



UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO
POSGRADO EN BIBLIOTECOLOGÍA Y ESTUDIOS DE LA INFORMACIÓN
FACULTAD DE FILOSOFÍA Y LETRAS
INSTITUTO DE INVESTIGACIONES BIBLIOTECOLÓGICAS Y DE LA INFORMACIÓN

*“ORGANIZACIÓN DE LA INFORMACIÓN Y METADATOS EN LA RECUPERACIÓN DE
RECURSOS ELECTRÓNICOS EN LA WEB”*

CON APOYO DEL PROGRAMA DE APOYOS PARA LA SUPERACIÓN DEL PERSONAL
ACADÉMICO (PASPA) DE LA DGAPA

T E S I S
QUE PARA OPTAR POR EL GRADO DE:
DOCTOR EN BIBLIOTECOLOGÍA Y ESTUDIOS DE LA INFORMACIÓN

P R E S E N T A

Miguel Ángel Amaya Ramírez

COMITÉ TUTORIAL

TUTOR PRINCIPAL

Dr. Filiberto Felipe Martínez Arellano. IIBI

COTUTORES

Dra, Georgina Araceli Torres Vargas . IIBI

Dr. Juan Voutssás Márquez . IIBI



Universidad Nacional
Autónoma de México



UNAM – Dirección General de Bibliotecas
Tesis Digitales
Restricciones de uso

DERECHOS RESERVADOS ©
PROHIBIDA SU REPRODUCCIÓN TOTAL O PARCIAL

Todo el material contenido en esta tesis esta protegido por la Ley Federal del Derecho de Autor (LFDA) de los Estados Unidos Mexicanos (México).

El uso de imágenes, fragmentos de videos, y demás material que sea objeto de protección de los derechos de autor, será exclusivamente para fines educativos e informativos y deberá citar la fuente donde la obtuvo mencionando el autor o autores. Cualquier uso distinto como el lucro, reproducción, edición o modificación, será perseguido y sancionado por el respectivo titular de los Derechos de Autor.

Tabla de contenido

	Págs.
Introducción	1
1. La organización de la información	7
1.1 Necesidad de organizar la información	7
1.2 Objetivos de la organización de la información	11
1.2.1 Localizar recursos de información	12
1.2.2 Identificar el objeto de información	24
1.2.3 Valorar y seleccionar el objeto de información	24
1.2.4 Obtener la información	25
1.3 Los metadatos	25
1.3.1 Definición de metadatos	28
1.3.2 Tipos de metadatos	34
1.3.3 Funciones de los metadatos	41
1.3.4 Importancia de los metadatos	48
1.3.5 Esquema de metadatos Dublin Core	49
1.4 Web semántica	51
1.4.1 Web semántica: definición	51
1.4.2 El uso del lenguaje de marcado XML (eXtensible Markup Lenguaje)	57
1.4.3 Marco de descripción de Recurso (RDF)	58
1.4.4 Ontologías	61
1.5. Características de la Web Semántica	66
1.6 Arquitectura de la Web Semántica	67
2. Recuperación de información	70
2.1 Recuperación de información	70
2.2 Herramientas de recuperación de información	71
2.2.1 Antecedentes	71
2.3 Tipología de herramientas de recuperación de información en la Web	74
2.3.1 Los buscadores	75
2.3.1.1 Directorios	76
2.3.1.2 Motores de búsqueda	78
2.3.1.3 Metabuscadors	79
2.3.2 Multibuscadores	81
2.3.3 Agentes inteligentes de búsqueda de información	82
2.3.4 Buscadores semánticos	83
2.4 Funcionamiento de las herramientas de búsqueda	84
2.5 Tendencia actual de las herramientas de búsqueda	86
2.6 Incremento de la información	88
2.6.1 Información académica	89
2.6.2 Internet Invisible	100
2.6.3 Los usuarios	105
2.6.4 Problemática para acceder a la información	109

2.6.4.1 Falta de datos suficientes para su indización	112
2.6.4.2 Inadecuado uso de estructuras de metadatos	114
2.6.4.3 El uso de protocolos inadecuados	115
3. Evaluación de herramientas de búsqueda en la Web	122
3.1 Selección	122
3.2 Fuentes utilizadas en la selección de las herramientas de búsqueda	123
3.3 Características de las herramientas de búsqueda seleccionadas	132
3.4 Estrategia de búsqueda	135
3.5 Criterios de evaluación para las herramientas de búsqueda	138
3.6 Resultados de la evaluación	141
3.6.1 Resultados de búsqueda en los motores	141
3.6.2 Resultados de la recuperación de información en Google	144
3.6.3 Resultados de la recuperación de información en Yahoo	149
3.6.4 Resultados de la recuperación de información en Bing	155
3.6.5 Resultados de la recuperación de información en Ask	162
3.6.6 Resultados de la recuperación de información en AOL	166
3.6.7 Resultados de la recuperación de información en WebCrawler	172
3.6.8 Resultados de la recuperación de información en Dogpile	177
3.6.9 Resultados de la recuperación de información en Zapmeta	183
3.6.10 Resultados de la recuperación de información en Info.com	188
3.6.11 Resultados de la recuperación de información en Exite	194
3.6.12 Resultados de la recuperación de información en Kngine	199
3.6.13 Resultados de la recuperación de información en Exalead	203
3.6.14 Resultados de la recuperación de información en DuckDuckgo	209
3.6.15 Resultados de la recuperación de información en Swoogle	215
3.7. Análisis de los resultados	219
Conclusiones	226
Referencias	229

DEDICATORIAS

A Dios por todas las bendiciones que he recibido de él a lo largo de mi vida, por las alegrías y por las tristezas vividas que me han permitido comprender que él nunca se equivoca.

Como homenaje póstumo a mi madre Guadalupe Ramírez, a mi padre Medardo Amaya, a mis suegros Laurencio Corona y Laura López.

A Alejandra Corona López mi querida esposa por ser el gran artífice de mi carrera profesional y hacer de nuestra vida algo maravilloso.

A mis hijos Miguel Ángel y Midori por ser dos jóvenes muy especiales y solo espero que sigan siendo como hasta ahora; estudiosos, soñadores, emprendedores y fuertes de carácter para que logren las metas que se propongan en la vida.

A mis hermanos: Martha, Rosa, José Luis, Guadalupe, Cruz, Roberto y Memo.

A mis alumnos de quienes semestre con semestre aprendo más de lo que puedo enseñarles. Me permito decirles que la distancia entre un sueño y la realidad es la disciplina, porque no existen sueños grandes sino soñadores pequeños.

AGRADECIMIENTOS

A la Universidad Nacional Autónoma de México, por permitirme pertenecer a la Máxima Casa de Estudios de nuestro País.

A la Facultad de Filosofía y Letras por darme la oportunidad de trabajar durante 25 años como profesor en sus aulas.

Al Dr. Filiberto Felipe Martínez Arellano, por haber aceptado dirigir esta investigación y demostrar en la práctica todo el conocimiento que tiene para realizar investigación en las Tecnologías de Información y Comunicación, principalmente en organización y recuperación de información, que lo han llevado a ser uno de los mejores investigadores en esta área en nuestro país.

A César, Braulio, Jaime, Florencio, Zuemi por todo el apoyo que siempre he recibido de ellos.

Al Dr. Juan José Calva González por todo su apoyo como Coordinador del Posgrado en Bibliotecología y Estudios de la Información.

A mis sinodales que después de la revisión de esta investigación permitieron que se enriqueciera con sus valiosos y acertados comentarios.

Dra. Georgina Araceli Torres Vargas

Dr. Juan Voutssás Márquez

Dra. Catalina Naumis Peña

Dra. Brenda Cabral Vargas

MUCHAS GRACIAS

INTRODUCCIÓN

El avance de la tecnología se ha manifestado en muchos ámbitos de nuestra vida desde la casa hasta las formas de enseñanza, y la información no ha sido ajena a este avance tecnológico, ni puede quedarse al margen de ella, esto lo vemos reflejado en la explosión de información en la red conocida como Internet.

En la actualidad estamos viviendo en la llamada era de la información, en donde ésta es un recurso estratégico que ha cobrado en la actualidad tanta importancia como la que tuvieron en otra época otros bienes.

Las Tecnologías de Información y Comunicación (TICs), han producido cambios insospechados respecto a cómo se produce y se utiliza la información en nuestros días. La información, un elemento que forma parte de nuestra labor profesional, ha sufrido cambios en la forma en que se genera, se distribuye y se obtiene.

El avance tecnológico que se generó a finales del siglo XX y se sigue incrementando en el presente, permite que se desarrollen nuevas formas de crear y tener acceso a la información, principalmente apoyadas en medios electrónicos, haciendo posible que casi cualquier persona pueda crear y publicar recursos de información en Internet.

Esto ha ocasionado que se incremente de manera incontrolable la información que se encuentra en la red, trayendo como consecuencia que los usuarios se enfrenten a una red con muchos problemas para poder recuperar la información que sea relevante para sus necesidades informativas.

Al buscar información en la Red, los usuarios se encuentran con dos problemas muy graves, el primero es que al realizar una búsqueda en Internet encuentran

mucha información y en ocasiones es muy difícil separar las páginas que cuentan con un valor académico.

El segundo problema es que en ocasiones no encuentran los documentos que buscan, esto ha ocasionado que mucha información en la red se encuentra perdida, debido a que los recursos de información son almacenados en la Web por sus propios creadores sin ningún tipo de organización para su posterior recuperación, y esto posiblemente sea uno de los puntos débiles de la red, porque millones de documentos no pueden ser recuperados por no contar con una adecuada organización que le permita a los buscadores su localización y recuperación.

Entre las causas más comunes a las que se enfrenta el usuario, al tratar de recuperar documentos electrónicos en la red, están las siguientes:

- La falta de datos suficientes para su indización.
- Inadecuado uso de estructuras de metadatos.
- El uso de protocolos inadecuados.

Es así como muchos autores han hecho diferentes propuestas para mejorar tanto la organización de la información, para poder recuperar eficientemente la información que necesitan los usuarios de la red.

En este sentido, los metadatos principalmente los de contenido temático, han sido propuestos para organizar la forma en que un sitio es indexado por los motores de búsqueda. Es importante mencionar que en la actualidad la mayoría de los principales buscadores utilizan metadatos para construir los índices de sus bases de datos.

Esto mejora la búsqueda de información, debido a que eliminan la basura y el material irrelevante recuperado, es decir que al buscar información

directamente en los campos que contienen las bases de datos del buscador se recupera información más precisa y relevante.

Existen algunas soluciones para mejorar la recuperación de información, entre ellas se encuentran:

- El uso de metadatos para describir la información.
- El uso de estructuras de metadatos como el Dublin Core.
- El uso del lenguaje de marcado XML con la sintaxis y la semántica para describir los metadatos.
- El uso de etiquetas META como la HTTP-EQUIV y la NAME.
- El uso de lenguajes controlados.

Por lo anteriormente expuesto, se plantea la siguiente problemática.

El incremento de información electrónica que se encuentra en Internet ha ocasionado un crecimiento incontrolable de la Web, además, la falta de organización de la misma ha traído como consecuencia que los usuarios de Internet se encuentren con una red carente de organización y la búsqueda de información académica en ella sea cada vez más complicada.

Por esta razón, el uso de estructuras de los llamados metadatos en particular los de contenido en la organización de la información puede ser una solución para la recuperación eficiente de información en formatos electrónicos.

Es importante mencionar que existen diferentes estructuras de metadatos; algunos de ellos organizan la información de manera que pueda ser recuperada fácilmente, pero otros no la organizan de la misma forma y es más complicado recuperarla. Por otro lado, al no existir estudios que analicen a fondo el valor de los metadatos en la organización de la información y su posterior recuperación en los buscadores que existen en la Web, se tomó la decisión de realizar un estudio en diferentes herramientas de recuperación de información, como son motores de búsqueda, metabuscadores y buscadores

semánticos, para detectar cuáles son los más eficientes para la organización y recuperación de información académica.

Para llevar a cabo la presente investigación, se contó con el siguiente objetivo:

Evaluar el funcionamiento de diferentes herramientas de búsqueda y recuperación de la información que se encuentran en la Web, para identificar si existe relación entre la recuperación de la información y el uso de metadatos de contenido.

Asimismo, en esta investigación se parte de la hipótesis de que:

Si se utiliza una estructura de metadatos adecuada en la organización de la información, se obtiene un mayor índice de recuperación de la información.

Sin embargo muchas de las herramientas de búsqueda de información de la Web no cuentan con estructuras de metadatos adecuadas, lo que dificulta la recuperación de la información, en particular la recuperación temática.

La metodología que se utilizó en la presente investigación para conocer si las herramientas previamente desarrolladas, recuperan adecuadamente la información, consistió en un estudio, que permitiese conocer el grado de efectividad con que las herramientas de búsqueda recuperan la información. Para tales efectos, se realizó una selección de los buscadores que se encuentran operando en la red. Lo cual resultó difícil, debido a que existe una gran cantidad de diferentes tipos y tamaños. Por lo cual, se utilizaron solo los motores de búsqueda, de mayor frecuencia de uso en la Web, previa identificación del ranqueo correspondiente, de los cuales se mencionan los tres sitios principales: eBizMBA(2016), ListofSearchEngines (2016) y Reliablesoft (2016). Con respecto a los metabuscadores, se tomó como referencia para su selección a los que aparecen en la página de ListofSearchEngines, en el rubro de metabuscadores y por ser los más utilizados. Para la selección de los buscadores semánticos, se utilizaron dos sitios Web: Search engine journal (2016) y Makeuseof(2016). Es importante aclarar que solo se utilizaron cuatro

buscadores semánticos por ser los únicos disponibles. Como resultado de este proceso de selección de herramientas de búsqueda de información, se consideraron cinco motores de búsqueda, cinco metabuscadores y cuatro buscadores semánticos.

Finalmente fue necesario estructurar una estrategia de búsqueda, que nos permitiera conocer si las herramientas de búsqueda recuperaban la información eficientemente. Como otro aspecto importante, en la investigación fue la selección del tema para efectuar las búsquedas en las diferentes herramientas, el tema seleccionado, fue:

En español: Las relaciones diplomáticas entre Estados Unidos y Cuba

En inglés: Diplomatic relations between the United States and Cuba

La decisión de incluir este tema obedeció a que es una temática muy relevante en estos momentos, tanto para los dos países involucrados, Estados Unidos y Cuba, así como para el ámbito internacional. Por otro lado, se decidió que la búsqueda debería realizarse en idioma Inglés y español, debido a que son los idiomas oficiales de los países protagonistas. Por lo que, se procedió a realizar tres tipos de búsqueda:

Búsqueda libre.

Búsqueda utilizando operadores lógicos boléanos.

Búsqueda utilizando operadores lógicos boléanos en el campo de título.

Los resultados de la presente investigación, conllevan a establecer las bases para realizar otros estudios que complementen el supuesto de la existencia de una relación entre el uso de estructuras de metadatos y la recuperación de información eficiente.

Para cumplir con el objetivo antes señalado y comprobar la hipótesis, esta investigación se dividió en tres capítulos.

En el primer capítulo, se aborda todo lo referente a la organización de la información, sus objetivos, así como las ventajas y desventajas de su uso en la organización de la información. Asimismo, en este capítulo se trata el tema de los metadatos y su importancia en la organización de la información para su posterior recuperación.

En el segundo capítulo se explica brevemente, el concepto de recuperación de la información, así como el tipo de herramientas que existen para recuperar información, asimismo se presenta una tipología de los buscadores, entre ellos los motores de búsqueda, metabuscadores y buscadores semánticos, además se explica el porque del incremento de la información electrónica en la red de redes, partiendo del uso de las tecnologías de información en la transformación de los materiales impresos a los materiales electrónicos, que día a día se incrementan en la Web y que en ocasiones no cuenta con un soporte académico, así como la que si cuenta con ese soporte que le brinda el respaldo de instituciones académicas, entre las cuales vemos a las revistas de libre acceso, las de acceso limitado, libros electrónicos y las herramientas para recuperarlas.

En el tercer capítulo se presentan los resultados de la evaluación en la recuperación de la información en las herramientas de búsqueda que se encuentran en la Web. Desde la selección de las herramientas de búsqueda, la elaboración de la estrategia de recuperación sobre la temática utilizada tanto en inglés como en español, asimismo se presentan los resultados de la búsqueda acompañada de pantallas de las diferentes herramientas utilizadas, así como el análisis de los resultados de la búsqueda de información de las herramientas seleccionadas que trabajan en la Web.

Finalmente, se presentan las conclusiones de la evaluación, asimismo se presentan las referencias utilizadas.

CAPITULO 1

LA ORGANIZACIÓN DE LA INFORMACIÓN

1.1 Necesidad de organizar la información

El avance de la tecnología ha permitido que se desarrolle un fenómeno en la Web, denominado como la explosión de información. Esto ha ocasionado que se incremente de manera incontrolable la información que se encuentra en Internet.

Adicionalmente a ello los usuarios se enfrentan a una red con muchos problemas y dificultades para poder recuperar la información que sea relevante para ellos, debido a que la información contenida en la Web no está organizada adecuadamente, esto ocasiona que las herramientas de búsqueda que se encuentran en la red no puedan recuperar la información eficientemente. Para evitar este problema de los usuarios de la Web es necesario que la información sea organizada para que dichas herramientas la recuperen adecuadamente.

En este sentido, Arlene Taylor (1999:2) en su libro *The organization of information*, hace mención sobre la necesidad del ser humano de organizar para poder recuperar lo que necesita, menciona también que los seres humanos aprenden analizando y organizando datos, información y conocimiento.

Por esta razón surge la necesidad de organizar la información. Si se toma en cuenta que para que los usuarios puedan obtener información y generar conocimiento, es necesario que la información este bien organizada, así será más fácil recuperarla.

En nuestra época es importante organizar la información, principalmente por el incremento de información que se ha generado gracias a las tecnologías de la información debido a los avances tecnológicos que se han desarrollado en las últimas tres décadas, como consecuencia de esto, el nacimiento y desarrollo de la red conocida como Internet y más específico uno de sus servicios como lo es el WWW. Como se sabe la Web tiene una gran cantidad de información la cual se ha incrementado por el desarrollo de la ciencia y la tecnología.

Estos dos aspectos, se pueden considerar como razones suficientes de la necesidad de organizar la información, si se toma como ejemplo el de las diferentes disciplinas del conocimiento, que, para que se incremente su desarrollo tiene que mostrar sus avances a sus propias comunidades y a la sociedad en general, esto lo llevan a cabo presentando los resultados de sus investigaciones, mediante la publicación de la información y para ello se vale del uso de la tecnología, que como ya se mencionó una de los principales herramientas es la Web que permite casi instantáneamente subir la información en los diferentes sistemas de información.

Como se puede observar, estos dos aspectos, el desarrollo de las diferentes disciplinas del conocimiento, apoyado por el uso de tecnología hacen cada vez más amplio el volumen de información en la Red, esta premura por presentar sus investigaciones, conlleva a otro problema grave para la Web, esta información u objetos de información en ocasiones no cuentan con elementos suficientes para poder describirlos adecuadamente y esto ocasiona un caos a los usuarios cuando tratan de recuperar los diferentes objetos de información.

Ahora bien, si es necesario organizar la información, ¿Cómo se puede lograr? Taylor (1999:2) aborda la organización de información, pero en lugar de utilizar términos tales como libro o artículo para referirse a la unidad organizable de información, utiliza términos tales como entidad portadora de información y paquete de información, asimismo toma las funciones de Hagler, aunque menciona que estas funciones tienen más énfasis en el trabajo del bibliotecario,

que fue el objetivo del autor, pero ella modifica la redacción y entonces se reflejan las principales actividades involucradas en toda la organización de la información.

Las funciones que identifico Hagler, son las siguientes:(citado por Taylor, 1999:3-4)

1. Identificar la existencia de todo tipo de entidades portadoras de información, y ponerlo a disposición.
2. La identificación de las obras contenidas dentro de dichas entidades portadoras de información o como parte de ellas.
3. Agrupar estas entidades portadoras de información sistemáticamente, en colecciones de bibliotecas, archivos, museos, archivos de comunicación de Internet y otros repositorios.
4. La producción de listas de estas entidades portadoras de información preparadas de acuerdo con las normas estándares de citas.
5. Asignar nombre del autor, título, materia, y otro acceso útil a estas entidades portadoras de información.
6. Proporcionar los medios para la localización de cada una de las entidades portadoras de información o una copia de la misma.

En el caso de identificar la existencia de todo tipo de entidades, se menciona que un libro puede ser publicado o un sitio Web puede ser establecido, pero si nadie sabe de su existencia, excepto la persona involucrada en su creación, de ninguna manera será de utilidad informativa para nadie. La existencia y la identidad pueden hacerse saber de muchas maneras; anuncios editoriales, anuncios por correo electrónico, comentarios, listas de asignatura relacionados, por nombrar algunos.

La mayoría de los editores crean catálogos anunciando sus productos, junto con los resúmenes de ellos. El producto de esta actividad son herramientas de referencia, como Books in Print.

Con respecto a la identificación de las obras contenidas, se menciona que un grupo de obras como cuentos o una agrupación de obras artísticas puede ser considerada como una entidad portadora de información en su conjunto, como alternativa, cada historia individual o artística pueden ser considerados como una entidad portadora de información también. Un sitio Web que trata de un personaje famoso puede tener obras digitalizadas individuales de la persona, material biográfico, las obras que tratan sobre el personaje escritos por autores contemporáneos. Lo escrito sobre la persona y los eventos pueden ser obras importantes por derecho propio y necesitan ser identificados por separado

Agrupar en colecciones ha sido tradicionalmente una actividad de las instituciones como bibliotecas, archivos y museos. Pero existen otros casos en donde también se crean colecciones, como el caso de documentos personales realizadas a causa de un intenso interés en un determinado tipo de información, documentos con información interna y la información necesaria para llevar a cabo el trabajo en una oficina, colecciones departamentales de materiales para la enseñanza de una disciplina universitaria en particular, etc., para posteriormente darlas a conocer públicamente.

Algunas colecciones en ocasiones incluyen recursos que no son de la institución. Por lo tanto muchas instituciones compran los derechos de algunas colecciones electrónicas para que sus usuarios puedan utilizarlos. Parte del proceso de organización es la determinación de si los recursos electrónicos necesitan ser incorporados a la colección de manera permanente.

Los productos creados por esta actividad incluyen bibliografías, índices, catálogos de bibliotecas, archivos, herramientas de búsqueda y registros de museos. Estas herramientas son importantes para la recuperación de paquetes de información individuales, porque si uno busca un artículo conocido, es necesario conocer su ubicación física, y las herramientas antes mencionadas nos pueden ayudar, asimismo estas pueden estar en forma impresa o en formato electrónico.

Finalmente en el punto seis se menciona que proporcionar la ubicación y el acceso a los documentos que se encuentran en una colección, ha sido un valor agregado en este siglo para las instituciones. Los catálogos u otras herramientas creadas en estas instituciones dan información sobre la ubicación física del documento, si no se ha prestado a domicilio a un usuario o no está siendo utilizado por alguien en las instalaciones.

Entonces ¿cómo organizar la información que contiene conocimientos y estos a su vez tiene flujos considerables de información?, como respuesta a esta pregunta se han propuesto por algunos autores los siguientes puntos para organizar la información.

- Identificando la existencia de información y cómo puede hacerse disponible
- Identificando el trabajo contenido dentro de esa información
- Sistematizar esa información en un repositorio
- Produciendo listas o tablas con de esta información organizada
- Proveer accesos a la información (como lo son, convencionalmente, el nombre, título, tema y otros tópicos útiles para recuperar información)
- Proveer un significado de localización de esa información.

