42*(2), 564–576. https://doi.org/10.1016/j.respol.2012.07.010

Thelwall, M. (2009). *Introduction to Webometrics: Quantitative Web Research for the Social Sciences*. *Synthesis Lectures on Information Concepts, Retrieval, and Services* (Vol. 1). https://doi.org/10.2200/s00176ed1v01y200903icr004

Torres-Salinas, D., Cabezas-Clavijo, Á., & Jiménez-Contreras, E. (2013). Altmetrics: New indicators for scientific communication in web 2.0. *Comunicar*, *21*(41), 53–60. https://doi.org/10.3916/C41-2013-05

Trajtenberg, M. (2006). A Penny for Your Quotes: Patent Citations and the Value of Innovations. *The RAND Journal of Economics*, *21*(1), 172. https://doi.org/10.2307/2555502

Trappey, A. J. C., Trappey, C. V., Wu, C. Y., & Lin, C. W. (2012). A patent quality analysis for innovative technology and product development. *Advanced Engineering Informatics*, *26*(1), 26–34. https://doi.org/10.1016/j.aei.2011.06.005

Tseng, Y. H., Lin, C. J., & Lin, Y. I. (2007). Text mining techniques for patent analysis. *Information Processing and Management*, *43*(5), 1216–1247. https://doi.org/10.1016/j.ipm.2006.11.011

Vaughan, L., & Thelwall, M. (2004). Search engine coverage bias: Evidence and possible causes. *Information Processing and Management*, *40*(4), 693–707. https://doi.org/10.1016/S0306-4573(03)00063-3

Venugopalan, S., & Rai, V. (2015). Topic based classification and pattern identification in patents. *Technological Forecasting and Social Change*, *94*, 236–250. https://doi.org/10.1016/j.techfore.2014.10.006

Wang, S., Lei, Z., & Lee, W.-C. (2014). Exploring Legal Patent Citations for Patent Valuation. https://doi.org/10.1145/2661829.2662029

White, M. (2010). Patent Searching: Back to the Future How to Use Patent Classification Search Tools to Create Better Searches. *Proceedings of the Canadian Engineering Education Association (CEEA)*, 1–6. https://doi.org/10.24908/pceea.v0i0.3155

WIPO. (2010). *World Intellectual Property Indicators 2010*. *World Intellectual Property Organization* (Vol. 1). Retrieved from http://www.wipo.int/export/sites/www/freepublications/en/intproperty/941/wipo_pub_941_2013.pdf

WIPO. (2012). Guía Para Bases De Datos Tecnológicas, 98. Retrieved from http://www.wipo.int/edocs/pubdocs/es/patents/434/wipo_pub_l434_11.pdf

Witty, M. (2017). Athenaeus describes the most ancient intellectual property. *Prometheus (United Kingdom)*, *35*(2). https://doi.org/10.1080/08109028.2018.1443619

Wolfram, D. (2000). Applications of informetrics to information retrieval research. *Informing Science*, *3*(2), 77–82. https://doi.org/10.28945/581

Yang, G. C., Li, G., Li, C. Y., Zhao, Y. H., Zhang, J., Liu, T., … Huang, M. H. (2015). Using the comprehensive patent citation network (CPC) to evaluate patent value. *Scientometrics*, *105*(3), 1319–1346. https://doi.org/10.1007/s11192-015-1763-7

Ye, F. Y., Huang, M. H., & Chen, D. Z. (2016). Comparative study of trace metrics between bibliometrics and patentometrics. *Journal of Data and Information Science*, *1*(2), 13–31. https://doi.org/10.20309/jdis.201611

Yoon, B., & Park, Y. (2004). A text-mining-based patent network: Analytical tool for high-technology trend. *Journal of High Technology Management Research*, *15*(1), 37–50. https://doi.org/10.1016/j.hitech.2003.09.003