1. 2 Objetivos de la organización de la información

Sin duda alguna, organizar la información persigue un fin. Al organizarla habría un beneficio importante para los usuarios, y los sistemas de información también se beneficiarían, principalmente en relación a las herramientas de búsqueda, debido a que el usuario tendría más opciones para buscar y recuperar la información eficientemente, esto se ha visto desde la antigüedad hasta nuestros días.

Los principales objetivos que persigue la organización de la información son, ayudar al usuario para:

- Localizar recursos de información
- Identificar el objeto de información
- Valorar y seleccionar los recursos de información útiles
- Obtener recursos de información

1.2.1 Localizar recursos de información

Este es quizá el principal objetivo para la organización de la información. Cuando un usuario busca información en cualquier sistema de recuperación, lo que persigue es encontrar recursos de información que le sean de utilidad para sus actividades escolares o de investigación. Esto se logra a través de los puntos de acceso, los que tradicionalmente han sido:

- Autor
- Título
- Tema

Cuando un usuario necesita información y conoce el nombre de algún autor o creador, lo primero que hace es tratar de encontrar documentos de dicho autor o creador. Como se sabe, el autor es la persona u organismo responsable del contenido intelectual de una obra; puede ser el escritor, compositor o artista. También se consideran como autores a los organismos que generan diversas obras bajo su propia responsabilidad, tales como:

- Gobiernos
- Universidades
- Asociaciones
- Instituciones

El título es otro punto de acceso que el usuario utiliza para recuperar información, el título es la palabra o frase con que el autor da a conocer su obra.

Finalmente, el punto de acceso mayormente utilizado desde hace muchos años por los usuarios que requieren información y la buscan en algún sistema de información (catálogo, buscador, metabuscador, etc.) es el tema como punto de acceso, para recuperar información. Esto es lógico, si tomamos en cuenta que un usuario trata de encontrar la mayor cantidad de recursos de información sobre una temática y al buscar por este punto de acceso le permite tener un panorama más amplio de la información que existe sobre ese tema en el sistema de información que esté utilizando. El tema es considerado como el contenido o asunto del que trata una obra o un recurso de información.

De esta manera, los puntos de acceso básicos e indispensables para organizar la información incluyen:

- Al creador o primer nombre de creador cuando se cita más de uno.
- Título propio o título del objeto de información.
- Materia, términos de materia.

Los puntos de acceso deben estar normalizados, por lo tanto es indispensable crear catálogos de autoridad, principalmente en el ámbito de los temas de que tratan los recursos de información.

Como se puede apreciar en los párrafos anteriores, existen elementos que si se utilizarán en la elaboración de las herramientas de búsqueda, como el caso de los temas como punto de acceso, tal vez la búsqueda de información en la Web, por parte de los usuarios, sería más sencilla y podrían recuperar solamente lo que necesitan, es decir que solo recuperarían los mejores documentos para cubrir sus necesidades de información.

Aunado a esto, otro factor que dificulta la búsqueda de información, es la aparición impresionante de nuevos objetos de información que han engrosado de manera importante la red de redes.

Para representar el contenido de una obra, se han utilizado los lenguajes controlados y los lenguajes libres, para representar correctamente los temas de qué trata un recurso de información.

Lenguajes controlados:

Los lenguajes controlados, son lenguajes documentales, entre los que encontramos a: tesauros, listas de encabezamientos de materia y clasificaciones. Esto debido a que utilizan un vocabulario previamente seleccionado, y admiten un limitado número de modificaciones en el momento de su utilización.

Es decir que se trata de vocabularios que controlan las ambigüedades propias del lenguaje, tanto para la representación de un tema como para la búsqueda y recuperación de información.

Las ambigüedades más características son:

- Sinonimia o “silencios” en la recuperación, porque un mismo tema puede figurar de distintas formas.
- Polisemia o “ruido”, dado que un mismo término o frase puede tener distintos significados.
- Homonimia, un caso similar al anterior, presente en términos o frases que se escriben igual pero pueden adquirir distintos significados según su contexto.

Taylor identifica tres tipos de herramientas que se utilizan para el manejo del lenguaje controlado (2004), estos son:

- Listas de encabezamientos de materia
- Esquemas de clasificación bibliográfica
- Tesauros

Para efectos de esta investigación, solo se mencionarán, dos de ellas, las listas de encabezamientos de materia y los tesauros.

Listas de encabezamientos de materia

En el contexto de los lenguajes documentales, las listas de encabezamientos de materia (LEM) son lenguajes controlados, formados por un conjunto de términos extraídos del lenguaje natural.

Los encabezamientos de materia son listas estandarizadas, representan como su nombre lo indica, la materia o asunto del que trata un recurso de información de forma jerárquica, es decir que si para describir el asunto de un documento, es necesario asignar varios encabezamientos de materia, primero se asigna el que cubra de forma general el contenido y después se agregan los encabezamientos secundarios, para especificar su contenido.

Se puede decir entonces, que los encabezamientos de materia son los signos que representan la materia o asunto del que trata el libro. Los signos son palabras o conjuntos de estas que representan conceptos; la materia o asunto es el objeto del libro, de lo que trata y no el discurso de este documento. Lo que equivale a decir que los encabezamientos de materia representan conceptos y no juicios, objetos de discurso y no discursos, esto se da precisamente por la forma en que los términos se relacionan. En este sentido podemos mencionar dos enfoques, estos son:

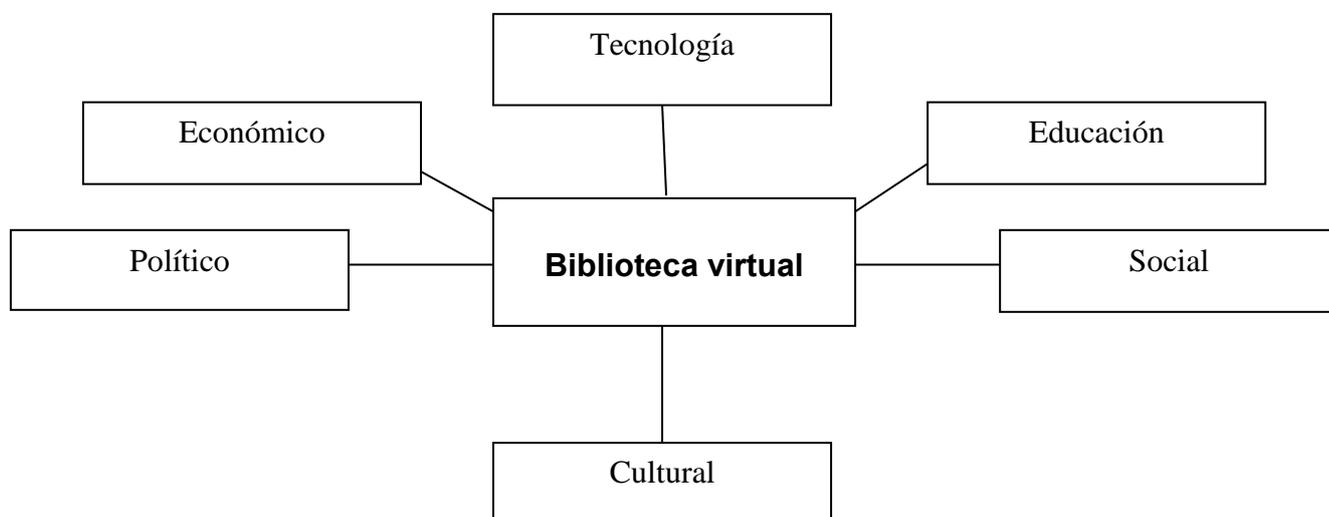
- Establecer una red de relaciones
- Construir una visión jerárquica

Cuando tratamos de establecer una red de relaciones, “El campo de relaciones del tema puede representarse gráficamente como una red de conceptos relacionados entre sí en distintas maneras y grados” (Jirón 2002)

Un tema nunca aparece en forma aislada, esto es que en ocasiones no se encuentran otros conceptos relacionados con el tema, probablemente porque desconocemos las conexiones reales del mismo, esto es muy común cuando realizamos una búsqueda sobre un tema nuevo para nosotros.

Esto es relativamente fácil de solventar, por medio del uso de sistemas postcoordinados de recuperación.

Ejemplo gráfico de un campo de relaciones del tema: "**Biblioteca Virtual**"

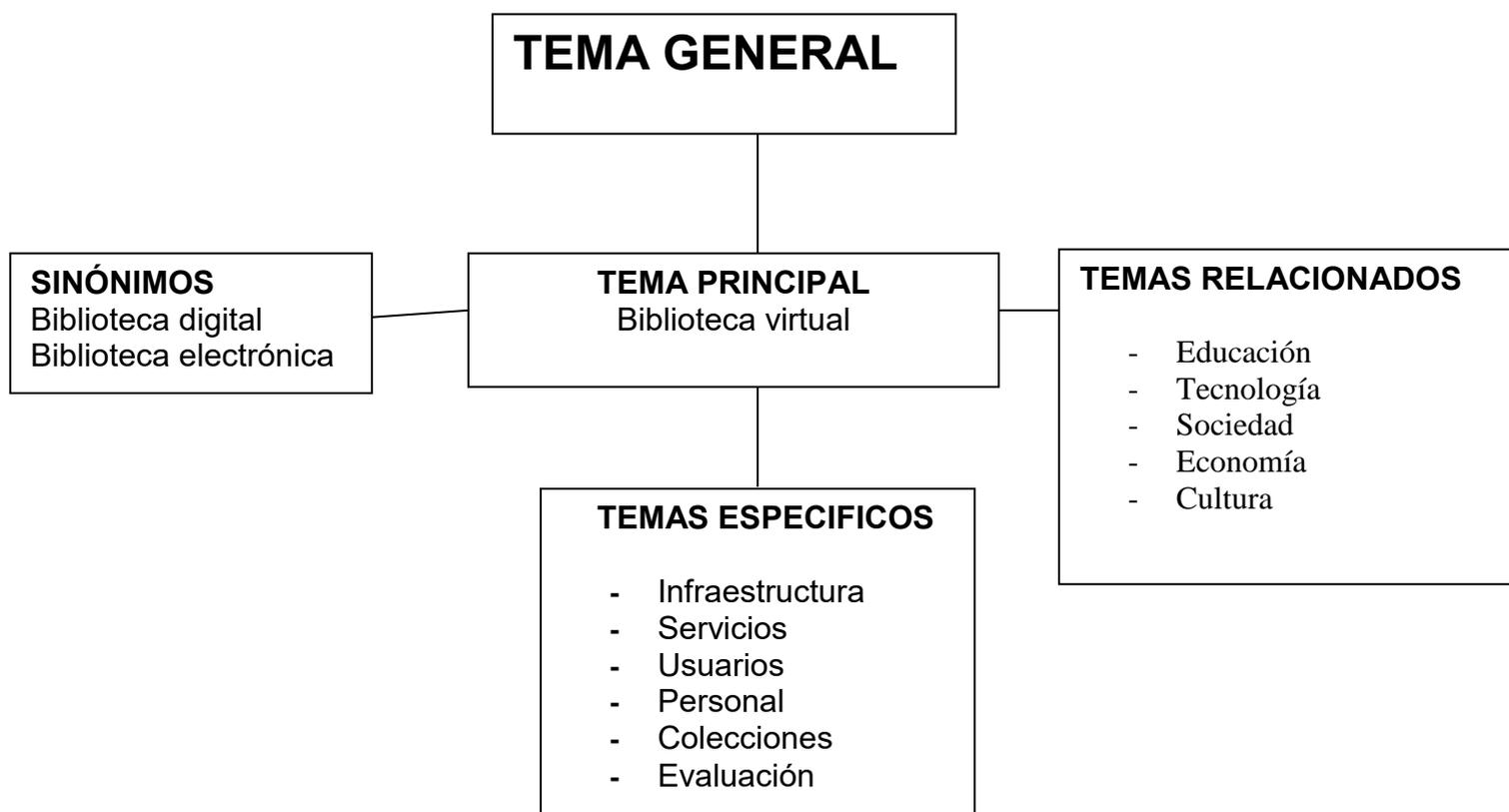


En esta red de relaciones se puede observar desde que punto de vista un usuario puede necesitar su información. Por lo tanto es importante conocer desde que punto de vista se tratar el tema, puede ser cultural, social, político, económico etc. Otro punto importante es que esta red de relaciones es parte fundamental para crear una visión jerárquica de un tema.

Al construir una visión jerárquica, se jerarquizan los temas, esto nos puede auxiliar para conocer las relaciones inherentes al tema básico y también nos facilita el uso del sistema de recuperación. (Jirón 2002) Esta visión nos ayuda a conocer las relaciones de un tema, si existen sinónimos y la profundidad que se

requiere sobre el mismo, para esto los temas específicos son una excelente opción.

Ejemplo de una visión jerárquica con el tema: “ **Biblioteca Virtual**”



Como se puede observar en la gráfica; la visión jerárquica nos ayuda a tener una idea clara sobre el tema del cual queremos encontrar y recuperar información.

Tesauros

Un tesauro es una herramienta para el control del vocabulario. Orienta a los indizadores y a los usuarios sobre los términos que pueden utilizar y, así ayuda a mejorar la calidad de la recuperación. El tesauro es una mezcla de listas de encabezamientos de materia y los esquemas de clasificación.

A partir de la norma UNE 50106:1990, Directrices para la creación y desarrollo de tesauros (equivalente de la ISO 2788:1986) los tesauros se pueden definir según su función y según su estructura:

- Por su función,
- Por su estructura

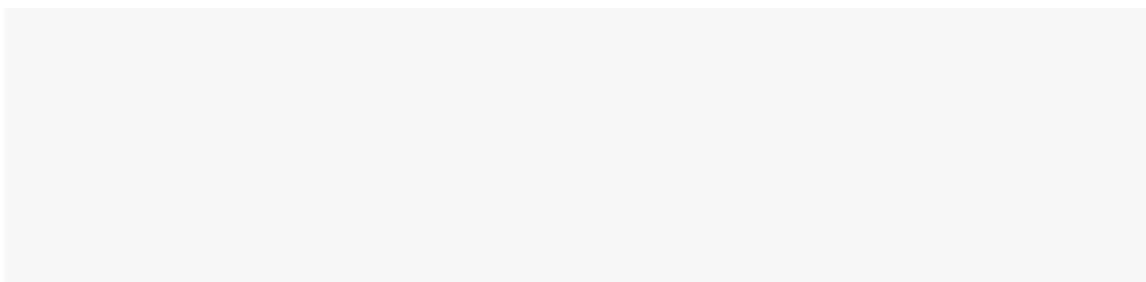
Por su función

Se pueden definir como instrumentos de control terminológico. Controlar el vocabulario significa identificar dentro de un campo semántico todos los conceptos que son representados por más de un término. La identificación de términos equivalentes hace posible minimizar la pérdida de información en las búsquedas realizadas en un sistema documental automatizado.

Por su estructura

Los tesauros permiten conocer todos los términos relacionados con un concepto determinado, lo que ayuda a añadir más términos adecuados para enriquecer tanto los análisis de contenido de los documentos como las estrategias de búsqueda para recuperar información.

Los tesauros son un prototipo de lenguaje de indización y recuperación controlado, que se basa en la postcoordinación de sus descriptores. Además representa de manera unívoca los conceptos de los documentos evitando así los problemas de homonimia, sinonimia, polisemia (relaciones de significado), así como la descripción de las relaciones entre los descriptores que la componen.



Ventajas y desventajas de los lenguajes controlados

Al utilizar los lenguajes controlados para la organización de la información, surgen una serie de ventajas y desventajas, entre las que podemos mencionar las siguientes:

Ventajas:

- Permiten representar un tema con el mejor término posible
- Reducen la diversidad de términos
- Ayudan a los usuarios a identificar los términos para realizar búsquedas
- Proporciona una mayor precisión y ahorro de tiempo a los usuarios

Desventajas:

- Los costos de las herramientas que utilizan estos lenguajes, debido a que son elaborados por profesionistas, por lo tanto los costos son altos
- La continua aparición de nuevos términos en las diferentes áreas del conocimiento
- El desconocimiento, por parte de los usuarios, de la existencia de los diferentes tipos de lenguajes controlados y el desconocimiento en su uso
- El vocabulario controlado proporciona al usuario un punto de búsqueda, en vez de dos o más, y reduce la posibilidad de que la búsqueda sea completa. Además puede perderse alguna información debido a la falta de clasificación y errores en el análisis.

Lenguajes libres:

Los lenguajes libres son vocabularios cuyas entradas están tomadas directamente del lenguaje natural, es decir, los utilizados por el autor o autores del documento. En ellos no existe ningún tipo de control y están constituidos por cada término o frase que se extrae de las obras analizadas. Las listas de palabras clave constituyen un ejemplo de lenguaje libre.

Lenguajes libres, se componen de términos no predefinidos que se van generando a partir de la realización de procesos de indización. Lenguajes altamente pertinentes, coherentes, de bajo coste y que proporcionan una enorme cantidad de resultados cuando se realizan búsquedas. De este tipo son las listas de descriptores libres y las listas de palabras clave. Los lenguajes libres no son propiamente lenguajes documentales puesto que para que reciban este nombre el vocabulario debe estar controlado.

Como sabemos el lenguaje libre avanza, en parte, por la rapidez que aportan las computadoras para localizar los términos deseados, pero aún no han superado a los sistemas documentales de terminología controlada, porque suele ser útil sólo cuando la información buscada es de tipo general, el principal ejemplo son las:

- Palabras clave

Las palabras clave suponen una herramienta imprescindible a la hora de realizar una búsqueda de información, porque permiten el acceso a los recursos de información que se encuentran en los sistemas de información. Las palabras clave pertenecen o son considerados lenguajes libres, esto debido a que son tomadas del título de los documentos, de los resúmenes o en ocasiones de la indización del documento completo.

En estos sistemas las nociones o conceptos que se extraen en la indización para expresar el tema o los temas del documento, tienen todas la misma categoría y no se expresarán en ningún orden determinado.

Serán únicamente las necesidades de información de los usuarios en la fase de recuperación los que obliguen a la combinación de conceptos que se realizará utilizando los operadores lógicos booleanos.

Al utilizar los lenguajes libres, para la organización de la información, también surgen una serie de ventajas y desventajas, entre las que podemos mencionar las siguientes:

Ventajas:

- Se indexa la mayor parte del documento, por esta razón se tiene una gran cantidad de palabras clave
- Se usan los resúmenes elaborados por los mismos autores, esto proporciona otros puntos de acceso, que permiten hacer búsquedas más extensas
- Los nuevos términos de un área aparecen antes en los lenguajes libres que en los vocabularios controlados, esto le da mayor actualidad a las búsquedas.
- El usuario puede utilizar términos naturales para realizar su búsqueda, puesto que estos salieron del mismo texto.
- Es más barato que los lenguajes controlados, esto debido a que en ocasiones la indización es automática.

- Los buscadores indizan las palabras que se encuentran, principalmente en los encabezados y las ponen en sus bases de datos en el campo de título.

Desventajas

- Al contar con muchos términos, no toda la información que se recupera es de relevancia
- Al indizar todo el texto, existen muchas palabras que son sinónimos o palabras vacías
- La existencia de muchos términos condiciona la precisión en la recuperación de información
- El usuario utiliza mucho tiempo para revisar y seleccionar los documentos pertinentes para su investigación

Lenguajes de información precoordinados y poscoordinados

En los sistemas de recuperación podemos encontrar, dos sistemas para la recuperación de información, estos son:

- Los precoordinados.
- Los postcoordinados

Autores como Martínez Arellano han denominado a estos lenguajes como sistemas de recuperación precoordinados y sistemas de recuperación postcoordinados.

Los sistemas precoordinados combinan los términos en el momento de la asignación. Las listas de encabezamientos de materia son considerados sistemas de recuperación precoordinados. En estos sistemas las distintas nociones o conceptos que se unen para expresar una materia o un tema ocupan un lugar determinado, es decir se introducen en el momento de la indización en un orden previamente establecido y la recuperación habrá de hacerse secuencialmente, siguiendo ese orden. Suelen ser muy precisos, pero también muy rígidos.

Por otro lado los sistemas de recuperación postcoordinados. Combinan los términos en el momento de la recuperación. Los tesauros utilizan sistemas de recuperación postcoordinados.

Martínez, agrega que los sistemas de recuperación precoordinados y los postcoordinados se pueden presentar, tanto en los lenguajes controlados como en los libres. Por lo tanto, cuando se usan estos tipos de sistemas de recuperación, encontramos que se pueden utilizar algunos enfoques que nos ayudan a clarificar el campo de relaciones que nos interesan estudiar, al puntualizar los términos de búsquedas pertinentes (palabras que serán utilizadas en el sistema de búsqueda), razón por la cual, la mayoría de los buscadores permiten utilizar operadores lógicos boléanos, para realizar esta combinación de términos al recuperar la información, desafortunadamente, para realizar búsquedas utilizando estos operadores, es necesario que se conozca adecuadamente su funcionamiento.

1.2.2 Identificar el objeto de información

Este tiene que ver principalmente con la cuestión de confirmar o estar seguro de que un recurso de información corresponda al documento que se busca y se necesita.

Ahora bien, ¿cómo puede un usuario saber, si el documento u objeto de información es el que verdaderamente necesita?, para esto es necesario que se auxilie de otros elementos que se pueden describir de un recurso de información. Estos son los aspectos que conocen de los catálogos de las bibliotecas, parte de lo que nosotros conocemos como las siete áreas de la catalogación y que sirven para describir los atributos o características del objeto de información:

- Título y de la mención de responsabilidad
- Mención de edición y responsabilidad relacionada con la edición
- Publicación, distribución, etc.
- Descripción física
- Mención de serie
- Notas
- ISBN

Si bien estas siete áreas se utilizan en la descripción de los atributos o características del objeto de información, también se tienen que tener normas o estándares para saber cómo registrar correctamente cada una de ellas.

1.2.3 Valorar y seleccionar el objeto de información

Corresponde a saber seleccionar un recurso de información que pueda solucionar las necesidades de información de los usuarios respecto al medio,

contenido, soporte, etc., o también descartar un recurso por ser inapropiado a las necesidades de los usuarios. En este sentido cuando un usuario realiza una búsqueda de un objeto de información en algún sistema de información y al utilizar los elementos como el tema, idioma, autor, edición, le es más fácil seleccionar el más relevante que se encuentra en el sistema de información y a su vez con los mismos elementos saber si la información que recuperó le es pertinente para sus quehaceres académicos.

1.2.4 Obtener la información

Accesar al ejemplar descrito (es decir, suministrar la información que permitirá al usuario adquirir un ejemplar por medio de la compra, el préstamo, etc., o acceder al ejemplar electrónicamente por medio de una conexión en línea a una fuente remota); o acceder, adquirir u obtener datos de autoridad o datos bibliográficos.

1.3 Los metadatos

En el ámbito bibliotecario se ha usado tradicionalmente la catalogación para organizar los documentos que llegan a las unidades de información, describiendo los recursos bibliográficos, así como identificando las obras que contiene y determinando los puntos de acceso por medio de los cuales la obra podrá ser recuperada por los usuarios.

Hasta aquí no existe ningún problema para los bibliotecarios, pero a partir de la aparición de las Tecnologías de la Información y Comunicación (TICs) y como se mencionó en el apartado anterior, han surgido nuevos formatos y recursos de información, que hacen muy complicada su representación para organizarlos, esto debido a que estos recursos de información cuentan con otros atributos que no contaban las obras impresas, entre ellos podemos mencionar:

- **Intangibilidad o inmaterialidad**, es decir sin soporte físico que los contenga
- **Accesibilidad masiva**, son utilizados por muchos usuarios al mismo tiempo
- **Velocidad de duplicación**, que se pueden tener infinitas copias sin tiempos editoriales
- **Dependencia de una plataforma**, acceso exclusivamente a través de equipo de cómputo y redes de comunicación
- **Inestabilidad**, escasa duración
- **No linealidad**, fragmentos que sumados constituyen el documento completo
- **Multimedialidad**, incluyen sonido y movimiento
- **Interactividad**, posibilidades de interacción entre el usuario y el recurso
- **Hipertextualidad**, Textos completos relacionados
- **Mutabilidad**, posibilidades de cambios permanentes en el recurso
- **Volatilidad**, desaparición del recurso

Como se puede observar, existe una gran cantidad de atributos en estos recursos de información electrónicos, por otro lado, estos objetos de información han aumentado considerablemente en la red, entre ellos podemos mencionar a:

- Periódicos
- Revistas
- Mapas
- Blogs
- Fotografías
- Videos
- Imágenes

- Bases de datos
- Catálogos
- Páginas Web
- Intranets
- Repositorios
- Buscadores
- Metabuscaradores
- Multibuscaradores, etc

Asimismo, los objetos de información que se encuentran en la red, están contenidos en diferentes formatos, entre los principales, podemos mencionar:

- HTML
- PDF
- Imagen
- Multimedia

Cada vez que aparecen en la red, nuevos objetos de información, el usuario tiene muchas dificultades para encontrarlos, debido a que no están organizados adecuadamente, principalmente porque cuentan con otros atributos diferentes a los que contaban los impresos, entre ellos se puede destacar:

- Formato en el que se encuentra el objeto de información original
- Características espaciales del objeto de información
- Características temporales del objeto de información
- Relación con otros objetos de información de la misma materia o área
- Hipertextualidad

De esta forma, se entiende la necesidad de organizar la información que se encuentra en la red con algún esquema de organización que tome en cuenta los nuevos atributos de los objetos de información actuales, en beneficio de los usuarios de la Web.

Tomando como base lo anterior, se puede decir que lo único que permitiría solucionar en alguna medida, que los usuarios puedan obtener la mejor información cuando la buscan, es que esta información contenga algunos elementos que permitan una adecuada descripción, para su mejor recuperación.

Por lo tanto el uso de los llamados metadatos se presenta como una solución para ayudar a los creadores de recursos de información, para asociar a ellos descripciones susceptibles de procesarse por las herramientas de búsqueda de información, permitiendo que los usuarios puedan recuperar información de calidad y relevante para sus investigaciones.

1.3.1 Definición de metadatos

Para mejorar el funcionamiento de las herramientas de recuperación de información es muy importante la representación y descripción de los recursos de información mediante esquemas de metadatos y algún lenguaje de marcado con la sintaxis y la semántica para describir estos metadatos, los cuales pueden mejorar la recuperación de información en el WWW. (Doan y otros 1998)

En el ámbito bibliotecario, el término fue acuñado por Jack Myers en la década de los 60 para describir conjuntos de datos. La primera acepción que se le dio (y actualmente la más extendida) fue la de dato sobre el dato, ya que proporcionaban la información mínima necesaria para identificar un recurso.

En el diccionario de computación (FOLDOC, 1993), Howe afirma que los metadatos pueden incluir información descriptiva sobre el contexto, calidad y condición o características del dato.

Pero en realidad ¿qué son los Metadatos?

Actualmente, la definición más sencilla de metadatos como ya se vio en líneas anteriores es la que los refiere como “datos sobre datos” de tal forma que lo que básicamente representarán será metainformación.

En las páginas y documentos HTML se incluye más información de la que los usuarios ven en sus pantallas, estos datos adicionales siempre están relacionados con la propia página, por lo que se denominan metainformación o *metadatos*. La metainformación siempre se incluye en la sección de la cabecera, es decir, dentro de la etiqueta <head>, aunque la metainformación más conocida y utilizada es el título de la propia página, en donde se puede incluir más información útil para los navegadores y para los buscadores.

Moyano explica que a pesar de esto, tradicionalmente, el término se ha empleado con referencia a datos sobre indización y catalogación creados por profesionales en la bibliotecología; no obstante, su definición estará subordinada tanto a su labor como al campo profesional que los maneje. (Moyano Grimaldo, 2004)

Entonces la función de los metadatos, rebasa lo simplemente descriptivo, permitiendo la localización eficiente de la información, esto hace que sean usados también en diferentes profesiones y no solamente en la Bibliotecología, además se utilizan en diferentes actividades, como en la elaboración de registros de adquisición, catálogos de exhibiciones, datos sobre utilización o incluso, información para la administración de información digital.

Los metadatos se han definido de diversas maneras. (1997). Tim Berners-Lee menciona que los metadatos son: “Información inteligible para el ordenador sobre recursos Web u otras cosas”. Esta definición enfoca a los metadatos aplicada a los recursos electrónicos y se refiere a los "datos" en un ámbito más

amplio que incluye no sólo lo textual, sino también gráficos, video, programas y música que puedan aparecer en formato electrónico, etc.

Está claro que los metadatos se pueden implementar para los objetos no digitales también. Pero como se mencionó, se refiere principalmente a la información digital, especialmente en la Web.

Otra definición de metadatos es la asignada por el Proyecto DESIRE : "Los datos relacionados con los objetos que satisface a los usuarios que tienen un conocimiento previo completo de su existencia y características "(Desire project, 2000).

Los objetivos principales de los metadatos están cubiertos por esta definición, debido a que al referirse a la información de objetos que satisfacen a un usuario, de inmediato se incluyen una amplia gama de actividades después de describirlos, entre éstas se encuentran las siguientes: encontrar, describir, gestionar y conservar a largo plazo los recursos de información.

Los metadatos también facilitan y mejoran el proceso de recuperación de la información (cuando se utiliza como criterio la precisión), mediante la identificación de los principales conceptos de la fuente de información.

Gastaminza menciona que el concepto metadatos es difícil de definir con claridad. El prefijo "meta" significa "sobre", "junto a", "después", lo que sugiere que los metadatos son compañeros de viaje de los datos, sean estos un artículo científico, una fotografía o un sitio Web. Los metadatos describen un recurso de información o facilitan el acceso a un recurso de información y es inherente a los metadatos el hecho de que hay una asociación de algún tipo entre los elementos de metadatos y el recurso de información al que se refieren. (Gastaminza, Sánchez, Pérez, 2005:3)

Pinto Molina y Lancaster dicen que “En el sentido actual, los metadatos son datos sobre los datos, esto es metaelementos sobre el contenido y el contexto de los documentos digitales que permiten marcar los campos utilizados por los robots de búsqueda para clasificar y encontrar la información. Descriptor básico de recursos que facilita su identificación y recuperación. (Pinto Molina y Lancaster FW, 2001)

Es decir que los metadatos, pueden ser considerados como elementos que marcan o etiquetan los campos que contienen información sobre el contenido de los documentos electrónicos y digitales, que además sirven para localizar los recursos de información, mediante su fácil descripción y recuperación.

La norma ISO 15489-1 del 2001 define a los metadatos, en el contexto de la gestión de documentos, como: “datos que describen el contexto, contenido y estructura de los documentos, así como su gestión a lo largo del tiempo, por lo tanto los metadatos son información estructurada o semiestructurada que posibilita la creación, registro, clasificación, conservación y disposición de los documentos a lo largo del tiempo y dentro de un mismo dominio o dominios diferentes”

Burnett define a los metadatos desde dos puntos de vista, el bibliotecológico y el informático, en el punto de vista bibliotecológico menciona que los metadatos son cualquier información que registra la caracterización y relaciones de los datos fuente, o el conjunto de elementos de datos que pueden utilizarse para describir y representar objetos de información. Y desde el punto de vista de la informática menciona que los metadatos son cualquier información que soporta la efectiva utilización de datos, incluida la información que pueda facilitar su gestión, acceso y análisis.

Los autores López, García y Pernías dicen que un registro de metadatos consiste en un conjunto de atributos o elementos necesarios para describir el recurso en cuestión. Además mencionan que el concepto de metadato existe

antes de la aparición de Internet, pero en los últimos años se ha popularizado mucho dada la necesidad que hay de organizar la información en la Web y para la estandarización con vistas a la interoperabilidad de los sistemas de información.

Judith Hopkins en su documento USMARC como una estructura de metadatos cita una conferencia que dio Robin Wendler, en la que menciona que los metadatos son la información necesaria para identificar, localizar, manejar y acceder un recurso electrónico. (Martínez Arellano, 2000)

Otra definición es la que se da en el documento llamado Perfil de Metadatos IDEBA en 2011 en la cual se menciona que un metadato describe información sobre un recurso (información espacial, capa digital o set de información). En esencia consiste en “datos acerca de otros datos”. Un metadato consiste en un conjunto de atributos ó elementos necesarios para describir y documentar un recurso en particular.

Aunque no resulte una herramienta diseñada específicamente para asegurar, entre otras cuestiones, los derechos de propiedad intelectual de un recurso desarrollado, su fin principal es asegurar, para el usuario interesado en un recurso, la documentación mínima indispensable para juzgar al recurso, su confiabilidad práctica y alcance, su grado de actualización y las responsabilidades involucradas en su creación, entre innumerables atributos posibles.

Asimismo se menciona en el documento de IDEBA que un perfil de metadatos consiste en un conjunto particular de descriptores, adoptados para la documentación de información en un contexto determinado. Diversos perfiles pueden diferir tanto desde la cantidad y tipo de descriptores utilizados, como en la forma de completarse los campos. Finalmente, como se puede identificar un metadato?

Para comprender mejor que es un metadato y como se puede identificar, se presenta un ejemplo de una pantalla de captura de información en el formato MARC 21, sobre un material del autor Richard J. Brenner, que lleva por título "Make the team".

```

Cabecera          01041cam  2200265 a 4500
No. de Control    001 ###89048230
Ident. de Sist.   003 DLC
FHUT              005 19911106082810.9
Datos Fijos       008 891101s1990    maua j      001 0 eng
NCBC              010 ## $a ###89048230
ISBN              020 ## $a 0316107514 :
                  $c $12.95
ISBN              020 ## $a 0316107506 (pbk.) :
                  $c $5.95 ($6.95 Can.)
Fuente de Cat.    040 ## $a DLC
                  $c DLC
                  $d DLC
Signatura LC      050 00 $a GV943.25
                  $b .B74 1990
Clas. Dewey       082 00 $a 796.334/2
                  $2 20
Nombre Personal   100 1# $a Brenner, Richard J.,
                  $d 1941-
Título            245 10 $a Make the team.
                  $p Soccer :
                  $b a heads up guide to super soccer! /
                  $c Richard J. Brenner.
Variante de tit.  246 30 $a Heads up guide to super soccer
Edición           250 ## $a 1st ed.
Pie de imprenta  260 ## $a Boston :
                  $b Little, Brown,
                  $c c1990.
Desc. Física      300 ## $a 127 p. :
                  $b ill. ;
                  $c 19 cm.
Nota General      500 ## $a "A Sports illustrated for kids
                  book."
Nota: Sumario     520 ## $a Instructions for improving soccer
                  skills. Discusses dribbling, heading,
                  playmaking, defense, conditioning,
                  mental attitude, how to handle
                  problems with coaches, parents,
                  and other players, and the history
                  of soccer.
Materia           650 #0 $a Soccer
                  $v Juvenile literature.
Materia           650 #1 $a Soccer.

```

Ejemplo tomado de: Materiales de Referencia MARC 21.

<http://www.loc.gov/marc/umbspa/um11a12.html>

Como se puede apreciar el formato MARC 21, está constituido por campos, por lo tanto, toda la información que se encuentra en dichos campos son los metadatos, es decir los datos que aparecen en el campo de autor: **Richard J. Brenner** y los datos que aparecen en el campo de título **Make the team**, son metadatos.

Después de revisar y hacer un pequeño análisis a los comentarios y del ejemplo anterior, podemos decir que los metadatos son información sobre recursos, que permiten describir, identificar y localizar un recurso de información contenido en la Web. Los metadatos deben contener información estructurada para que las máquinas las entiendan, en este sentido los metadatos son datos altamente estructurados que describen información, contenido, la calidad, proceso de producción de datos, detalles de formatos y otras características de los datos. Es decir los metadatos son un conjunto de información de los atributos de los objetos de información que les proveen significado, contexto y organización de los mismos. Asimismo las estructuras de metadatos se encuentran conformadas por campos, lo que permite su fácil utilización en la organización y recuperación de la información.

1.3.2 Tipos de metadatos

Como hemos podido darnos cuenta, los metadatos se utilizan en diferentes áreas del conocimiento, si no es que la mayoría. En el entorno de las bibliotecas, los metadatos son de uso general para cualquier esquema de descripción de recursos, aplicado a cualquier tipo de objetos, digitales o no digitales.

Después de observar y analizar las definiciones del apartado anterior nos damos cuenta que la mayoría de los autores enfocan a los metadatos en tres grandes rubros, ya antes mencionados por Tim Berners-Lee, estos son:

- Metadatos descriptivos

- Metadatos estructurales
- Metadatos administrativos

Los metadatos descriptivos, en primera instancia y como su nombre lo indica, sirven para describir e identificar los recursos de información en un sistema, están relacionados con el contenido, es decir recogen información sobre lo que un recurso de información contiene o trata. Su principal utilidad está basada en la localización debido a que aportan información que los robots de los sistemas de búsqueda procesan y esto permite a los usuarios realizar una búsqueda y recuperación eficiente de información.

Los metadatos estructurales facilitan la navegación y presentación de recursos electrónicos, además establecen las relaciones intrínsecas o extrínsecas de un documento o conjunto, es decir que proporcionan información sobre la estructura interna de los recursos y establecen las relaciones entre distintos capítulos de un libro, artículos de una revista, o el número de una revista dentro de una colección, incluyen página, sección, numeración, índices y tablas de contenido.

Y finalmente, los metadatos administrativos, estos están relacionados con el contexto de un determinado recurso de información, como principal objetivo, es facilitar la administración y procesamiento de las colecciones digitales, incluyen datos sobre la creación y el control de calidad; comprenden la gestión de derechos y requisitos de control de acceso y utilización, además de la acción de preservación.

A partir de esta clasificación de metadatos, han aparecido una gran cantidad de tipos de metadatos, nombrados de acuerdo a cada uno de los autores que ha realizado, investigaciones acerca de su utilización o funciones.

Isabel Daudinot, menciona que son múltiples los tipos y funciones de los metadatos existentes actualmente, porque estos dependen de varios factores, entre los cuales menciona:

- El tipo de información que describen
- El nivel de estructuración de esta información
- El lugar donde se encuentren los metadatos
- Su ámbito de aplicación
- El tipo de usuarios que los utiliza y sus fines.

Por lo tanto, son múltiples las clasificaciones de metadatos realizadas hasta ahora por diferentes autores.

Bill Prothman identifica dos tipos de metadatos, el primer tipo son los metadatos que describen la entidad de datos o el objeto de información, el segundo tipo son los metadatos que describen su almacenamiento y uso, pero además señala cinco categorías más específicas, estas son:

- Metadatos de acceso: permiten la interrogación, navegación y recuperación de información; describen. Entre otros aspectos cómo se estructuran lógicamente los datos
- Metadatos semánticos: sirven para dotar a la información almacenada de un significado o propósito específico.
- Metadatos de calidad: posibilitan un análisis cualitativo de los datos.
- Metadatos de transferencia: se refiere a cómo los datos pueden transferirse entre aplicaciones.
- Metadatos de almacenamiento: revelan cómo y dónde se almacenan los datos de un sistema

Kashyap y Sheth (2000:20) consideran dos tipos de metadatos, estos son:

- Metadatos independientes del contenido
- Metadatos dependientes del contenido

Los metadatos independientes del contenido son aquellos que recogen la información que no depende del contenido del documento, como la localización, la fecha, modificación, etc.

Los metadatos dependientes del contenido: aquellos que dependen del contenido asociado al objeto de información que describen, recogen información sobre la representación y estructura y facilita la interoperabilidad. Este tipo de metadatos se divide a su vez en:

- Metadatos basados en el contenido directo: aquellos que dependen directamente del contenido del objeto de información.
- Metadatos descriptivos del contenido: describen la información de un documento sin utilizar su contenido.

Para Burnett, Bor Ng y Park existen dos tipos de metadatos:

- Metadatos intrínsecos.
- Metadatos extrínsecos.

Los metadatos intrínsecos incluyen atributos como; materia, título, autor, editor, lugar de publicación, fecha, tipo de objeto de información, forma del identificador URN, relación, fuente, idioma, cobertura, resumen, versión, notas, firma, clasificación, descriptores, etc.

Los metadatos extrínsecos, abarcan información como los requerimientos del sistema, modo de acceso, accesibilidad, costos, control, extensión o tamaño del documento, descripción codificada y descriptores de la revisión.

Méndez identifica diferentes tipos de metadatos, entre los que encontramos:

- Metadatos internos, generados en el momento de creación o digitalización del documento.

- Metadatos externos, relacionados con un documento, pero creados por alguien distinto al que crea la información.
- Metadatos generados automáticamente por el ordenador
- Metadatos creados manualmente en el encabezamiento del documento
- Metadatos creados por personas que no son expertos
- Metadatos creados por especialistas o por expertos
- Metadatos por el control semántico de los datos
- Metadatos controlados; vocabulario estándar; tesaurus, etc
- Metadatos no controlados que no responden a las herramientas anteriores
- Metadatos inmersos en el propio documento
- Metadatos separados del documento
- Metadatos altamente estructurados
- Metadatos no estructurados
- Metadatos estáticos, que no cambian después de creados
- Metadatos dinámicos, que pueden cambiar con el uso y la manipulación del documento.
- Metadatos de larga duración
- Metadatos de corta duración

Gilliland Swetland también presenta una serie de diferentes tipos de metadatos, entre ellos podemos mencionar: (Gilliland, 2000)

- Metadatos Administrativos
- Metadatos descriptivos
- Metadatos preservación
- Metadatos de uso
- Metadatos técnicos

Además de la existencia de diferentes tipos de metadatos, Gilliland también ofrece una serie de tipos de metadatos, por sus características, entre ellos podemos mencionar: (Gilliland, 2000)

- Fuente de los Metadatos

- Método de creación de los Metadatos
- Naturaleza de los metadatos
- Metadatos de status
- Metadatos de estructura
- Metadatos de Semántica
- Metadatos de nivel

Won Kim, menciona que en la práctica hay varios tipos de metadatos, estos son: (Won Kim, 2005:42-43)

- Sistema de catálogos de metadatos
- Metadatos de relación
- Metadatos de contenido
- Metadatos de linaje de datos
- Metadatos técnicos
- Metadatos de uso de datos
- Metadatos del sistema
- Metadatos de proceso

Doug Laney sugiere que hay tres tipos de metadatos: (Doug Laney, 2011)

- Metadatos de definición
- Metadatos de navegación,
- Metadatos administrativos.

Metadatos de definición expresan el significado de los objetos de la empresa (si se dispone de datos, aplicaciones, procesos o recursos) para asegurar su uso constante y el intercambio de toda la empresa y de su evolución. Los metadatos de definición promueven una mejor comprensión de las fuentes de información.

Metadatos de navegación ofrecen enlaces y expresan las relaciones entre objetos de información para garantizar su accesibilidad y facilidad de uso. Los metadatos de navegación ayudan a localizar la información que la cadena de suministro de la información evoluciona.

Metadatos administrativos contiene datos técnicos tales como tipo y modelo de escáner, resolución, profundidad de bit, espacio de color, formato de archivo, compresión, fuente de luz, propietario, fecha del registro de derecho de autor, limitaciones en cuanto al copiado y distribución, información sobre licencia, actividades de preservación (ciclos de actualización, migración, etc.).

Como se aprecia en la lista de los autores anteriores, cada uno de ellos presentan diferentes propuestas de metadatos, por supuesto que esto en lugar de aclarar la situación de los metadatos la hace cada vez más confusa para las personas que no tienen tanta experiencia con el uso de los metadatos, para aclarar un poco más la situación de los tipos de metadatos, utilizaremos la clasificación que propone Tim Berners-Lee, estas son:

- Metadatos administrativos
- Metadatos estructurales
- Metadatos descriptivos

En estas tres grandes categorías, se engloba cada uno de los metadatos antes descritos, desde el punto de vista del tipo de información que describen, su nivel de estructuración, el lugar en el que se ubican, así como su aplicación y finalmente el tipo de usuarios que los usan, esto con el fin principal de evitar confusiones cuando se utilicen para organizar y recuperar información.

TIPOS DE METADATOS

Tipo 1	Tipo 2	Tipo3
Descriptivos	Estructurales	Administrativos
Creados manualmente	Acceso	Almacenamiento
Creados por no expertos	Contenido	Calidad
Externos	Corta duración	Controlados
Generados automáticamente	Control semántico	Creados por especialistas
Independientes del contenido	Dependientes del contenido	Definición
Intrínsecos	Estructura	Dinámicos
No controlados	Estructurados	Estáticos
No estructurados	Inmersos en el propio documento	Extrínsecos

	Método de creación de los Metadatos	Fuente de los Metadatos
	Navegación	Internos
	Nivel	Larga duración
	Relación	linaje de datos
	Separados del documento	Naturaleza de los metadatos
	Técnicos	Preservación
	Trasferencia	Proceso
	Sistema	Semánticos
		Status
		Uso de datos

Finalmente se puede decir que la gran cantidad de tipos de metadatos que existen, condiciona su uso para organizar y recuperar la información en sistemas automatizados, pero es importante conocerlos, así como saber su función o funciones, para utilizarlos de acuerdo a nuestras necesidades.

1.3.3 Funciones de los metadatos

Otro aspecto importante, con respecto a los metadatos son sin duda alguna sus funciones, las cuales según algunos autores se pueden estudiar desde dos puntos de vista, desde el nivel del sistema y del nivel del usuario. (Burnett 1997).

Desde la perspectiva del sistema, los metadatos facilitan la interoperabilidad y la capacidad de compartir datos entre las herramientas de recuperación de recursos, acelerando la concreción de proyectos, mejorando las investigaciones y su utilidad, así como la toma de decisiones y reduce costos al minimizar la duplicación de esfuerzos. Desde la perspectiva de los usuarios, puede facilitar la capacidad de determinar, que datos están disponibles, si satisface necesidades de información específicas y cómo adquirirlo y cómo transferirlo a un sistema local. (Burnett 1997).

Steele menciona que existen dos funciones principales de los metadatos, estos son:

- Proveer un medio para descubrir qué datos existen y cómo se pueden obtener y accederse.
- Proveer un mecanismo de búsqueda para reunir metadatos

Martínez Arellano, menciona tres funciones de los metadatos:

- 1) Proporcionar una descripción de una entidad de información junto con otra información necesaria para su manejo y preservación.
- 2) Proporcionar los puntos de acceso a esa descripción.
- 3) Codificar esa descripción.

Kate Beard menciona cuatro funciones principales de los metadatos, estas son:

- Búsqueda: los metadatos deben proporcionar suficiente información, bien para descubrir si existen datos de interés dentro de la colección de datos disponibles, o simplemente, para saber que existen.
- Recuperación: los metadatos deben proporcionar información a los usuarios para que puedan adquirir la que sea de su interés. información sobre los costes.
- Transferencia: los metadatos deben facilitar la información necesaria para que los usuarios hagan uso de los archivos recuperados en sus máquinas.
- Evaluación: esta función de los metadatos es quizá la más compleja. Los metadatos de apoyo a la evaluación pueden consistir en cualquier información que asista a los usuarios a determinar si los datos van a ser útiles para una aplicación.