Zhou, X., Zhang, Y., Porter, A. L., Guo, Y., & Zhu, D. (2014). A patent analysis method to trace technology evolutionary pathways. *Scientometrics*, *100*(3), 705–721. https://doi.org/10.1007/s11192-014-1317-4

Anexo I

Tabla 50: Recopilación de URL de acceso a las bases de datos de patentes Nacionales o Regionales

Oficina	URL de acceso a Base de datos sobre patentes
Alemania	https://depatisnet.dpma.de/DepatisNet/
Australia	http://pericles.ipaustralia.gov.au/ols/auspat/
Austria	http://see-ip.patentamt.at/
Canadá	https://brevets-patents.ic.gc.ca/opic-cipo/cpd/eng/search/number.html
Chile	http://ion.inapi.cl:8080/Patente/ConsultaAvanzadaPatentes.aspx
China	http://www.pss-system.gov.cn/sipopublicsearch/ensearch/searchEnHomeIndexAC.do
Dinamarca	https://onlineweb.dkpto.dk/pvsonline/Patent
Eslovaquia	http://www.upv.sk/?databases-and-registers
Eslovenia	http://www2.uil-sipo.si/dse.htm
España	http://invenes.oepm.es
Estados Unidos	http://tess2.uspto.gov/bin/gate.exe?f=login&p_lang=english&p_d=trmk
Estados Unidos de América	http://www.uspto.gov/patft/index.html
Eurasia	https://www.eapo.org/en/publications/publicat/publicat.php
Finlandia	http://patent.prh.fi/patinfo/default2.asp
Francia	http://bases-brevets.inpi.fr/en
Hong Kong	https://esearch.ipd.gov.hk/nis-pos-view/#/pt/quicksearch
Hungría	http://www.hipo.gov.hu/kereso/
India	http://ipindiaservices.gov.in/publicsearch
Irlanda	https://eregister.patentsoffice.ie/query/PTQuery.aspx
Italia	https://www.uibm.gov.it/bancadati/
Japón	https://www.j-platpat.inpit.go.jp/
Latino América	http://lp.espacenet.com/
Lituania	https://vpb.lrv.lt/lt/veiklos-sritys/isradimu-patentai/patentuotu-isradimu-paieska/lietuvos-respublikos-patentu-duomenu-baze
Moldavia	http://www.db.agepi.md/inventions/Search.aspx

Noruega	https://search.patentstyret.no/
Nueva Zelanda	http://www.iponz.govt.nz/app/Extra/IP/PT/Qbe.aspx?sid=635310401987053816
Oficina de Patentes Europea	http://www.espacenet.com
Oficina para la Armonización del Mercado Internacional	https://oami.europa.eu/ohimportal/en/databases
Paises Bajos	https://mijnoctrooi.rvo.nl/fo-eregister-view/
Reino Unido	https://www.ipo.gov.uk/p-ipsum.htm
Republica Checa	http://www.upv.cz/en/client-services/online-databases/patent-and-utility-model-databases.html
Republica de Corea	http://www.kipris.or.kr/enghome/main.jsp
Rumania	http://bd.osim.ro/cgi-bin/invsearch8
Rusia	https://www.fips.ru/en/informational-resources/information-retrieval-system/databases.php
Singapur	https://www.ip2.sg/RPS/WP/CM/SearchSimpleP.aspx?SearchCategory=PT
Suiza	https://www.swissreg.ch/
Tailandia	http://203.209.117.243/DIP2013/simplesearch.php

Anexo II

Resumen del esquema XML seguido por los documentos publicados por la USPTO:

<us-patent-grant>	<us-bibliographic-data-grant>	<publication-reference>	<document-id>	<country>
				<doc-number>
				<kind>
				<date>
		<us-sir-flag sir-text=>		
		<application-reference>	<document-id>	<country>
				<doc-number>
				<date>
		<us-application-series-code>		
		<priority-claims>	<priority-claim>	<country>
				<doc-number>
				<date>
		<us-issued-on-continued-prosecution-application>		
		<rule-47-flag/>		
		<us-term-of-grant>	<prior-disclosure-affidavit-filed>	
			<lapse-of-patent>	<country>
				<doc-number>
				<date>
				<text>
			<us-term-extension>	
			<length-of-grant>	
			<disclaimer>	<text>
		<classifications-ipcr>	<classification-ipcr>	<ipc-version-indicator> <date>
				<classification-level>
				<section>
				<class>
				<subclass>
				<main-group>
				<subgroup>
				<classification-value>
				<action-date> <date>
				<generating-office> <country>
				<classification-status>
				<classification-data-source>
				<edition>

		<classification-locarno>	<main-classification>		
		<classification-national>	<country>		
			<main-classification>		
			<further-classification>		
		<invention-title>			
		<us-botanic>	<latin-name>		
			<variety>		
	<references-cited>	<citation>	<patcit>	<document-id>	<country>
					<doc-number>
					<kind>
					<name>
					<date>
			<nplcit>	<othercit>	
			<category>		
			<classification-national>	<country>	
				<main-classification>	
	<number-of-claims>				
	<us-exemplary-claim>				
	<us-field-of-classification-search>	<classification-cpc-combination-text>			
		<classification-national>	<country>		
			<main-classification>		
			<additional-info>		
		<us-classifications-ipcr>	<ipc-version-indicator>	<date>	
			<classification-level>		
			<section>		
			<class>		
			<subclass>		
			<main-group>		
			<subgroup>		
			<symbol-position>		
			<classification-value>		
			<action-date>	<date>	
			<generating-office>	<country>	
			<classification-status>		
			<classification-data-source>		
	<figures>	<number-of-drawings-sheets>			
		<number-of-figures>			

	<us-related-documents>	<related-publication>	<document-id>	<country>	
				<doc-number>	
				<kind>	
				<date>	
		<us-provisional-application>	<document-id>	<country>	
				<doc-number>	
				<kind>	
				<date>	
		<division>			
		<continuation>			
		<continuation-in-part>			
		<reissue>			
		<parent-doc>	<document-id>	<country>	
				<doc-number>	
				<kind>	
				<date>	
			<parent-status>		
			<parent-grant-document>		
			<parent-pct-document>		
		<child-doc>	<document-id>	<country>	
				<doc-number>	
				<kind>	
				<date>	
	<parties>	<applicants>	<us-applicant>	<addressbook>	<last-name>
					<first-name>
					<address>
					<city>
					<state>
					<country>
				<nationality>	<country>
				<residence>	<country>
				<us-rights>	
		<inventors>	<inventor>	<addressbook>	
				<deceased-inventor>	
		<agents>	<agent>	<addressbook>	<orgname>
					<last-name>
					<first-name>
					<address>
					<country>
		<assignees>	<assignee>	<addressbook>	<orgname>
					<role>

					<addres s>
					<city>
					<countr y>
		<examiners>	<primary-examiner>	<last-name>	
				<first-name>	
				<department >	
			<assisitant-examiner>	<last-name>	
				<first-name>	
	<pct-or-regional-filling-data>	<document-id>	<country>		
			<doc-number>		
			<kind>		
			<date>		
		<us-371c124-date>	<date>		
		<us-371c12-date>	<date>		
	<pct-or-regional-publishing-data>	<document-id>	<country>		
			<doc-number>		
			<kind>		
			<date>		
<abstract>					
<drawings>					
<description>					
<claims>					

Anexo III

Enlaces a documentos con datos brutos Bloque Outlink:

https://github.com/crifonju/PatentLinkAnalysis

Debido a las restricciones de la herramienta Majestic, los datos relativos al Bloque Inlink no se encuentran a disposición pública en internet. En caso de estar interesado en ellos puede contactar con la autora en la dirección de correo electrónico: crifonju@upv.es