Caplan identifica cinco funciones de los metadatos,

- Describir las características y las funcionalidades de los objetos a lo largo del ciclo de vida, desde la creación, las distintas versiones, etc.
- Organización (registro, catálogo e indexación),

- Búsqueda y recuperación,
- Usabilidad (reproducción, modificación)
- Preservación (refresco, migración, chequeo de integridad),

Senso y Rosa Piñero (2003:102), proponen cuatro funciones de los metadatos, estas son:

- Uso: Relativos al nivel y tipo de uso que se hace con los recursos informáticos.
- Técnica: Relativos a cómo funcionan los sistemas o el comportamiento de los metadatos. - Preservación: Para salvaguardar los recursos de la información.
- Descriptiva: Utilizados para representar recursos de la información.
- Administrativa: Usados en la gestión y administración de recursos de la información.

Jones y Taylor (2003) proponen las siguientes funciones de los metadatos:

- Búsqueda en la Web (discovery): Los metadatos de búsqueda en la Web proporcionan suficiente información para permitir discernir el contenido, formato y alcance de un conjunto de datos.
- Exploración: Una vez localizado un conjunto de datos, se tiene que evaluar si son adecuados para satisfacer las necesidades de los usuarios
- Explotación: se relaciona con el proceso de obtención y uso de un conjunto de datos. Estos pueden contener información relacionada con la fuente de los datos y las restricciones de uso (Parker et al. 1996).

La Infraestructura de Información Biológica Nacional de Medioambiente de Canadá, (ECNBII) propone las funciones de:

- Búsqueda en la Web: La mayoría de las partidas de información medioambiental se localizarán en este nivel que es el más elemental, ya se trate de una colección de objetos simples o una base de datos.
- Acceso: Utilizando el perfil geoespacial completo y/o el perfil biológico, este nivel asegurará la descripción detallada y la localización de los datos.
- Uso: Este nivel permitirá el uso de los metadatos biológicos o geoespaciales a nivel de puesto de trabajo, para su visualización y extracción por medio de servicios Web.

Nebert (2004) plantea las funciones como las respuestas a cuestiones para las que el usuario necesita respuestas:

- Metadatos de localización - ¿Qué conjuntos de datos contienen el tipo de datos en el que estoy interesado? Esto permite a las organizaciones conocer y hacer pública la cartera de datos disponibles.
- Metadatos de exploración - ¿Contienen los datos identificados suficiente información como para permitir hacer un análisis razonable según los objetivos? Ésta es la documentación que ha de ser facilitada con los datos para garantizar que otros utilicen los datos correcta y juiciosamente.
- Metadatos de explotación - ¿Cuál es el proceso de obtención y utilización de los datos que se requieren? Esto ayuda a los usuarios finales y a las organizaciones proveedoras a almacenar eficientemente, reutilizar, mantener y archivar sus colecciones de datos.

Danko (2004), Oosteron (2004), Ostensen y Danko (2005) abordan las funciones de los metadatos desde el punto de vista de los elementos añadidos a los metadatos para permitir desarrollar las funciones:

- Localizar: Son ítems de metadatos que permiten a los usuarios localizar la información que buscan y a los productores publicitar sus datos.
- Evaluar: Una vez localizados los datos, se necesitan otros ítems de metadatos que permitan determinar si los datos se adecuan al uso que pretende dar.
- Extraer: En muchas ocasiones los usuarios necesitan acceder a los datos una vez localizados y evaluada su adecuación.
- Emplear: Una vez descargados los datos los usuarios necesitan conocer cómo manejar los datos. Por esta razón se incluyen algunos ítems de metadatos adicionales que describen, entre otros, cómo fusionar y combinar los datos con los suyos propios, aplicarlos correctamente y comprender plenamente sus propiedades y limitaciones.

Johnston (2005) propone las siguientes funciones:

- Localizar: encontrar los recursos de interés, utilizando las descripciones de los recursos que han sido publicadas por sus creadores, distribuidores o terceras partes.
- Evaluar: deben permitir al usuario evaluar la pertinencia del recurso.
- Acceder: deben proporcionar información sobre los mecanismos necesarios para acceder a los recursos una vez que han sido evaluados.
- Interpretar: pueden ayudar a un usuario a interpretar los datos.
- Gestionar: pueden ayudar en otras tareas de gestión a los responsables de los datos, permitiéndoles preservar toda la información relacionada con la propiedad intelectual y los derechos.

Moellering y Brodeur (2006), mencionan las siguientes funciones:

- Localizar: encontrar la ubicación de un conjunto de datos geográficos en referencia a un conjunto específico de características,.

- Evaluar: asegurar que los datos geográficos de la base de datos espacial tienen las características deseadas por el usuario: precisión, vigencia, etc.
- Extraer: transferir la base de datos espacial desde su ubicación, habitualmente a través de la Web, a una localización adecuada para el usuario de los datos.
- Emplear (aplicar): utilizar los metadatos para procesar con éxito la base de datos geográfica, analizar, y quizá resolver, el problema que se tiene entre manos.

Gayatri y Ramachandran (2007) mencionan las siguientes funciones de los metadatos:

- Encontrar / localizar y acceder a recursos:
- Identificación digital: Los identificadores digitales como el nombre del archivo URL y DOI (Identificador Digital de Objeto), que forman parte de los elementos de los metadatos, ayudan en la identificación de los recursos.
- Gestión y organización de los recursos
- Interoperabilidad: los metadatos facilitan la interoperabilidad, puesto que se han definido estándares de metadatos y existen protocolos compartidos.
- Archivo y conservación: Los objetos digitales se corrompen y pueden alterarse fácilmente, de ahí la necesidad de su conservación.

Como se puede apreciar en este apartado también existe una gran cantidad de autores que hablan sobre cuales deben de ser las funciones de los metadatos, pero la mayoría de ellos coinciden en tres aspectos fundamentales, y que podríamos englobar en la propuesta de Martínez Arellano, que sería la siguiente:

- 1) Proporcionar una descripción de una entidad de información junto con otra información necesaria para su manejo y preservación.
- 2) Proporcionar los puntos de acceso a esa descripción.
- 3) Codificar esa descripción.

En este sentido, Isabel Daudinot, menciona “ que los metadatos, aun cuando cumplen con las funciones tradicionales de los catálogos -- Identificación y descripción de la información, búsqueda, recuperación y ubicación de la información – también desempeñan otras funciones inherentes a la información en formato digital como objeto de información, entre las que pueden citarse: limitaciones de uso, valoración del contenido, formas de acceso a la información, autoría y propiedad intelectual, actualización de la información, accesibilidad de los contenidos, preservación y conservación, visibilidad de la información y formas de acceso a la información”. (Daudinot, 2007)

Esta misma autora dice que según su experiencia, las funciones fundamentales de los metadatos son: Describir los objetos de información, informar a los usuarios la existencia, el contenido, la calidad y los objetivos de los recursos de información descritos, con el fin de que ellos puedan determinar si un recurso de información es potencialmente útil, sin necesidad de acceder a este, A su vez, los metadatos deben garantizar el punto de acceso a la información que los usuarios desean consultar. (Daudinot, 2007)

Finalmente podemos decir, que los metadatos se basan principalmente en dos principios fundamentales de la organización de la información, estos son: la identificación del recurso, mediante la descripción del mismo, utilizando los atributos o características de los recursos de información, y la localización de los recursos de información, este principio desde mi muy particular punto de vista es el principal objetivo de los metadatos, debido a que es fundamental para los usuarios recuperar información pertinente para sus actividades académicas y se basa en los puntos de acceso del documento para poder recuperarlo de la manera más sencilla y eficiente.

1.3.4 Importancia de los metadatos

Como podemos observar los metadatos tienen una gran importancia tanto en la organización como en la recuperación de la información. Tomando en cuenta a diferentes autores tales como; Gilliland-Swetland, Ortiz-Repiso, Milstead, Feldman, Senso y De la Rosa Piñeiro entre otros, se presentan algunos aspectos que hacen que los metadatos tomen ese valor de importancia dentro del ámbito de la recuperación de información:

Incrementa el acceso a la información

La existencia de un conjunto de metadatos que describa correctamente uno o varios objetos aumenta la posibilidad de acceder a ellos. Por otro lado, los metadatos hacen posible la búsqueda de información en múltiples colecciones a la vez.

Disminuye el tráfico de la red

Al indizar la representación del objeto, y no el objeto en sí, no requiere demasiado ancho de banda para hacer las búsquedas o generar los índices, por lo tanto la recuperación de información es más rápida.

Expande el uso de la información

Facilitan la difusión de versiones digitales de un único objeto.

Control de versiones

No sólo en lo que se refiere a gestionar la vida de un objeto, sino también en lo que tiene que ver con su difusión, es decir: generar diferentes metadatos con

distintas cantidades de información sobre un mismo objeto con el fin de distribuirla a un público heterogéneo.

Aspectos legales

Los metadatos permiten establecer claramente las restricciones de explotación, informan sobre los derechos de autor, controlan el uso de todo, o una parte, del objeto, tiene control del pago por su uso, controlan el acceso a información restringida.

Preservación del objeto original

La utilización de metadatos junto al uso de lenguajes controlados permite aumentar la precisión en la mayoría de búsquedas en Internet.

1.3.5 Esquema de metadatos Dublin Core

Existen muchos esquemas de metadatos que sirven para organizar la información, de entre ellos El Dublin Core es uno de los esquemas más comunes y usados actualmente en los servicios de información por sus características, así como por la simplicidad para la creación y mantenimiento de metadatos.

La Iniciativa Dublin Core (DCMI <http://dublincore.org/>) comenzó en 1995 en un encuentro en Dublin, Ohio (USA) en el que participaron el NCSA (National Center for Supercomputing Applications) y OCLC (On Line Library Computer Center), junto con representantes de la IETF (Internet Engineering Task Force) y en el que bibliotecarios, proveedores de contenido y expertos en lenguajes de marcado pretendieron desarrollar estándares para describir los recursos de información y facilitar su recuperación.

Así nació un pequeño conjunto de descriptores, en principio pensados para que fuera el propio autor el que los incluyera en el documento o recurso, pero que rápidamente adquirieron alcance global porque también se interesaron en ellos numerosos y variados proveedores de información pertenecientes a diferentes sectores como el de las artes, las ciencias, la educación, los negocios y la administración.

En la actualidad, los metadatos Dublin Core se han convertido en uno de los estándares más extendidos para la recuperación de información en la World Wide Web, además se ha convertido en un vocabulario muy utilizado no sólo en el ámbito bibliotecario y documental, sino en otros muchos sectores.

Además, este conjunto de metadatos se puede utilizar no sólo con HTML, sino sobre otros lenguajes estructurados como XML y conjuntamente con otros lenguajes de descripción como RDF. The Dublin Core metadata element set se convirtió en norma ISO 15836/2003 en febrero de 2003. (Lamarca Lapuente)

Los metadatos Dublin Core tratan de ubicar, dentro de Internet, los datos necesarios para marcarlos, así como describir, identificar, procesar, encontrar y recuperar un documento introducido en la red.

A continuación se presenta el Esquema Dublin Core que cuenta con 15 elementos:

1.- Título	Etiqueta: Title
2. Creador	Etiqueta: Creator
3. Materia	Etiqueta: Subject
4. Descripción	Etiqueta: Description
5. Editor	Etiqueta: Publisher
6. Colaborador	Etiqueta: Contributor
7. Fecha	Etiqueta: Date
8. Tipo de recurso	Etiqueta: Resource Type
9. Formato	Etiqueta: Format
10. Identificador	Etiqueta: Resource Identifier
11. Fuente	Etiqueta: Source.
12. Idioma	Etiqueta: Language
13. Relación	Etiqueta: Relation
14. Cobertura	Etiqueta: Coverage
15. Derechos	Etiqueta: Rights Management

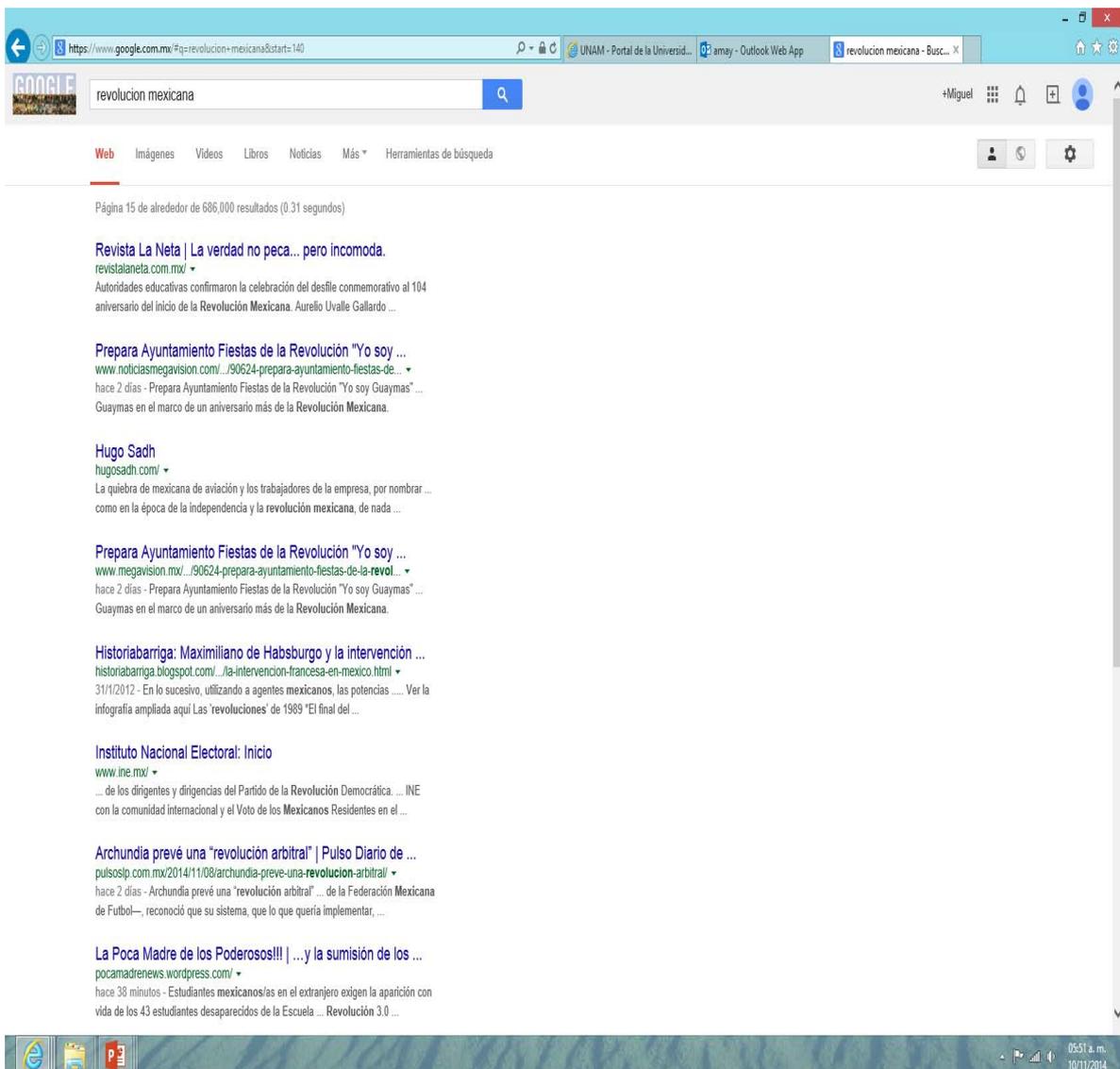
Como se aprecia en párrafos anteriores, utilizar metadatos, así como esquemas de metadatos en la organización y recuperación de información, permite tener una amplia gama de beneficios, como son el control de la información y mejorar las posibilidades de los usuarios al obtener la información más pertinente para sus investigaciones, en este sentido también ha despertado la curiosidad de algunos autores que han mencionado que al recuperarse la información más eficientemente por usar metadatos en la red, se podría pensar en la Web semántica.

1.4 Web semántica

1.4.1 Web semántica: definición

Desde 1989, año en el que nació la World Wide Web (WWW) en los laboratorios del CERN, la Web se ha convertido en un espacio de información global formado por miles de millones de páginas. Este gran número de páginas Web existentes, garantiza la existencia de información sobre casi cualquier tema, al tiempo que complica cada vez más encontrar la información que se busca.

La causa de este problema proviene del hecho de que la principal forma que utilizan los usuarios para buscar información es por el tema, esto ha ocasionado que los buscadores de información recuperen mucha información, debido a que los temas carecen de una estructura semántica, es decir que están fuera de contexto, cuando un usuario busca información sobre cualquier tema en un buscador convencional, hace lo siguiente: si busca información sobre la revolución mexicana, utiliza los términos **revolución mexicana** y el buscador recupera 850,000 resultados, una cantidad impresionante, que por supuesto no todos estos resultados tienen que ver con la Revolución Mexicana, entre los resultados que no tienen que ver con el tema deseado, encontramos los siguientes:



Hospital Regional "Centenario de la Revolución Mexicana"
PREMIO · En Investigación Histórica sobre La Revolución Mexicana
Letras de Corridos Mexicanos - Crónica de la revolución
INSTITUTO NACIONAL DE ESTUDIOS HISTÓRICOS DE LA REVOLUCIÓN MEXICANA, INEHRM
Hugo Brehme (1882 – 1954). Fotógrafo de la Revolución Mexicana.
Chocan Tiida y Platina sobre calzada Revolución Mexicana
Esc. Revolución Mexicana - México LinkedIn
Libreprensa - Parque Ecológico Revolución Mexicana
Estafeta Mexicana S.A. de C.V.
Revolución Tres Punto Cero –

Se observa en este ejemplo, como el buscador recupera muchos documentos o páginas Web que no tienen nada que ver sobre el tema de “Revolución Mexicana” y esto se debe precisamente a que los buscadores convencionales recuperan la información por los términos que utilizamos pero no permiten identificar el contexto al que se refiere o necesita el usuario, esto ocasiona que el usuario pierda tiempo e interés al revisar los resultados de su búsqueda.

Por lo tanto la función principal de las herramientas que deberán integrar la Web semántica será la de identificar semánticamente los temas, para que los usuarios puedan recuperar la información que necesitan sin tener que revisar miles de documentos, recuperando solamente los que sean relevantes y pertinentes para satisfacer sus necesidades de información, por esta razón estas herramientas tendrán que apoyarse en tres aspectos importantes, estos son:

1. El uso de lenguajes de marcado, como el XML (eXtensible Markup Language)
2. El uso de RDF
3. El uso de ontologías

Antes de explicar cada uno de estos tres elementos, es indispensable conocer como definen algunos autores el concepto de Web semántica, esto principalmente porque algunos autores mencionan que la Web semántica ya está entre nosotros de diversas formas.

En 1998, Tim Berners-Lee publicó en la página principal del *World Wide Web Consortium* (World Wide Web Consortium, 2010a) (W3C en adelante) su conocido artículo —*A Roadmap to the Semantic Web*” (Berners-Lee, T., 1998). En dicho artículo se introdujo por primera vez el término Web Semántica (*Semantic Web*) destacándose la necesidad de expresar la información de forma que ésta fuera procesable por máquinas. El artículo presentaba un conjunto de pasos hasta lograr una Web en la que el razonamiento fuera automático, llevado a cabo y distribuido por máquinas.

En 2001, Berners-Lee publica un nuevo artículo *“The Semantic Web”* (Berners-Lee, T.; Hendler, J.; Lassila, O., 2001), en el que presenta las principales características que tendrá la futura Web y afirma que las máquinas facilitarán nuevas prestaciones al mejorar su capacidad de procesar y comprender la información dispersa por la Web. Como solución a las limitaciones semánticas de la actual Web, propone hacer procesable de forma automática el contenido de la Web.

Asimismo, Berners-Lee, menciona que la Web semántica es una extensión de la actual Web, cuyo objetivo es que no sólo los humanos, sino también las máquinas, sean capaces de 'comprender' el contenido de los documentos. Es decir que la Web Semántica no sería una nueva Web sino la extensión de la existente, mediante la adición de metadatos que describan la semántica de las páginas de forma procesable por máquinas.

Ya se ha mencionado, que la Web semántica es el proyecto del W3C para transformar la Web en la Web del futuro, y la definen de la siguiente manera.

La Web semántica proporciona un marco común que permite que los datos sean compartidos y reutilizados a través de aplicaciones, empresas y fronteras comunitarias. Es un esfuerzo colaborativo liderado por el W3C con la participación de un gran número de investigadores y socios industriales. Está basado en Resource Description Framework (RDF) e integra una variedad de aplicaciones utilizando XML para la sintaxis y URI para las denominaciones (www.w3.org/2001/sw/)

Rodríguez Perojo y Ronda León mencionan que la Web semántica es una extensión de la Web cuya idea básica es tener los datos definidos y relacionados para que su uso sea más efectivo y sea posible su automatización, integración y re-utilización por medio de diferentes aplicaciones, es decir, pretende proporcionar una infraestructura que permita que las páginas Web, las bases de datos, los programas y aplicaciones, los dispositivos, tanto personales como los empleados en el hogar, puedan consumir y producir datos, sin los problemas causados por los diferentes

protocolos de acceso a la información que hacen de la transferencia de contenidos una tarea ardua y difícil. El soporte principal para la organización, almacenamiento y distribución de la información siguen siendo los sistemas taxonómicos, es decir, sistemas de clasificación que respondan a necesidades concretas de las empresas, bibliotecas y centros de información en favor no sólo de los usuarios humanos, sino de las máquinas también.(Rodríguez Perojo y Ronda León,)

Lluís Codina y Cristòfol Rovira proponen dos definiciones separadas de la Web semántica:

Definición 1. La visión de la Inteligencia Artificial:

La Web semántica es un conjunto de iniciativas destinadas a promover una futura Web cuyas páginas estén organizadas, estructuradas y codificadas de tal manera que los ordenadores sean capaces de efectuar inferencias y razonar a partir de sus contenidos.

Definición 2. La visión del procesamiento robusto:

La Web semántica es un conjunto de iniciativas destinadas a convertir la World Wide Web en una gran base de datos capaz de soportar un procesamiento sistemático y consistente de la información. (Codina, Rovira 2006:2)

Contrastando estas definiciones, con la que ofrece Tim Barners-Lee, principalmente con la primera, se contradicen estos autores debido a que la Web semántica no tiene nada que ver con los algoritmos complejos y poderosos que se plantean en la inteligencia artificial.

Así lo menciona el mismo Tim Barners-Lee, cuando dice que la Web semántica no es inteligencia artificial, el concepto de documentos que sean entendidos por máquinas no implica alguna inteligencia artificial mágica, que permita a las máquinas comprender balbuceos humanos. Solo indica la habilidad de la máquina para resolver un problema bien definido haciendo operaciones bien

definidas sobre datos existentes bien definidos. En vez de pedir a las máquinas que entiendan el lenguaje de los humanos, esto involucra pedirle a los usuarios un esfuerzo extra, el esfuerzo consistirá en estructurar los datos de una manera más ordenada y estandarizada. (Barners-Lee, 1998)

Otra definición la dan los autores Daconta, Obrst y Smith, quienes mencionan que la Web semántica podría expresarse como una Web de datos inteligentes procesables por máquinas, donde los datos inteligentes son independientes de las aplicaciones, compuestos, clasificados, y parte de un gran ecosistema de información.(Daconta, Obrst y Smith, 2003)

Asimismo estos autores mencionan que la clave de la Web semántica esta en darle mayor importancia a los datos, de la que tienen en la actual estructura de la Web. Esto se debe a que tradicionalmente, los datos han estado en segundo plano con respecto al procesamiento de los mismos. De esta forma la Web semántica propone un cambio de paradigma, en donde el poder no esté en las aplicaciones, sino en los datos. (Daconta, Obrst y Smith, 2003)

Castells opina que la idea es que la Web semántica esté formada (al menos en parte) por una red de nodos tipificados e interconectados mediante clases y relaciones definidas por una ontología compartida por sus distintos autores.

Asimismo, dice que la adopción de ontologías comunes es clave para que todos los que participen de la Web semántica, contribuyendo o consumiendo recursos, puedan trabajar de forma autónoma con la garantía de que las piezas encajen. Es decir, contar con un programa que navegue por una red como ésta puede reconocer las distintas unidades de información, obtener datos específicos o razonar sobre relaciones complejas.

Castell también menciona que la Web semántica propone describir los recursos de la Web con representaciones procesables (es decir, entendibles) no sólo por personas, sino por programas que puedan asistir, representar, o reemplazar a las personas en tareas rutinarias o inabarcables para un humano. Las tecnologías de la Web semántica buscan desarrollar una Web más

cohesionada, donde sea aún más fácil localizar, compartir e integrar información y servicios, para sacar un partido todavía mayor de los recursos disponibles en la Web.(Casstell)

Se puede apreciar en las definiciones anteriores que, tanto Berners como los otros autores coinciden, en primera instancia en que la Web semántica, no es una nueva Web, sino que sería una extensión de la ya existente y que la información contenida en ella tiene que ser procesada por máquinas, además se menciona que en la Web semántica lo importante será la forma en que estén almacenados los datos, y no la forma en que se buscan. Es decir que la Web semántica todavía no es una realidad, además para que sea una realidad lo que tendría que cambiar son las herramientas de recuperación de información, como son los buscadores, bases de datos, metabuscadores, etc., por lo tanto lo que tiene que evolucionar en el Web son las herramientas de recuperación de información con un enfoque semántico.

Entonces se puede decir, que la Web Semántica es una extensión de la World Wide Web en la que los contenidos de la Web pueden ser expresados mucho más que en un lenguaje natural, es decir con un enfoque semántico y que se identifique el contexto de los datos, también en un formato que pueda ser entendido, interpretado y usado por diferentes softwares, permitiéndoles buscar, compartir e integrar información más sencilla y más fácil , además su objetivo principal será crear un medio universal que permita el intercambio de datos y brindar un mayor significado a la misma para que puedan ser interpretadas por las máquinas.

1.4.2 Lenguaje de marcado XML (eXtensible Markup Language)

XML o eXtensible Markup Language, es un lenguaje de marcado derivado del lenguaje SGML que proporciona una sintaxis superficial para documentos estructurados, pero que no impone restricciones semánticas sobre el

significado de los mismos. XML permite suministrar, recibir y procesar información en la Web y puede interoperar tanto con SGML como con HTML.

XML es un estándar desarrollado por el Consorcio World Wide Web (W3C). Se trata de una versión reducida y especializada en la Web de la norma SGML. Su caracterización como "extensible" se deriva de la no limitación en el número de marcas o etiquetas, pues permite crear todas aquéllas que sean necesarias. Otra de sus características principales es que permite enlaces multidireccionales (esto es, que apuntan a varios documentos). (Lamarca Lapuente)

El estándar XML, es uno de los componentes básicos de la Web semántica y del nuevo entorno de la organización de la información, puesto que permite una codificación "semántica" de cada uno de los atributos o elementos de una página Web o de cualquier otro documento para que éstos tengan un sentido propio, de manera que puedan ser interpretados y manejados por una máquina. XML constituye un lenguaje de marcado altamente estructurado, a través del cual "podemos disponer de etiquetas como <título>, <subtítulo>, <capítulo>, <autor>, <institución>, <ciudad>, etc.", o bien, "si una empresa o institución necesita almacenar y procesar información sobre los currículum vitae de sus empleados, puede desarrollar un Schema XML que le permita crear documentos XML bien formados que dispongan de etiquetas como <lugar_nacimiento>, <titulación_académica>, <idiomas_hablados>, <experiencia_laboral>, etc." (Codina & Rovira, 2006:9).

1.4.3 Marco de Descripción de Recursos (RDF)

RDF es un lenguaje para representar información sobre recursos en la World Wide Web. Está particularmente indicado para representar metadatos sobre recursos Web, tales como el título, autor, modificaciones de los datos de la página Web, copyright y otras licencias de información sobre documentos

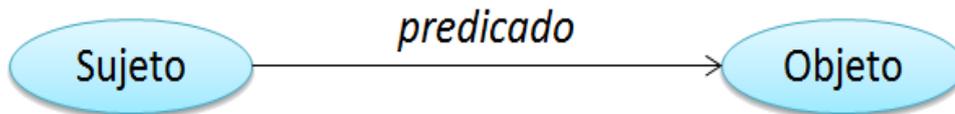
Web, así como la disponibilidad para algunos recursos compartidos. Se trata de un modelo de datos para objetos (recursos).

RDF puede también usarse para representar información sobre ciertas cosas que pueden ser identificadas en la Web, aunque no puedan ser directamente recuperadas en la misma, por ejemplo, información sobre artículos disponibles desde servicios online (como información sobre especificaciones, precios, disponibilidad, etc.), o la descripción de las preferencias de los usuarios de la Web para obtener información.

El lenguaje RDF o Infraestructura para la Descripción de Recursos es muy útil en situaciones en las que la información necesita ser procesada por aplicaciones que intercambian información legible por máquina, más que por humanos. RDF provee un marco común de trabajo para expresar esta información y para intercambiarla entre aplicaciones distintas mediante una serie de "parsers" o analizadores RDF y otras herramientas de procesamiento automatizado.

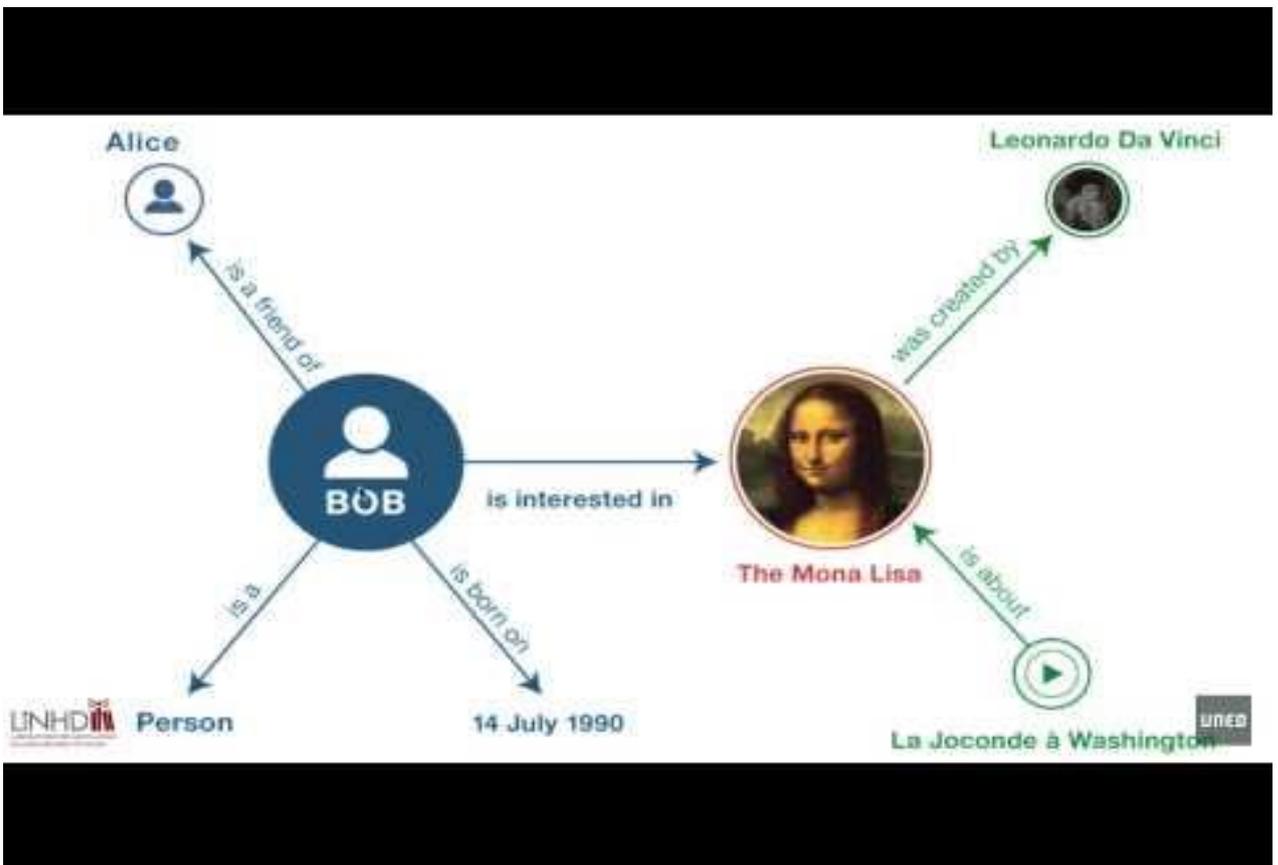
RDF puede utilizarse en diferentes áreas como en la recuperación de recursos para los buscadores, robots y agentes inteligentes, catalogación para describir el contenido y las relaciones de contenido disponibles en un sitio Web o en una colección de documentos, para describir los derechos de propiedad intelectual o las políticas de privacidad de un sitio Web, etc.

RDF está basado en la idea de identificar los recursos en la Web usando los Uniform Resource Identifiers o URIs, y describiendo los recursos en términos de propiedades simples y valores. Una descripción RDF es un conjunto de proposiciones simples (también llamadas sentencias o declaraciones) y una proposición se conoce también como una tripleta, porque está compuesta de 3 cosas: un sujeto, un predicado y un objeto.



Componentes de una sentencia o una proposición

- **Sujeto:** Es el recurso al cual nos estamos refiriendo.
- **Predicado:** Es el recurso que indica lo que estamos definiendo.
- **Objeto:** Puede ser un recurso que puede considerarse el valor definido



Ejemplo de una tripleta, tomado de:

https://www.youtube.com/watch?v=VX8hYjJs4AQ&feature=player_detailpage

En este ejemplo se puede explicar de mejor manera la función de una tripleta en un sistema de información (catalogo, buscador, metabuscador, directorio, etc.), en este ejemplo el sujeto es representado con el nombre de Bob, que a su vez tiene otras relaciones, representadas como datos, su fecha de nacimiento, además Bob es una persona y tiene una amiga denominada Alice. A continuación sigue el predicado el cual une al sujeto con el objeto,

representado con la imagen de la Mona Lisa una pintura, la cual tendría relación con el video con un video que habla de ella, llamado la Gioconda a Washington y finalmente otra relación sería con Leonardo Da Vinci, su creador.

Por lo tanto, se puede decir que el RDF se basa en la idea de convertir las declaraciones de los recursos en expresiones con la forma sujeto-predicado-objeto (conocidas en términos RDF como tripletes).

El sujeto es el recurso, es decir, es lo que se está describiendo. El predicado es la propiedad o relación que se desea establecer acerca del recurso. Por último, el objeto es el valor de la propiedad o el otro recurso con el que se establece la relación.

La combinación de RDF con otras herramientas como RDF Schema y OWL permite añadir significado a las páginas, y es una de las tecnologías esenciales de la Web semántica.

1.4.4 Ontologías

En este apartado se trata de explicar a detalle el primer elemento que constituye o deberá integrarse por las herramientas de recuperación de información que se encontraran o conformaran la Web semántica.

Por lo tanto para entender un poco más hacia donde se dirige este concepto de Web semántica, es necesario conocer que aspectos son importantes para explicar más a detalle la función de la Web semántica. Para esto se tendrá como apoyo el esquema de arquitectura de la Web semántica que presenta Tim Barners-Lee en el cual utiliza un rubro conocido como vocabulario ontológico.

Para dar inicio con la definición, no de vocabulario ontológico, sino, el de Ontología, primero tenemos que mencionar que el término ontología proviene

del campo de la Filosofía. Por lo tanto la Ontología es una rama de la metafísica que se interesa en la identificación de los tipos de cosas que existen y cómo describirlos. Provee una clasificación exhaustiva de los objetos y sus relaciones. Sin entrar en la validez de su significado, constituye un tipo de instrumento que permite la representación del conocimiento en un área determinada en clara conexión con su recuperación en entornos informáticos.

Desde hace ya varias décadas el término ontología ha tenido una gran relevancia en el ámbito bibliotecario, esto se debe al interés en el desarrollo de la Web semántica, es decir que se busca que la Web tradicional no sea solamente un lugar donde hay información, sino sea un lugar donde se adquiera conocimiento.

En este contexto, se debe aclarar que existe una gran cantidad de definiciones sobre la ontología, entre las cuales se puede mencionar a:

Neches (1991) quien dice que una ontología es un vocabulario acerca de un dominio: términos + relaciones + reglas de combinación para extender el vocabulario.

En este sentido, Gruber (1993) menciona que en realidad el término ontología, desde el punto de vista de la Filosofía, es una explicación sistemática de la existencia, pero desde el punto de vista de las ciencias de la computación, ésta es una representación formal del conocimiento por un conjunto de conceptos dentro de un dominio y las relaciones entre los conceptos y desde el punto de vista de la Bibliotecología la ontología es una especificación formal y explícita de una conceptualización compartida.

Por lo tanto la ontología proporciona un vocabulario compartido, que puede ser utilizado para modelar un dominio por el tipo de objeto, y / o concepto, sus propiedades y relaciones.

Asimismo Gruber (1993), dice que la ontología se utiliza para razonar acerca de las propiedades de ese dominio y puede ser usado para describir el dominio mismo.

Por otro lado, Borst (1887) comenta que una ontología es una especificación formal de una conceptualización compartida. Borst, 1997. (Aquí el término forma se refiere a que es procesable por ordenador).

Weingand (1997) dice que una ontología es una base de datos que describe los conceptos generales o sobre un dominio, algunas de sus propiedades y cómo los conceptos se relacionan unos con otros.

Otros autores como, Studer, Benjamins y Fensel (1998:161-198), mencionan que una ontología es una especificación explícita y formal de una conceptualización.

Rodríguez Perojo y Ronda León mencionan que las ontologías se utilizaron tradicionalmente como modelo de representación de conocimiento en la inteligencia artificial. En este sentido estos mismos autores, aseveran que una ontología estará formada por una taxonomía relacional de conceptos y por un conjunto de axiomas o reglas de inferencia mediante los cuales se podrá inferir un nuevo conocimiento.

Otra definición, es la que ofrece Neches, quien menciona que una ontología es un instrumento que define los términos básicos y relaciones a partir del vocabulario de un área de conocimiento, así como las reglas de combinación de estos términos y relaciones para definir extensiones a un vocabulario.

Una Ontología describe formalmente un dominio de discurso, consiste en una lista finita de términos y las relaciones entre esos términos. Los términos denotan conceptos importantes (o clases de objetos) del dominio. (Antoniou y otros, 2004.)

Breitman (2007) menciona que el objetivo de una Ontología es capturar y hacer explícito el vocabulario utilizado en aplicaciones semánticas, garantizando de esta manera una comunicación libre de ambigüedades y un dominio es un área de conocimiento específica que es centro de atención de una comunidad de interés, por lo tanto una ontología define los términos usados para describir y representar un dominio.

De acuerdo a Seidenberg y Rector, las ontologías pueden añadir un gran valor a las tecnologías Web, debido a que “el conocimiento capturado en las ontologías puede ser usado, entre otras cosas, para anotar datos, distinguir entre homónimos y polisemia, generalizar o especificar conceptos, dirigir interfases de usuario inteligentes e incluso inferir completamente nueva información” .

Maedche y Staab entienden por Ontologías a aquellos “esquemas de metadatos, que proveen de un vocabulario controlado de conceptos, cada uno con una semántica explícitamente definida y procesable por la máquina”.

Siendo así que las Ontologías ayudan tanto “a la gente y a las máquinas a comunicarse consistentemente, soportando los cambios semánticos y no solamente los cambios en sintaxis”. Sin embargo la elaboración de Ontologías contiene tres tipos de problemas:

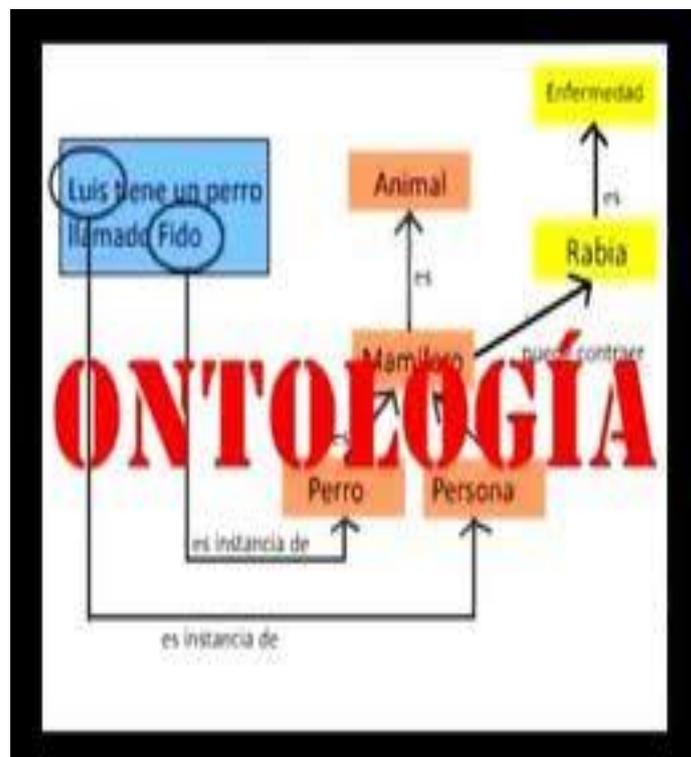
1. De tiempo
2. De dificultad
3. De confidencialidad.

Asimismo, Gruber (1993) asegura que las ontologías son utilizadas en:

- Inteligencia Artificial
- La Web Semántica
- En Ingeniería de Sistemas
- En Ingeniería de Software
- Informática Biomédica

- Arquitectura de la Información y
- Bibliotecología

En términos bibliotecarios, se puede decir que una ontología podría considerarse algo parecido a un tesoro, pero más estructurado, ambos son considerados sistemas de organización del conocimiento, debido a que los dos son listas relacionadas, es decir, que ambos sirven para estructurar el conocimiento por medio de vocabularios controlados. Lamarca Punte menciona que la diferencia entre los tesauros y las ontologías radica en la complejidad, ya que las ontologías introducen un mayor nivel de profundización semántica y proporcionan una descripción lógica y formal que puede ser interpretada tanto por las personas, como por las máquinas, mientras que los tesauros sólo pueden ser interpretados por humanos, asimismo menciona que las ontologías permiten la interoperabilidad entre distintos sistemas.



Ejemplo de una Ontología, tomado de:

<http://administracionconocimiento6.blogspot.mx/2010/11/64-herramientas-para-crear-y-mantener.html>

En este ejemplo, se puede ver, la relación que existe entre dos términos, desde el punto de vista de su conceptualización. Luis es una persona, que además es mamífero. Tiene un perro que se llama Fido, que también es un mamífero, Fido es un animal, que puede contraer rabia y la rabia es una enfermedad.

Por lo tanto las ontologías son herramientas que sirven para estructurar conceptualmente determinados ámbitos del conocimiento por medio de vocabularios controlados, las ontologías introducen un mayor nivel de profundización semántica y proporcionan una descripción lógica y formal que puede ser interpretada tanto por las personas, como por las máquinas. Las ontologías permiten, además, la interoperabilidad entre distintos sistemas.

1.5 Características de la Web semántica

Definitivamente hablar de la Web semántica, es hablar de un área importante en donde convergen las tecnologías Web, que propone introducir descripciones explícitas sobre el significado de los recursos, para permitir que las propias máquinas tengan un nivel de comprensión de la Web suficiente como para procesar un trabajo que actualmente realizan manualmente los usuarios que navegan por la Web.

Frente a la semántica implícita, el incremento de los recursos, y la falta de organización de la Web, la Web semántica propone clasificar, dotar de estructura y anotar los recursos con semántica explícita procesable por una computadora.

De esta manera surgen algunas características que identifican a la Web semántica, entre ellas podemos mencionar las siguientes:

- Estructura mejor la información con el fin de poder localizar y encontrarla de forma más sencilla

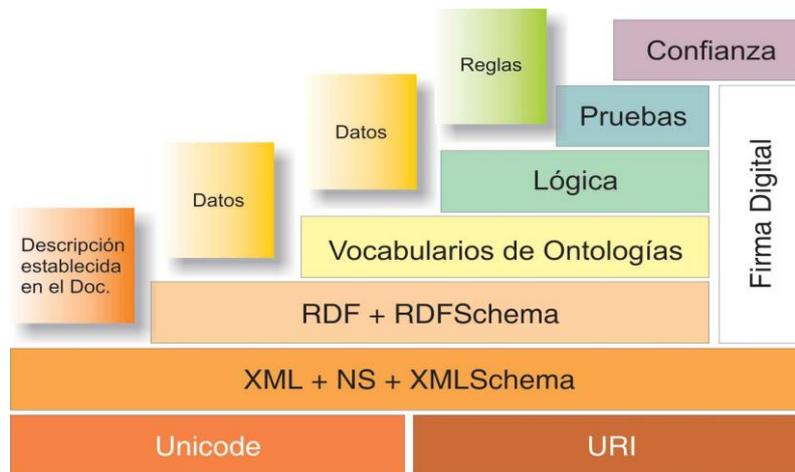
- Se basa en metadatos que permiten razonar en la Web
- Utiliza un lenguaje universal que permite el intercambio con otros
- Posee herramientas capaces de procesar la información de manera sencilla
- Es estándar y permite que los contenidos puedan ser utilizados y entendidos por cualquier software.
- Los navegadores distinguen más que páginas HTML , con lenguajes como XMTL Y RDF
- Posee orden y flexibilidad.

1.6 Arquitectura de la Web Semántica

La idea de una Web, en donde las máquinas puedan entender el significado de su contenido, suena bastante interesante y además sería de gran utilidad para los usuarios que necesitan información.

Rodríguez Perojo y Ronda León, en su documento “Web semántica: un nuevo enfoque para la organización y recuperación de información en la Web”, hacen una esquematización del contenido del diagrama presentado por Berners-Lee en la Conferencia XML del año 2000, puede servir como aproximación visual al conjunto de tecnologías que forman el esquema de capas.

En la siguiente imagen se muestra una taxonomía de los capas de la Web semántica:



Modelo de capas propuesto por Berners-Lee para la Web semántica

Los componentes mencionados, son:

- Estándares para la localización de recursos de información en el Web de forma inequívoca y única como son los URIs (Uniform Resource Identifiers) y la norma internacional Unicode para la codificación de caracteres a nivel internacional.
- XML (eXtensible Markup Language), como base sintáctica para la estructuración del contenido en el Web, así como el empleo de espacios de nombres (Namespaces) para asociar con precisión cada propiedad con el esquema que define dicha propiedad y esquemas (XML Schema) para definir qué elementos debe contener un documento XML, cómo están organizados, qué atributos y de qué tipo pueden tener sus elementos.
- Un modelo básico para establecer propiedades sobre los recursos, para el que se empleará RDF (Resource Description Framework), así como un modelo para definir relaciones entre los recursos por medio de clases y objetos, que se expresan mediante esquemas en RDF
- Lenguajes para la representación de ontologías que permitan la interoperabilidad y reutilización entre ontologías de diversos dominios del conocimiento en la Web, cuya base se encuentra en RDF Schema .

- Una capa lógica que permita realizar consultas e inferir conocimiento, donde estarían las ontologías, agentes software y web services como estructuras para lograr interoperabilidad entre aplicaciones y sistemas de información heterogéneos.
- Una capa de seguridad que permita asignar niveles de fiabilidad a determinados recursos, de forma comprobable posteriormente por los agentes, para lo que se utilizarán firmas digitales y redes de confianza.

Se aprecia que la Web Semántica, se basa principalmente en un principio de la organización de la información y este es: el de localizar los objetos de información, para esto se ayuda de las ontologías, que son una lista finita de términos y las relaciones que existen entre ellos, asimismo necesita como describir estos términos y para ello se auxilia de un Marco de descripción de Recursos (RDF) este es un Marco para metadatos en la World Wide Web, creado por W3C "World Wide Web Consortium".

Se puede mencionar, que definitivamente existe una relación muy estrecha entre la organización de la información y la recuperación de la misma, por lo tanto, para que la Web, pueda ser esa Web que ayude a los usuarios a recuperar la información que necesitan de una manera más precisa, es necesario que se utilicen las herramientas que se detallaron brevemente en este capítulo, en la construcción de ella, como son: el uso de metadatos, una estructura de metadatos, como el Dublin Core, también es necesario que se tome en cuenta un lenguaje de marcado como el XML y un marco para la descripción de recursos como el RDF y también se creen ontologías que permitan que le den un significado y una semántica a la información que se encuentra en la Web y en las bibliotecas digitales. Finalmente, esta técnica de crear ontologías y utilizar las tripletas de RDF, se puede aplicar a la construcción de la Web semántica, debido a que todavía ésta no es una realidad visible en la red y para poder llegar a tenerla es indispensable que se generen herramientas que organicen la información y puedan recuperarla eficientemente y se pueda transformar en realidad la Web a una Web semántica.

CAPITULO 2

RECUPERACIÓN DE INFORMACIÓN

2.1 La recuperación de información

En la actualidad es muy común utilizar la Web para buscar e investigar sobre un tema en particular, de esta forma cobra vital importancia el término recuperación de información. Es pertinente señalar que existe una gran variedad de definiciones sobre la recuperación de información, una de las definiciones más acertada es la que ofrece Salton (1983), cuando menciona que “la recuperación de la información tiene que ver con la representación, almacenamiento, organización y acceso a los ítem de información”.

Por otro lado autores como Croft (1987); Feather y Storges (1997); Tramullas Sanz (1997); Baeza-Yates (1999), coinciden en que el concepto de recuperación de información, tiene que ver con la labor fundamental del profesional de la información, en las cuestiones de organización de la información, para su posterior recuperación.

Por lo tanto podemos decir que la recuperación de información, es el proceso de organización, representación, almacenamiento y recuperación eficiente de la información, para satisfacer las necesidades de información de los usuarios que tienen dudas sobre algún tema en particular.

Asimismo el proceso de recuperación de información se lleva a cabo mediante consultas a diferentes herramientas de búsqueda en donde se almacenan los índices estructurados, de la información que se encuentra en la Web, mediante un lenguaje de interrogación adecuado. Entre los elementos que permiten realizar una búsqueda con un mayor grado de pertinencia y precisión, se encuentran: los índices, las palabras clave, los tesauros, etc.

Si bien es cierto que la Web cuenta con una gran cantidad de información que se puede utilizar, también cuenta con diferentes herramientas que nos permiten

recuperar la información pertinente para satisfacer las necesidades de información de los usuarios además de las propias.

2.2 Herramientas de recuperación de información

La importancia que tienen las herramientas de búsqueda de información en el mundo de la World Wide Web, sin duda alguna es definitiva, esto debido principalmente a que ellas son el pasaporte con el que cuentan los usuarios de Internet para poder acceder a la información que se encuentra almacenada en la red de redes.

Existen muchas herramientas de búsqueda de información en la WWW, con características propias que las hace diferentes unas de otras, pero sin duda alguna las herramientas más conocidas y más utilizadas por los usuarios son los llamados buscadores.

Antes de definir las características de las diferentes herramientas de búsqueda de información, comenzaremos por presentar sus antecedentes.

2.2.1 Antecedentes

Las herramientas de búsqueda de información están asociadas a Internet desde los mismos comienzos de la Red de Redes, de entre ellas las más comunes son los llamados buscadores, conocidos como motores de búsqueda, en un estudio que se realizó en la Universidad de Leiden en Holanda se menciona que el primer buscador fue llamado "Archie", el cual fue creado por Alan Emtage en 1990, estudiante de ciencias informáticas de la Universidad McGill (Montreal, Canadá).

Este indexaba nombres de archivos, creando una base de datos abierta que daba como resultado de la búsqueda, los archivos que coincidían con la palabra buscada.

Al año siguiente en 1991 Mark McCahill crea Gopher, que permitió la aparición de dos nuevos programas de búsquedas: Veronica y Jughead. Estos programas buscaban nombres de archivos alojados en los sistemas de índices de Gopher.

En junio de 1993, Matthew Gray presentó la primera araña o robot rastreador. El Robot, bautizado como el "Errante de la Red Global" World Wide Web Wanderer. Tardó dos años en rastrear toda la Red. El índice que produjo se denominó Wandex.

Unos meses después, el 2 de Septiembre de 1993 apareció lo que puede considerarse propiamente como el primer motor de búsquedas: W3Catalog. Fue desarrollado por Oscar Nierstrasz en la Universidad de Ginebra y fue retirado el 8 de diciembre de 1996.

El motor de búsqueda, Aliweb, se lanzó en Noviembre de 1993 por su creador Martijn Koster, en la Primera Conferencia Internacional sobre el mundo Web. Aliweb era un motor de búsqueda de datos, lo que le permitía agregar descripciones y palabras claves escritas para encontrar las páginas y no la dirección donde se encuentra (que era como se utilizaba Wandex). Además no utilizaba un robot de búsqueda, eliminando así la lentitud del motor por afluencia masiva, tal como sucedía con Wandex. Actualmente es considerado como el buscador más antiguo activo de la red.

En diciembre de 1993, un mes después, surgió JumpStation, creado por Jonathon Fletcher y se encontraba alojado en la Universidad de Stirling (Escocia, Reino Unido).

Este contaba ya con las características básicas de los actuales buscadores. JumpStation procesaba sólo los títulos de las páginas Web y sus principales encabezados para construir su índice; lo hacía sirviéndose de un robot para capturar la información, ofrecía resultados con base a palabras clave y presentaba esos resultados en listas de direcciones Web que coincidían con la palabra buscada, a finales de 1994 el motor contaba con un poco más de 270 000 páginas Web indizadas.

El 20 de abril de 1994, con Webcrawler, el primer motor de búsqueda de texto completo, que rastreaba e indexaba todas las palabras de una página Web y no sólo las principales. En la actualidad ya no se le considera un buscador, sino un metabuscador que recupera los principales resultados en sus búsquedas de Google, Yahoo, Bing.

También en el año de 1994, surgido como un proyecto de desarrollo de un motor de búsqueda liderado por el Dr. Michael Mauldin de la Universidad Carnegie Mellon, nació otro buscador llamado Lycos, con un gran impulso comercial. Llegó a ser uno de los sitios Web más visitados de internet con una presencia global en alrededor de 40 países.

Poco tiempo después de la aparición de estos dos últimos buscadores, aparecieron nuevos buscadores como Maguellan, Infoseek, Excite, Inktomi, Northern Light, AltaVista y Yahoo. En 1995 su uso ya comenzaba a generalizarse en Estados Unidos y también se comenzaba a utilizar en México, en especial en la UNAM, en el departamento de consulta de la Dirección General de Bibliotecas.(gaceta unam)

A finales de la década de los 90s, en el año de 1998, un buscador llamado Google crecía a pasos agigantados. Por su parte, en ese mismo año, la empresa Microsoft, lanzó su buscador MSN Search, usando como base el motor de Inktomi. Seis años después, Microsoft comenzó la transición a su propio buscador, que usa su propio robot (msnbot), relanzado con fuerza en

2009 con el nombre de Bing y Google se ha consolidado como uno de los buscadores con más usuarios.

2.3 Tipología de herramientas de recuperación de información en la Web

Para organizar la información y recuperarla se han desarrollado herramientas de búsqueda, que pretenden hacer posible una recuperación rápida, efectiva y eficiente de la información.

Como es sabido, son muchos millones de documentos los que se encuentran accesibles mediante las herramientas de recuperación de información que existen para facilitar el acceso a ellos, entre los cuales, podemos mencionar los siguientes tipos de herramientas de recuperación de información:

- **Los buscadores, entre los que se encuentran:**
 - **Directorios**
 - **Motores de búsqueda**
 - **Metabuscadore**s

- **Multibuscadores**

- **Agentes inteligentes de búsqueda de información**

- **Buscadores semánticos**

Como se puede apreciar en la lista anterior, existen diferentes herramientas de búsqueda de información, con diferentes características, lo cual ha ocasionado que los usuarios no conozcan qué tipo de herramienta es la que están utilizando y por lo tanto también usan el término de buscador o motor de

búsqueda de manera indistinta para referirse a cualquier herramienta de búsqueda que utilicen.

Por esta razón, es necesario explicar la conceptualización de cada una de estas herramientas que se utilizan en la Web para recuperar información, asimismo se mencionaran algunos ejemplos de cada una de ellas, mencionando las más populares, así como su dirección electrónica,

2.3.1 Los buscadores

Como se mencionó en líneas anteriores, en la década de los 90s los buscadores tuvieron un desarrollo impresionante. Desde los primeros días de la Web, estos se han convertido en una herramienta indispensable para los usuarios de Internet. Por lo tanto los buscadores son los que contienen los índices de la información que se encuentra en la Web, además de ser la llave necesaria para, que los usuarios que buscan información y para los creadores de recursos de información, puedan utilizar estos índices que se encuentran almacenados en cada buscador.

El buscador ha sido definido como un programa informático diseñado para la búsqueda de diferentes formatos de archivos digitales como páginas Web, documentos de texto, archivos de video y música, imágenes. Los cuales se encuentran alojados en un disco duro o servidor local así como en la red de Internet, presentando los resultados de la búsqueda en una serie de enlaces que conectan con los archivos o la información solicitada.

Por lo tanto, en términos generales podemos decir que un buscador es un programa informático, que permite consultar una base de datos en la cual se encuentran los índices que relacionan direcciones de páginas Web con su contenido. Su uso facilita enormemente la obtención de un listado de páginas Web que contienen información sobre algún tema en particular.

Asimismo, algunos autores (Levene 2000) mencionan que existe una gran cantidad de buscadores, clasificados en tres grandes rubros, como ya se mencionó en párrafos anteriores.

- Directorios
- Motores de búsqueda
- Metabuscadore

Estos tres tipos de buscadores, son los más comúnmente utilizados por los usuarios que necesitan y buscan información en la WWW, asimismo cada uno de ellos cuenta con sus características propias, ventajas y desventajas en la recuperación de información.

2.3.1.1 Directorios

Los directorios fueron las primeras herramientas de recuperación de información utilizadas en Internet, su funcionamiento tiene que ver con el trabajo humano, principalmente bibliotecarios, que indexan en una base de datos, una serie de datos relevantes sobre un recurso de información, como son:

- El título
- El resumen del contenido
- La dirección electrónica o URL

Todo esto, dividido en categorías, con el objetivo de que cuando un usuario realice una búsqueda, la base de datos le devolverá una lista de enlaces hacia las páginas Web que tienen relación con las palabras o frases buscadas.

Los directorios, son índices temáticos en los que se pretende clasificar los sitios Web mediante categorías temáticas jerarquizadas. Generalmente son indexados por personal bibliotecario, aunque en la actualidad, existen algunos

directorios como Yahoo, que iniciaron su indización de manera manual, pero con el incremento de información en la red, migraron su forma de indizar la información y actualmente conservan su estructura de directorio pero ahora utilizan los robots de Google y Bing para indizar automáticamente los recursos de información en su base de datos. Los directorios son útiles para los usuarios que acceden a la red sin una clara orientación de cómo buscar información. Hay muchos directorios temáticos, entre ellos podemos mencionar a:

- Yahoo <http://www.yahoo.com/>
- About <http://www.about.com>
- The open Directory Proyect <http://www.dmoz.com./>
- Academic info <http://academicinfo.com/>
- Look Smart <http://search.looksmart.com/>

Dentro de las características que podemos mencionar de estos directorios se encuentran las siguientes:

- Construidos por selección humana. (en algunos casos)
- Organizados por categorías temáticas.
- Alta calidad en contenido de los sitios Web indizados
- Los hay de diferentes tamaños desde pequeños hasta más amplios y especializados, pero siempre más pequeños que los motores de búsqueda.

2.3.1.2 Motores de búsqueda

Estas herramientas de búsqueda, son las más comunes en la Red. Los motores de búsqueda son un tipo de buscador basado en programas informáticos que automatizan el rastreo y la indexación de las páginas Web disponibles, éstos nacieron de la necesidad de organizar la información anárquica contenida en Internet, a diferencia de los directorios los cuales necesitan de una parte humana para la verificación e indexación de las páginas Web. El funcionamiento de los motores de búsqueda se basa en 3 etapas o fases:

- Rastreo
- Indexación y Ranqueo
- Consulta

En la primera fase un programa informático denominado robot o araña rastrea y recoge los datos de todos los recursos de información en el Web, a través de los enlaces compartidos, permitiendo ampliar enormemente el campo de rastreo a un gran número de información.

En el segundo paso un algoritmo toma la información captada por el robot o araña y la valora rankeandola según criterios establecidos por el algoritmo, posteriormente se indexa la información en gigantescas bases de datos.

Por último cuando se realiza una búsqueda el programa consulta a la base de datos mostrando una serie de enlaces o datos numerados y jerarquizados según el ranqueo o relevancia calculada por el algoritmo.

Cuando un motor efectúa una búsqueda, la información es rastreada en el contenido de una base de datos (no se busca directamente en el World Wide Web), por lo que ninguno de estos buscadores incluye en su base de datos

todos los índices de la información existente en el WWW y por lo tanto se obtienen resultados diferentes, entre los más populares podemos mencionar a:

- Altavista <http://www.altavista.com>
- Google <http://www.google.com.mx/>
- Go.com <http://www.go.com/>
- Lycos <http://www.lycos.com/>
- Ask Jeeves, entre otros. <http://www.ask.com>

Dentro de las principales características que podemos mencionar de estos buscadores se encuentran las siguientes:

- Construyen sus bases de datos con la ayuda de programas llamados spiders, robots, etc.
- No están organizados por temas
- No son evaluados, contienen páginas de todo tipo.
- Sus bases de datos son enormes.

Es importante mencionar que las diferencias en los resultados obtenidos en las búsquedas son debido a las características particulares de cada uno de los buscadores.

2.3.1.3 Metabuscaadores

Con la continua aparición de diferentes herramientas y para agilizar la búsqueda y recuperación de información en internet, han aparecido otras herramientas de búsqueda de información en la red, como es el caso de los metabuscadores.

Los metabuscadores son programas informáticos que permiten la realización de una misma búsqueda en diversos buscadores de forma simultánea, su funcionamiento se basa en lanzar las consultas requeridas por los usuarios sobre un conjunto de motores de búsqueda y directorios, aprovechando los algoritmos de búsqueda desarrollados por cada uno de ellos, posteriormente el metabuscador rankea todos los resultados mostrándolos según criterios de relevancia establecidos, su interés radica en la reducción del tiempo de búsqueda por la simultaneidad de herramientas que revisa.

Algunos de los metabuscadores más comunes son:

- Dogpile <http://www.dogpile.com>
- Metacrawler <http://www.metacrawler.com>
- Mama <http://www.mamma.com/>
- Search.com <http://www.search.com/>
- Kartoo <http://www.kartoo.com>
- Ixquick <http://www.ixquick.com>

Asimismo dentro de las principales características de los metabuscadores encontramos las siguientes:

- Eliminan los resultados duplicados
- Eliminan los links rotos, o que no están disponibles
- Utilizan la lógica de recuperación de otros buscadores.
- No cuentan con base de datos propia

Estos metabuscadores cada vez se han ido mejorando, en la actualidad algunos trabajan por medio de clúster (clustering), que permite reducir el volumen de información mediante la categorización o agrupamiento de los datos con características similares, la cual es una tarea natural del intelecto humano.

El clustering aprovecha esta característica aportando herramientas que faciliten la construcción automática de taxonomías y minimizando la intervención humana en este proceso. Este tipo de clasificación automática es conocida también como taxonomía numérica, debido a que fue desarrollada originalmente para ser aplicada en clasificaciones biológicas. Existen alrededor de 150 técnicas diferentes de clustering. En documentación se han utilizado sobre todo técnicas jerárquicas. Algunas versiones de estos metabuscadores son: (Pinto 2005)

- Vivísimo <http://www.vivisimo.com>
- Clusty <http://www.clusty.com>
- Turbo10 <http://www.turbo10.com/>

2.3.2 Multibuscadores

Al existir una gran cantidad de diferentes buscadores en la Web, y para ayudar a los usuarios a consultar en una sola pantalla, sin salir de ella, diferentes directorios, motores de búsqueda y metabuscadores, aparecieron los llamados multibuscadores, ellos se limitan a colocar en una página Web una lista con las herramientas más populares y con una ventana de texto para poder enviar la cadena de búsqueda introducida por el usuario a cada uno de ellas. Al usar un multibuscador se envía la cadena de búsqueda a una serie de motores de búsqueda y directorios previamente definidos.

Una vez realizada la consulta a los buscadores prefijados, el multibuscador devolverá la información de todos los enlaces encontrados. Con la utilización de multibuscadores, el usuario se evita tener que ir de buscador en buscador hasta encontrar la información deseada. El usuario elige los buscadores que quiere utilizar y desde allí realiza su consulta que ofrecerá las respuestas buscador por buscador. La única ventaja es la posibilidad de consultar un gran número de buscadores partiendo de una única página. La exhaustividad privará sobre la precisión, ya que el usuario encontrará un gran número de enlaces y

muchas páginas repetidas. Un buen ejemplo de multibuscador es el ofrecido por:

Cyberjournalist <http://cyberjournalist.ourtoolbar.com/xpi>

Deeperweb <http://deeperweb.com/>

2.3.3 Agentes inteligentes de búsqueda de información

En párrafos anteriores se ha mencionado del gran crecimiento de internet y de la proliferación de buscadores y metabuscadores, como una ayuda para la recuperación de la información, a estos se han sumado otras herramientas de búsqueda denominadas agentes inteligentes de búsqueda de información. Estas herramientas constituyen la evolución lógica de los metabuscadores, debido a que buscan también en varios motores de búsqueda y directorios, solo que para su utilización es necesario la instalación de *softwares* clientes.

Dentro de las principales características de los agentes inteligentes encontramos las siguientes:

- Eliminan los documentos duplicados.
- Permiten el trabajo offline, esto es, sin estar conectados a internet.
- Permiten editar y guardar las estrategias de búsqueda

Entre los agentes inteligentes más conocidos se encuentran:

- Copernic <http://www.copernic.com>
- Bull's eye <http://www.intelliseek.com>
- Matahari o WebFerret <http://www.webferret.com/>
- Bingoo <http://www.bingooo.com>

2.3.4 Buscadores semánticos

Este tipo de buscadores, se pueden considerar como la evolución en la recuperación de información, es decir, son las herramientas del futuro, para llegar no propiamente a la Web semántica, sino a una técnica semántica de recuperar la información en cualquier sistema de información.

Un buscador semántico deberá efectuar la búsqueda atendiendo al significado del grupo de palabras que se solicita en una búsqueda. Los buscadores semánticos interpretarán las búsquedas de los usuarios y los textos de los documentos en la Web a través del uso de algoritmos que representan comprensión o entendimiento, ofreciendo resultados correctos de una manera rápida sin que el usuario tenga que abrir los documentos e inspeccionarlos por sí mismo. De esta manera, los buscadores semánticos reconocen el contexto de las palabras o sentencias de búsqueda.

Con los buscadores semánticos se persigue que la búsqueda que realice el usuario sea más precisa y origine resultados relevantes con la mínima intervención de éste.

Ahora bien, que elementos necesitan este tipo de buscadores, para poder realizar búsquedas semánticas. En primer lugar necesitan utilizar metadatos codificados, es decir, que deben utilizar un esquema de codificación que podría ser el formato MARC o bien un lenguaje de marcado como el XML, los cuales permitirían codificar los metadatos.

En segunda instancia, es necesario que el buscador semántico, cuente con un esquema que le permita dotar a los documentos de un significado de relaciones entre diferentes aspectos de un documento, como podría ser qué relación tiene el tipo de recurso, con el título y el autor, en este sentido el esquema que podría solucionar los aspectos antes mencionados, es el RDF (Marco de Descripción de Recursos)

Y finalmente debe contener también ontologías, las que introducirán un mayor nivel de profundización semántica y proporcionaran una descripción lógica y formal que pueda ser interpretada tanto por las personas, como por las máquinas, asimismo las ontologías permiten la interoperabilidad entre distintos sistemas, como se observó en el capítulo dos.

En la actualidad existe una serie de buscadores llamados semánticos, sin embargo es necesario evaluarlos para saber si reúnen las características que debe de tener un buscador, para realizar búsquedas con una técnica semántica apropiada para recuperar eficientemente la información, entre los que podemos mencionar a:

Kngine	http://www.kngine.com/
Exalead	http://www.exalead.com/search/
DuckDuckgo	https://duckduckgo.com/
Swoogle	http://swoogle.umbc.edu/

La creación y uso de estos buscadores en la Web todavía es incipiente, pero se siguen haciendo esfuerzos para llegar a tener herramientas que ayuden de manera importante a los usuarios de la red a recuperar la información de manera más eficiente.

2.4 Funcionamiento de las herramientas de búsqueda

Una herramienta de búsqueda funciona con base en cuatro elementos básicos:
(Lara y Martínez, 2006:14)

1. Un programa denominado robot, que recorre el WWW buscando recursos de información y sus respectivas URLs.
2. Un sistema automático de análisis de contenidos e indexación de los documentos localizados por el robot.

3. Un sistema de interrogación, generalmente basado en la lógica booleana, que permite al usuario expresar su necesidad de información.
4. Un programa que sirve como puente entre el servidor de documentos html y la base de datos.

La herramienta de búsqueda, recibe la solicitud de los usuarios, la cual está conformada por uno o varios términos, en ese momento la herramienta realiza una consulta en la base de datos que contiene los índices de los recursos Web, finalmente presenta una lista de recursos que contienen los términos solicitados por los usuarios en el título o en el cuerpo del documento. Los resultados aparecen ordenados según algunos criterios que el programa asocia automáticamente a cada recurso.

Levene (2000) menciona que una buena herramienta de búsqueda será capaz de ordenar los resultados de manera que aparezcan antes las páginas más relevantes atendiendo a varios indicadores, entre los cuales se encuentran:

- **Localización**
- **Frecuencia de aparición**
- **Popularidad**
- **Precio**
- **Interfaz**

En este sentido, Levene (2010) menciona que el motor de consultas proporciona la interfaz entre el índice de búsqueda, el usuario, y la Web. De este modo, el motor de consultas procesa una consulta del usuario en dos pasos.

Primer paso, el motor de consulta recupera de la información del índice de búsqueda de páginas Web potencialmente relevantes que coinciden con las palabras clave en la consulta del usuario, y en la segunda etapa se produce un

ranking de los resultados, a partir de las más relevantes a las menos relevantes.

Por esta razón, algunos autores mencionan que la interfaz más utilizada es la basada en páginas Web con formularios, entre ellos podemos mencionar:

- **Formularios**
- **Páginas Web de resultados**

2.5 Tendencia actual de las herramientas de búsqueda

La escasa calidad de la información recuperada es un inconveniente de las actuales herramientas de búsqueda. Los mecanismos para aumentar la precisión en la búsqueda (refinamientos, búsquedas avanzadas, acotación por dominios, etc.), a veces, no funcionan como cabría esperar. A ello, hay que añadir el mínimo valor académico de algunos de los sitios Web recuperados, el porcentaje de recursos repetidos y el porcentaje de recursos inactivos (que ya no existen físicamente en la red aunque continúan indexados).

Debido a estos problemas que se presentan en las herramientas de búsqueda en cuanto a su funcionamiento y las cantidades impresionantes de resultados que ofrecen, algunos autores han vislumbrado algunas vías de solución futura que marcan la tendencia en la evolución de estas herramientas.

Los resultados de la búsqueda pueden ser satisfactorios o no tanto. Las herramientas de búsqueda ofrecen resultados muy diferentes ante una misma cuestión inicial; este hecho demuestra la poca exhaustividad de estas herramientas en la indexación de los recursos Web y pone de manifiesto los problemas derivados de la escasez de control lingüístico.

Actualmente se aboga por la incorporación de metadatos en la organización de la información, así como la incorporación de aspectos de análisis lingüístico y control terminológico en las herramientas de búsqueda, de forma que sea posible efectuar una recuperación menos ligada a la comparación de cadenas de caracteres y más vinculada a la comparación de conceptos.

En este sentido se empieza a hablar de una Web para el público en general y una Web de los recursos culturales, científicos y técnicos. La especialización de las herramientas de búsqueda es una buena vía para conseguir mejores servicios de información, la especialización conduce a la concentración del conocimiento en ciertos lugares donde los usuarios pueden encontrar fácilmente los recursos relacionados con su ámbito de conocimiento.

Por lo tanto, al hablar de una Web semántica, es necesario hablar de herramientas de búsqueda semántica, las cuales efectúan la búsqueda atendiendo al significado del grupo de palabras que se solicita en una búsqueda. Las herramientas semánticas interpretan las búsquedas de los usuarios y los textos de los documentos en la Web a través del uso de algoritmos que representan comprensión o entendimiento, ofreciendo resultados correctos de una manera rápida sin que el usuario tenga que abrir los documentos e inspeccionarlos por sí mismo. De esta manera, estas herramientas semánticas reconocen el contexto correcto para las palabras o sentencias de búsqueda. Otro aspecto importante es el uso de metadatos que permitan describir el esquema en el que se almacenan los datos, organizar la información de tal manera que sea más fácil su recuperación, además es necesario utilizar un modelo básico para establecer propiedades sobre los recursos, como el RDF (Resource Description Framework), así mismo el empleo de Ontologías, permitiendo que las herramientas semánticas puedan recuperar la información más eficientemente.

Después de haber analizado todos los aspectos importantes de los buscadores y metabuscadores, como son sus características, sus componentes y la forma como indexan la información, es necesario que se lleve a cabo una evaluación sobre la forma en la que recuperan la información estas herramientas.

2.6 Incremento de la información

La información, un elemento que utilizamos como parte de nuestra labor profesional, ha sufrido cambios en la forma en que se genera, se distribuye y se obtiene.

El avance tecnológico que se generó a finales del siglo XX y se sigue incrementando en el presente, permite que se desarrollen nuevas formas de crear información, principalmente apoyadas en medios electrónicos, haciendo posible que casi cualquier persona pueda crear y publicar en Internet.

Hoy en día, Internet, está viviendo su propio big bang. Año con año, la numerosa información que cerca de mil 100 millones de usuarios suben a las redes sociales, sitios Web y otros se multiplica considerablemente, marcando un crecimiento nunca antes visto. Es decir que la información del Web crece el 60% al año, del cual los usuarios comunes y corrientes aportan el 75% de lo publicado. (Worldwide Information Growth Ticker, 2010)

Estos y otros datos fueron recopilados del estudio "Extracting Value from Chaos" realizado por IDC, una consultora de mercado que, patrocinada por **EMC** (multinacional especialista en soluciones para la infraestructura de la información) realiza año a año un estudio que mide la cantidad de información que existe en la Web, y en vista a los resultados ofrece un pronóstico de lo que sucederá a futuro. Para explicar todo esto, ya hace mucho que dejaron de hablar de *Gigabytes*: ahora la medida va en *Zettabytes*(es decir, un millón de millones de *Gigabytes*). (Worldwide Information Growth Ticker, 2010)

Durante el 2010, la cantidad de datos superó holgadamente el zettabyte, y ya se espera que para este año la información alcance los 1,8 ZB.. Ya para el 2020, esta información debiera alcanzar 35.2 ZB, producto de múltiples factores que van aportando a este crecimiento, (El Universo Digital: ¿Cuánta información existe en Internet? 2010).

2.6.1 Información académica

Como se ha mencionado anteriormente, existe una gran cantidad de información en Internet, a la cual se tiene acceso por medio de la Web, pero uno de los problemas que se presenta tiene que ver con la búsqueda y recuperación de información, pues encontramos tanta información que en ocasiones es difícil discernir entre lo que sirve o no .

Es así como la localización de la información académica juega un papel fundamental para las actividades de docencia e investigación, pero también esta información crece cada día a pasos agigantados y recuperarla no es tan sencillo, esto debido a que los buscadores convencionales en ocasiones no pueden acceder a estos recursos.

Macías y Michán mencionan que la cantidad de información que se produce ahora sobre temas científicos y académicos es diversa y colosal, asimismo mencionan que puede ser electrónica o impresa, así como en texto, imágenes y sonidos.

Esta información académica o científica, está contenida principalmente en libros, revistas, tesis doctorales y en ocasiones en reportes de investigación, que no necesariamente se hallan publicadas en papel y solamente las encontramos en línea.

Macias y Michán, también, hacen mención, que esta información se localiza principalmente sistematizada en:

- Bases de datos
- Catálogos o listas

Además las podemos encontrar en directorios de repositorios. Su consulta puede libre o restringida, trata sobre los seres vivos o sus partes, fenómenos y explicaciones, esta información versa sobre:

- Las publicaciones
- Los investigadores
- Los proyectos
- Los grupos y las líneas de investigación
- Los convenios
- Los subsidios
- La producción científica
- Las instituciones de investigación y enseñanza
- Las instituciones de enseñanza
- Las sociedades científicas

Esto ha repercutido de manera dramática en la visión del mundo contemporáneo, la práctica científica, las relaciones científicas, sociales, económicas, políticas y culturales. (Macías y Michán, 2009)

En 1998 se comenzó a publicar en formato electrónico vía Internet la información científica que se encontraba en papel, existiendo para 2001 más de 4000 artículos en línea, actualmente existe una proliferación de literatura científica virtual, diferenciándose revistas de acceso libre y revistas de acceso limitado.

Revistas de acceso abierto

El libre acceso a la producción científica u open access se refiere al acceso a través de Internet a los trabajos derivados de la actividad científica y académica libre de barreras tanto económicas como de las derivadas de los derechos de copyright. Las vías para alcanzar el open access según la Budapest Open Initiative, son la “ruta verde”(green route) mediante el depósito de las publicaciones u otros objetos digitales de todo tipo en repositorios, o la “vía dorada” (golden route) referida a la publicación en revistas open access.

(Budapest Open Access Initiative, 2014)

Herramientas para recuperar artículos de revistas de acceso abierto

En la Web existen muchos sitios considerados de libre acceso u open access, como ejemplo se pueden considerar los siguientes sitios, por las grandes cantidades de información con que cuentan:

- DOAJ, por sus siglas en inglés, es el Directorio de Revistas de Acceso Abierto es un servicio que proporciona acceso a revistas de calidad controlada y acceso abierto. El Directorio pretende ser global y abarcar todas las revistas de acceso abierto científicas y académicas que utilizan un sistema de control de calidad adecuado, y no se limita a idiomas o áreas temáticas en particular. El objetivo del Directorio es aumentar la visibilidad y la facilidad de uso de las revistas de acceso abierto científicas y académicas, promoviendo así su mayor uso e impacto. Actualmente este directorio cuenta con 9937 revistas de 120 países y cuenta con 1,157,054 artículos en texto completo. (Directory of Open Access Journals, 2014)
- e-revistas, Plataforma open access de revistas científicas electrónicas españolas y latinoamericanas, que cuenta con aproximadamente 708 revistas y un promedio de 221823 artículos en texto completo, entre los países latinoamericanos que ofrecen el mayor número de revistas electrónicas se encuentran: Colombia, Brasil, Venezuela y México.(e-revist@s, 2014)

Así como estos sitios podemos encontrar una gran cantidad de ellos en la red y poder utilizarlos sin ninguna restricción o costo.

Revistas de acceso limitado

En el caso de las revistas que cuentan con un acceso limitado, es muy complicado utilizarlas, esto debido a que por su naturaleza pueden encontrarse en sitios que tienen ciertas restricciones de uso como pueden ser, Instituciones que solo permiten el acceso a los miembros de su comunidad o que pertenecen a organizaciones que ofrecen servicios de información y que tienen un costo.

En este sentido se suele considerar el comienzo de la literatura científica basada en revistas en el año 1665, con la publicación del **Journal des savants**. Asimismo, los ingleses, insisten en que sus **Philosophical Transactions of the Royal Society** son al menos tan antiguos.

A partir de mediados del XVII aparecen las revistas y su crecimiento es exponencial desde el primer momento, han pasado los siglos y a partir de la segunda guerra mundial (gracias al éxito del modelo que supuso el **proyecto Manhattan**) la ciencia se ha profesionalizado. Ahora ya no eran solo unos pocos caballeros ingleses o franceses los que dedicaban su tiempo al conocimiento científico, sino miles de personas en muchísimos países; y financiados en algunos casos por ellos mismos.

En el año 2010, M'hamed el Aisati, jefe de nuevas tecnología de Elsevier, realizó un estudio sobre la explosión de la investigación científica basada principalmente en publicaciones únicas contenidas en la base de datos Scopus de Elsevier, en esta investigación se presenta que alrededor de 39 millones de artículos fueron publicados en revistas científicas entre 1817 y 2010.

Asimismo presento la infografía denominada La genealogía de la ciencia según Scopus (Genealogy of Science According to Scopus) en este mapa se presenta una línea del tiempo de las publicaciones de casi 30 000 revistas

desde su fecha de creación. La información está, además, ordenada en 27 categorías y en la gráfica se pueden observar las 15 categorías más activas.

Entre ellas se encuentran:

- Medicina
- Física y astronomía
- Ingeniería
- Bioquímica, Genética y Biología molecular
- Química
- Ciencias de la tierra
- Computación
- Ciencia de los materiales
- Ciencias sociales
- Agricultura y ciencias biológicas
- Medio ambiente
- Matemáticas
- Ingeniería química
- Neurociencia
- Energía

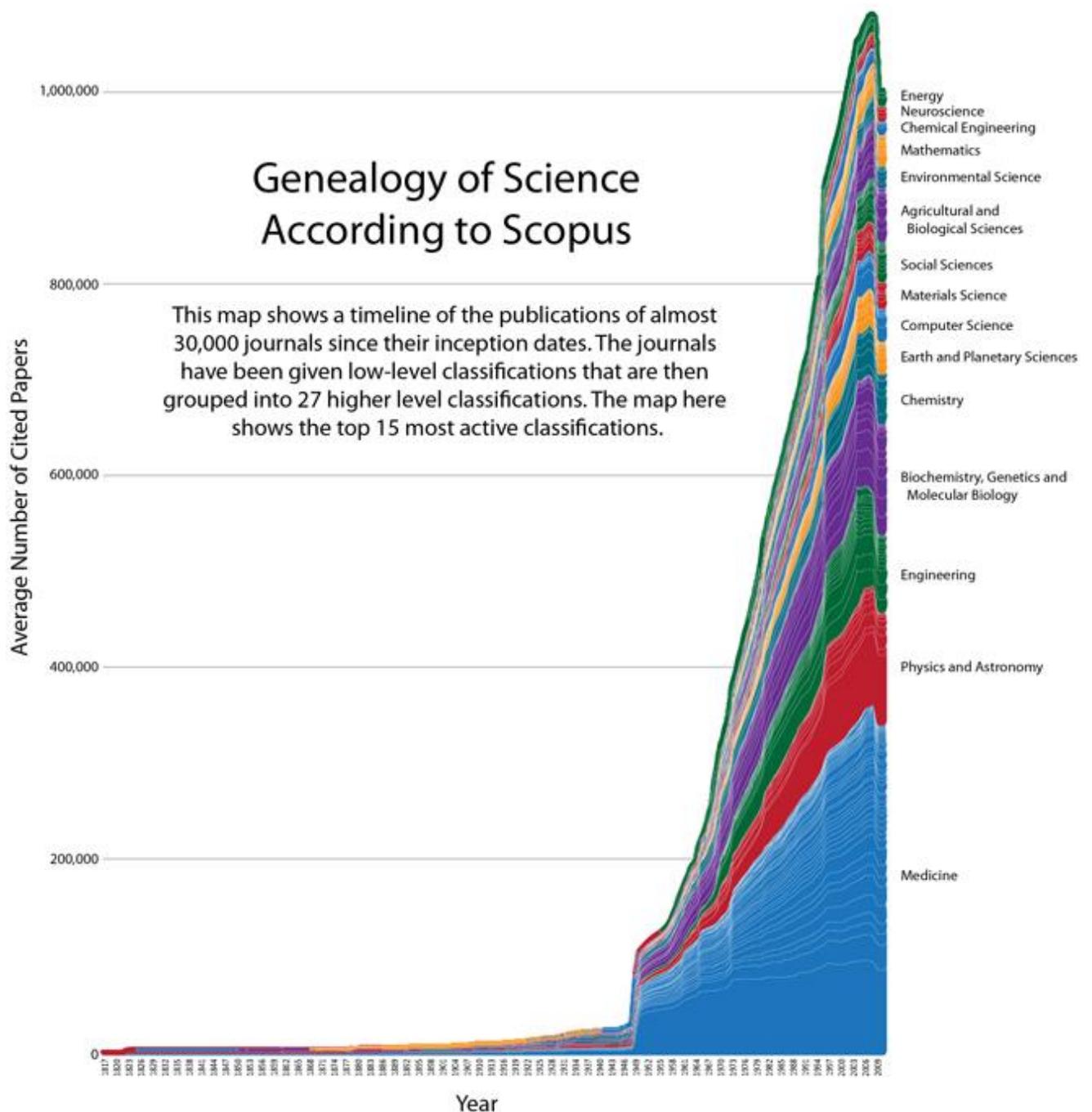


Imagen # 2.1 Genealogy of Science According to Scopus

Como se puede apreciar en la imagen # 1, el crecimiento de información científica es impresionante debido a que en todos los países del mundo se realiza investigación de casi todas las áreas del conocimiento estas investigaciones se publican en revistas académicas y científicas, así como en sitios especializados, en este sentido, El National Center of Biotechnology Information de los Estados Unidos reportaba que para el año 2010, su base de datos PubMed-Medline (PubMed, 2014) contaba con más de 19 millones de

referencias bibliográficas, estimando un incremento de 800 mil referencias cada año.

En Internet y principalmente en la Web existe una gran cantidad de sitios que ofrecen servicios de recuperación de información, para poder ingresar a ellos es necesario tener una suscripción, es decir que es necesario pagar un costo por utilizar sus servicios. También es importante aclarar que algunos sitios de este tipo permiten que se utilice sus servicios sin costo, pero algunos documentos son propiedad de otros servicios y si uno necesita un artículo en particular y este se encuentra en otra compañía no se puede visualizar el documento completo y es necesario pagar un costo por obtenerlo.

A continuación se presentan dos sitios representativos, estos son:

SCIRUS, en su portal se puede observar que cuenta con más de 575 millones de artículos científicos indexados en el último recuento. (<http://www.scirus.com/>)

Otro servicio de información es Sciencedirect que reporta un promedio de 11,803,683 artículos y capítulos de libros de más de 2500 revistas científicas y más de 11 000 libros. (ScienceDirect, 2014)

Libros electrónicos

Otro rubro importante, hablando del incremento de información, son los denominados libros electrónicos, estos se han integrado de una manera importante en la red, teniendo una gran aceptación por parte de los usuarios que necesitan información relevante para sus actividades académicas y de investigación.

En el portal de Data-Planet by Conquest System. Inc (2013) aparece un gráfico que presenta las tendencias, sobre el incremento de libros electrónicos en bibliotecas académicas entre los años 2002 a 2010, estos datos provienen del

National Center for Education Statistics (NCES), que realiza una compilación de datos de unos 3,700 Centros de Instituciones de Educación Superior de los Estados Unidos, para tener una visión general del comportamiento de las Bibliotecas Universitarias de ese país.

Este gráfico proporciona diferentes opciones para ver detalles sobre los servicios, personal, colecciones y en especial información de libros electrónicos que se encuentran en las bibliotecas universitarias de los Estados Unidos.

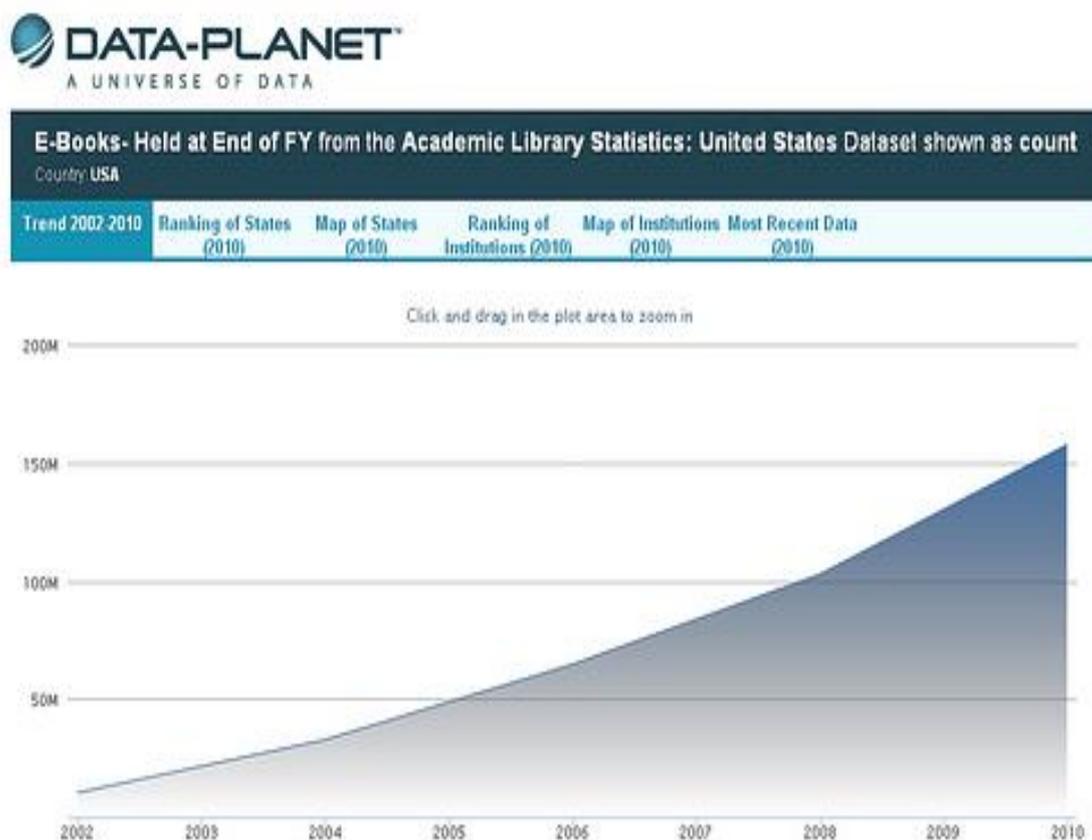


Imagen # 2.2 tomada de -Planet by Conquest System. Inc (2013)
http://farm9.staticflickr.com/8383/8641390823_1662c628f6_d.jpg

En la siguiente gráfica se presenta el incremento de los libros electrónicos en las bibliotecas universitarias de los Estados Unidos, este incremento es en cuanto a unidades, es decir a libros electrónicos por unidad no por título, está información que se presenta es de cada dos años, debido a que la compilación de los datos por el NCES se realiza con esa frecuencia.

AÑOS	UNIDADES DE e-BOOKS
2002	10 318 157
2004	32 775 291
2006	64 681 892
2008	102 768 275
2010	157 673 888

Datos tomados de Data-Planet by Conquest System. Inc (2013)

<http://www.data-planet.com>

Como se puede apreciar en la tabla anterior el incremento de unidades de libros electrónicos solamente en las bibliotecas universitarias de los Estados Unidos se ha dado de manera impresionante. En el año 2002 había un promedio de 10, 318 157 libros electrónicos, entre los años 2002 a 2004 se triplicó el número de libros electrónicos con un total de 32, 775 291 libros disponibles, para el año 2006 se duplicó la cantidad de libros con un aproximado de 64, 681 892, para el año 2008 se volvió a duplicar nuevamente la cantidad de libros con aproximadamente 102, 768 275, y finalmente se presenta un incremento del 50 % de libros electrónicos en 2010 con un total de 157.673.888.

Existe un incremento impresionante, de las colecciones de libros electrónicos en bibliotecas estadounidenses entre 2002 y 2010. Una muestra evidente del incremento de información que se sigue dando en el siglo XXI en la llamada red de redes.

Los Repositorios

Sin duda alguna hablar del incremento de la información académica en la Web, no tendría tanta validez, si dejamos fuera a los repositorios de información, tanto institucionales, como a los especializados, estos archivos digitales que pretenden recoger toda la producción científica de un país, una institución o una región en un sitio Web que esté a disposición de los usuarios en forma gratuita y sin restricciones. (Córdoba González, 2010)

Nacen a partir de la Iniciativa de Acceso Abierto de Budapest, en 2001, y su propósito fundamental fue acelerar el progreso del esfuerzo internacional para lograr que los artículos científicos en todos los campos del conocimiento estuvieran disponibles sin costo algunos para los usuarios vía Internet.

En la gráfica # 3, que se encuentra en el Directorio de Repositorios de Acceso Abierto (DOAR) por sus siglas en inglés, se presenta la siguiente información, acerca de la distribución de los repositorios por continente.

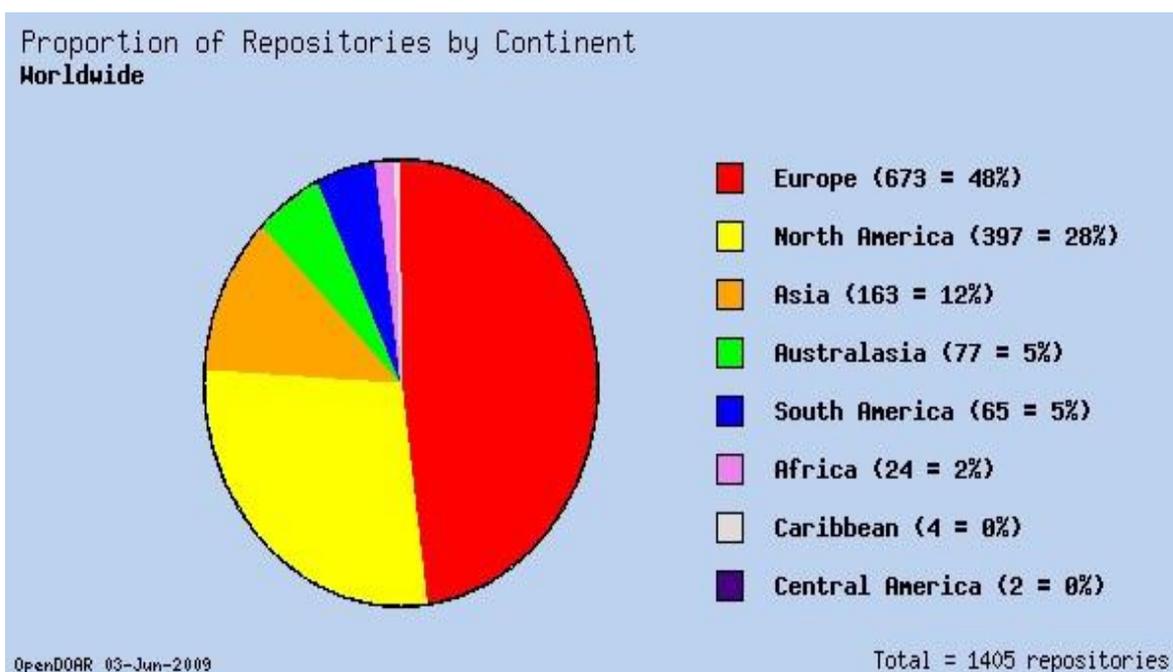


Imagen #2.3 Open DOAR: <http://www.opendoar.org/>

PROPORCIÓN DE REPOSITORIOS POR CONTINENTE

CONTINENTE	No. DE REPOSITORIOS	PORCENTAJE
EUROPA	673	48%
NORTE AMERICA	397	28%
ASIA	163	12%
AUSTRALIA	77	5%
AMERICA DEL SUR	65	5%
AFRICA	24	2%
CARIBE	4	.8%
CENTRO AMERICA	2	.8%
TOTAL	1405	100 %

Como se puede observar en la gráfica # 3, en Europa se encuentran el 48% de los repositorios con un total de 673, en Norte América existe el 28% con 397 repositorios, en Asia se tienen 163 repositorios que conforman el 12%, en Australia existen aproximadamente 77 repositorios que suma el 5%, América del sur aporta también un promedio del 5% con 65%, África, el caribe y Centro América, con 24 repertorios, 2%, 4 repertorios, .8% y 2 repertorios, equivalente al .8% respectivamente, con un total de 1405 repositorios, tanto institucionales como especializados, por la cantidad que se presenta al final, pareciera que no es muy grande la cantidad de información que aportan estos recursos de información a Internet, pero lo interesante es ver cuanta información de artículos y reportes de investigación se encuentran en cada uno de ellos.

En los repositorios se encuentra información completamente académica, dentro de los recursos que la contienen se encuentran:

- Artículos
- Libros
- Conferencias
- Tesis
- Conjunto de datos
- Objetos de aprendizaje
- Multimedia
- Patentes
- Referencias
- Software
- Especial
- Documentos Inéditos

En definitiva, se puede decir que así como ha crecido la información general en la Web, también existe un crecimiento exponencial de información académica y científica, la cual es generada principalmente por académicos y científicos que pertenecen a instituciones de enseñanza, instituciones de investigación y sociedades científicas.

Uno de los problemas para poder consultar esta información es que se necesita conocer las direcciones electrónicas de los sitios en los que se encuentran indizados y es muy difícil que se pueda recuperar la información mediante los buscadores convencionales de la Web, como se mencionó anteriormente.

A esta información que no puede ser recuperada en primera instancia por los buscadores convencionales, muchos autores la han denominado con diferentes nombres, entre ellos se encuentran: Internet invisible, Web profunda, Web invisible, etc.,

2.6.2 Internet Invisible

El término Web invisible fue utilizado por primera vez por Jill Ellsworth para denominar a la información que resultaba invisible para los buscadores convencionales en la Web. (Bright Planet 2000)

Para Isidro Aguillo, la Web invisible es “ una fracción muy voluminosa de información que, aunque accesible a través de la red, por distintas razones no es indexada por los motores de búsqueda” (Aguillo, 2000).

Bergman (2000) señala que el hecho de buscar información en la Web es como pescar en el océano con una red, en donde se captura mucha información de la superficie, pero los contenidos de las profundidades del océano, que es donde se encuentra frecuentemente la información valiosa y de índole académica, no se recuperan tan fácilmente o difícilmente son arrastradas por esa red.

El mismo autor también la ha denominado Web profunda (Deep Web) para diferenciarla de la Web superficial (surface Web), cuya información puede recuperarse por medio los buscadores de Internet (Bergman, 2000).

Se denomina “Internet Invisible” o Infranet (en inglés Deep Web) al conjunto de recursos accesibles únicamente a través de algún tipo de pasarela o formulario Web, que no pueden ser recuperados por los motores o robots de los buscadores convencionales.

Es decir, que los recursos de información que debemos considerar, como Internet Invisible, son: Artículos, libros, conferencias, tesis, objetos de aprendizaje, documentos multimedia, patentes, referencias bibliográficas, softwares, documentos especiales y documentos inéditos o que no han sido publicados, contenidos en diferentes herramientas de recuperación.

Asimismo las organizaciones académicas, también participan en el crecimiento de Internet Invisible aportando una gran cantidad de información que se ve reflejada principalmente en:

- Bases de datos
- Páginas dinámicas
- Intranets de las organizaciones.
- Repositorios, etc.

Tamaño de Internet Invisible

Un estudio realizado en el año 2000 por Michael K. Bergman arrojó una serie de datos que nos pueden dar una idea del tamaño de Internet Invisible. En este estudio se menciona que la información pública contenida en el Internet invisible es 400 a 550 veces más grande que el segmento común de la Web, y crece a mucha mayor velocidad (Bergman, 2000)

Otros estudios calculan que el tamaño de la Web profunda es 275 veces mayor que la Web visible. Otras estimaciones posteriores señalan que el tamaño de la Web invisible es sólo entre dos y 50 veces mayor que el de la Web visible (Sherman y Price, 2001). Las diferencias en las cifras se deben a las diferentes metodologías utilizadas por los autores.

El internet invisible contiene 7.500 Terabytes de información, comparados con los 19 Terabytes de información en la Web superficial. La cifra de la Web invisible equivale aproximadamente a 550 000 millones de documentos

individuales, comparados con los 8 000 millones de la Web visible que reporta Google.

En el mismo estudio de Bergman se menciona que existen más de 200.000 sitios conocidos como internet invisible. Como un ejemplo se menciona que 70 de los sitios más grandes de la Web invisible contienen, en conjunto, cerca de 750 terabytes, lo que significa 40 veces más que la contenida en la Web visible. Esto se debe a que los sitios de la Web invisible tienden a ser más densos, en cuanto información, con un contenido más profundo que los sitios indizados en la Web convencional.

Es decir, que un sitio que es considerado como de Internet Invisible, contiene más información que un sitio común, esto se debe a que un sitio de Internet Invisible puede ser una base de datos o un catálogo de biblioteca y este contiene más información, esto debido a que cuando se da de alta una nueva bases de datos puede contener en su interior una gran cantidad de información que aumenta considerablemente el tamaño de la información de la Web invisible.

Esto permite entender por qué algunos autores mencionan que la cantidad del contenido total de la Web invisible es de 1000 a 2000 veces mayor que la de la Web superficial. Con respecto al contenido de la Web invisible Isidro F. Aguillo agrupa los siguientes recursos:

- Catálogos de bibliotecas y bases de datos bibliográficas.
- Bases de datos no bibliográficas.
- Revistas electrónicas.
- Depósitos de documentos en formatos no indexables
- Obras de referencia: diccionarios, enciclopedias, etc.

Como se puede apreciar en las líneas anteriores la información que existe en Internet se puede dividir en información que no cuenta con un soporte académico y la información que está respaldada por instituciones que cuentan con un respaldo académico, como son:

- Las Instituciones educativas, entre ellas colegios, universidades, etc
- Las Instituciones gubernamentales, secretarías, direcciones, consejos etc.
- Organizaciones privadas, laboratorios, fábricas etc.
- Organizaciones no gubernamentales, paraestatales, etc.

Estas organizaciones generan una gran cantidad de información con tintes académicos que día con día engrosan la cantidad de documentos que contiene la Internet.

Como se ha dicho en párrafos anteriores, algunos autores mencionan que entre el contenido del Internet invisible se encuentran los catálogos de bibliotecas y las bases de datos bibliográficas y no bibliográficas, en este sentido se han realizado algunos estudios en los cuales se han revisado algunos sitios que son considerados como directorios de Internet invisible y que contienen una gran cantidad de bases de datos y sitios especializados que contienen información netamente académica y que están auspiciados por Instituciones gubernamentales, universidades, etc.

En el estudio denominado *Accessing the Deep Web: A Survey* elaborado por Bin He y otros, (2004) revisaron cuatro sitios que contienen una gran cantidad de bases de datos Web de diferentes áreas del conocimiento. La encuesta utilizó el método de muestreo IP para recoger muestras de servidores al azar para estimar la escala global, así como facilitar su posterior análisis.

Durante abril de 2004, se recogieron muestras aleatorias de 1,000,000 de IPs (de un universo de 2.230,124,544 direcciones IP) (Bin He, 2004). Estas direcciones fueron revisadas para saber si pertenecían a servidores que contenían bases de datos con información académica y que son consideradas como Internet Invisible, como resultado de estas revisiones se encontró lo siguiente:

	Resultados de muestreo	Estimación total	99% Intervalo de confianza
Sitios Web profunda	126	307 000	236 000 - 377 000
Bases de datos Web	190	450 000	366 000 - 535 000
Bases de datos estructuradas	147	348000	275 000- 423 000
Bases de datos no estructuradas	43	102 000	62 000- 142 000
Interfaces de consulta	40	1258000	1097000 - 1419000

Como se puede observar en esta grafica se revisaron los sitios muestra y se realizó un estimado total considerando un intervalo de confianza del 99%, en ese sentido se pueden considerar que en ese momento existían 2, 465, 000 sitios considerados de Internet invisible. Asimismo se encontraron que los sitios estaban divididos en 14 categorías, entre las que se encuentran:

- Negocios y Economía
- Computadoras e Internet
- News & Media
- Entretenimiento
- Deportes y Recreación
- Salud
- Gobierno
- Regional
- Sociedad y Cultura
- Educación
- Artes y Humanidades
- Ciencias
- Referencia
- Otros

Como se muestra en el listado anterior los sitios que se revisaron, prácticamente cubren la mayoría de áreas del conocimiento

Los sitios que se revisaron fueron:

- Completeplanet.com, que contiene más de 70,000 bases de datos

- lii.org (Librarian Indexing to the Internet) cuenta con 14,000 bases de datos
- Turbo10.com com, contiene aproximadamente 2,300 bases de datos, e
- Invisible-web.net con 1,000 bases de datos

Esto puede dar una idea de las grandes cantidades de información que se encuentran en lo que se denomina Internet Invisible, debido a que estos sitios contienen bases de datos y no artículos, si se recuerda en párrafos anteriores en los que se mencionaban tres bases de datos que contienen artículos y capítulos de libros, se entenderá mejor a que nos referimos, estos sitios eran; PubMed-Medline (www.pubmed.com) con más de 19 millones de referencias bibliográficas, SCIRUS (<http://www.scirus.com/>) con más de 575 millones de artículos científicos y fialmente Sciencedirect .(<http://www.sciencedirect.com/>) reporta 11,803,683 artículos y capítulos de libros. Esto únicamente en tres bases de datos, imaginemos la cantidad de información ya sea artículos o capítulos de libros, que existirá en 70 000 bases de datos. Ahora bien es importante mencionar que para efectos de este trabajo solo se recuperaron sitios que contienen información netamente académica pero también es necesario mencionar que dentro de Internet invisible se encuentran sitios que no son académicos y que es necesario tomarlos con cuidado, sobre todo para las audiencias de internautas jóvenes que podrían perturbarlos.

2.6.3 Los usuarios

Sin duda alguna el crecimiento de la información en Internet para algunos autores es una preocupación constante, debido a que entre más información sin elementos para poder organizarla, hace que en lugar de beneficiarnos nos perjudique, esto claro, porque cuando se necesita información y se utilizan los motores de búsqueda de la red, se encuentra mucha información, haciendo cada vez más difícil encontrar información de calidad. Para conocer cuanta información se genera cada año en Internet se han realizado varios estudios en

los cuales se puede apreciar la cantidad de usuarios de Internet, que son de alguna manera los que generan dicho crecimiento.

Un estudio realizado por Neilson Online* en 2010 mencionaba que existían más de 1,733,993,741 usuarios en ese momento de Internet.

* <http://www.nielsen-online.com/>

En septiembre de 2009 había un aproximado de 1,730,000,000 de usuarios los cuales se habían incrementados el 18% con respecto al año anterior, estos usuarios estaban distribuidos de la siguiente manera:

USUARIOS	DISTRIBUCION GEOGRAFICA
738,257,230	ASIA
418,029,796	EUROPA
252,908,000	AMERICA DEL NORTE
179,031,479	AMERICA LATINA Y EL CARIBE
67,371,700	AFRICA
57,425,046	ORIENTE MEDIO
20,970,490	OCEANIA / AUSTRALIA

En México, la Asociación Mexicana de Internet (AMIPCI) presentó su noveno estudio sobre los hábitos de los internautas en México. En este estudio se actualizaron los principales indicadores que se han presentado en los estudios anteriores, tales como:

- El universo de internautas en México, es decir la cantidad de usuarios que se conectan a Internet, desde el 2006 hasta la fecha.
- El perfil del internauta mexicano. Usuarios de Internet por género y por grupos de edad.
- Datos de conexión del internauta.
- Principales actividades del internauta mexicano en la red.

En el caso de los usuarios que se conectan a Internet, encontramos lo siguiente:

AÑOS	USUARIOS DE INTERNET EN MÉXICO (CIFRAS EN MILLONES)
2006	20.2 millones
2007	23.9 millones
2008	27.6 millones
2009	30.6 millones
2010	34.9 millones
2011	40.6 millones
2012	45.1 millones

Como se puede observar en la tabla # las cantidades de usuarios que se presentan son impresionantes, debido que del año 2011 al 2012 se incrementó el 10% de usuarios conectados a la red, es decir que en un año se agregaron 4 y medio millones de nuevos usuarios a la red.

Con respecto al perfil del usuario mexicano, se encuentran dos indicadores, el primero tiene que ver con el género, en este sentido se presenta que los usuarios que más utilizan el Internet son hombres con un 51% y con el 49% son mujeres, esto resulta interesante porque nos podemos dar cuenta que el sexo femenino ha tenido un repunte impresionante en el uso de tecnologías de la información casi igualando al sexo masculino.

El segundo indicador tiene que ver con la distribución de usuarios de Internet por grupos de edades, los datos que se presentan son los siguientes:

EDAD AÑOS	PORCENTAJE %
6 – 11 años	11 %
12 – 17 años	22 %
18 – 24 años	21 %
25 – 34 años	19 %
35 – 44 años	12 %
45 – 54 años	10 %
Más de 55 años	4 %

En esta grafica encontramos datos interesantes, puesto que la mayoría del porcentaje se encuentra entre las edades de 18 a 54 años con un 62 %, en

este sentido son interesantes estos datos ya que si tomamos en cuenta que en México la mayoría de los alumnos que entran a la Universidad cuentan con 18 años en adelante, por lo tanto son usuarios que utilizan la red y también utilizan información para realizar sus actividades académicas.

Con respecto a los datos de conexión del internauta mexicano en 2013, encontramos que los usuarios se conectan aproximadamente 5 horas y un minuto, 67 minutos más que en 2012 y el día que más porcentaje de conexión hay entre ellos, es el viernes.

Finalmente en el rubro de principales actividades de los usuarios mexicanos en la red, encontramos aspectos relevantes, como son:

ACTIVIDADES EN LA RED	PORCENTAJES
Enviar/recibir mails	87 %
Búsqueda de información	84 %
Acceder a Redes Sociales	82 %
Enviar/recibir mensajes instantáneos	61 %
Comprar en línea	37 %
Operaciones de Banca en línea	33 %
Búsqueda de empleo	20 %
Acceder/crear/mantener sitios propios	17 %
Acceder/crear/mantener blogs	16 %
Visita a sitios para adultos	15 %
Otras actividades	11 %
Encuentros online (búsqueda de pareja)	8 %

Definitivamente algo que resulta completamente increíble es el porcentaje de usuarios mexicanos que utilizan el Internet para buscar información en la Web, con un 84% solamente abajo del 87% de usuarios que utilizan la red para enviar y recibir correos o e-mails, es decir que en México se utiliza más la red para las comunicaciones y después para buscar información, en este sentido los bibliotecarios debemos estar contentos pues como ya se ha mencionado el Internet abre nuevas posibilidades para los bibliotecarios, en los aspectos de la organización y recuperación de la información.

2.6.4 Problemática para acceder a la información

El incremento incontrolable de la información que se encuentra en la red, ha traído como consecuencia que los usuarios se enfrenten a una red con muchos problemas y carente de organización para poder recuperar la información que sea relevante para sus necesidades.

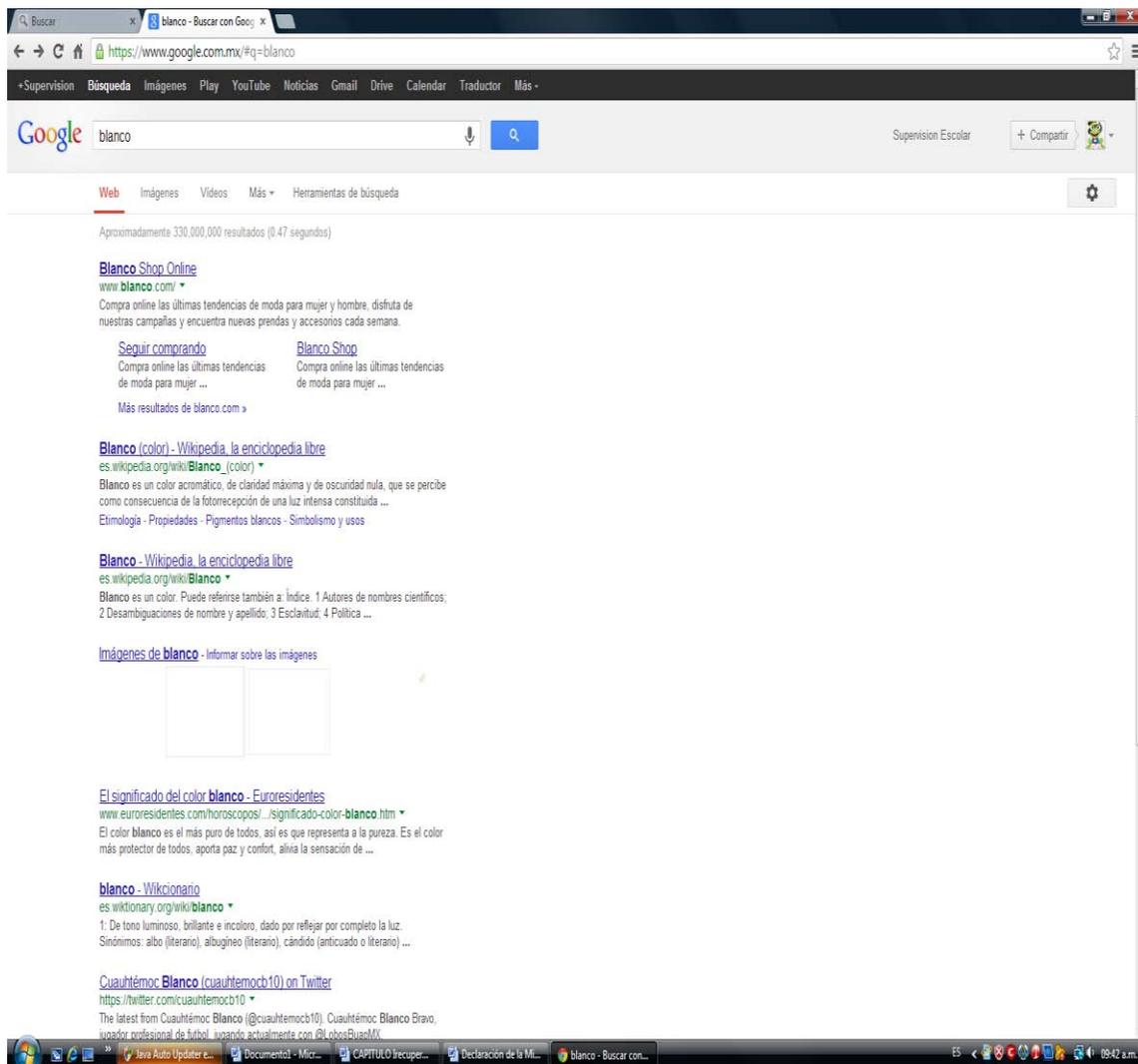
Asimismo, la recuperación de información solía ser una actividad que sólo unas pocas personas se dedicaban a realizarla, entre ellas los bibliotecarios de referencia, asistentes legales y algunos otros profesionistas. Ahora el mundo ha cambiado, y cientos de millones de personas participan en la recuperación de información todos los días cuando se utiliza un motor de búsqueda para rastrear la información en la Web. En este sentido podemos mencionar que en la recuperación de información en sistemas automatizados y en especial en la Internet, existen muchos problemas que impiden que los usuarios puedan localizar fácilmente la información relevante y pertinente para cubrir sus necesidades de información.

Al buscar información en la Red, los usuarios se encuentran con dos problemas muy graves, el primero es que al realizar una búsqueda en Internet encuentran mucha información y en ocasiones es muy difícil separar las páginas que cuentan con un soporte académico de las que no lo tienen.

En el ámbito de la recuperación de información los bibliotecarios trabajan con dos aspectos importantes, que son: la relevancia y la precisión. Es decir que si al realizar una búsqueda de información en algún buscador de la red se encuentra poca información relevante, entonces la tasa de relevancia es baja, y si por el contrario se encuentra mucha información irrelevante es decir mucha basura, entonces la precisión es nula. Esto solamente lo pueden detectar los profesionales de la información, debido a que su formación les permite identificar cuando se esta realizando una buena estrategia de recuperación de información, en el caso de los usuarios comunes es difícil que detecten esta situación debido a que ellos solamente revisan los primeros documentos que

aparecen en la lista de búsqueda del el buscador y no les interesa saber si hay más documentos que les puedan servir.

Un ejemplo de este problema podría ser que al realizar una búsqueda de información en Google, utilizáramos la palabra “Blanco” tratando de encontrar un autor que lleve ese apellido, y los resultados serían estos:



En la imagen anterior se puede ver que existen 360 000 000 de resultados de esta búsqueda, en los cuales se hace referencia en primera instancia a definiciones sobre el color blanco, sitios de futbolistas o de grupos de personas

que se dedican a otras actividades que no tienen nada que ver con la necesidad de información que tenemos.

El segundo problema es que en ocasiones no encuentran los documentos que buscan, esto ocasionado a que mucha información en la red se encuentra perdida, debido a que los recursos de información son almacenados en la Web por sus propios creadores sin ningún tipo de descripción que permita su organización y posterior recuperación, y esto posiblemente sea uno de los puntos débiles de la red, porque millones de documentos no pueden ser recuperados por no contar con una adecuada descripción de sus características que le permita a los buscadores su localización y recuperación.

Existen varios problemas cuando se trata de recuperar la información. Por un lado los servicios de búsqueda (motores) recuperan una gran cantidad de documentos que, en la mayoría de los casos, no satisfacen las necesidades de información de los usuarios por no ser pertinentes para su investigación, y esto es debido a que los documentos en la Web carecen de datos suficientes para su indización. (Senso, 2003)

Desde hace varios años los bibliotecarios tienen que afrontar el reto de solucionar los problemas de la recuperación de información ocasionados por el aumento vertiginoso de la información en formato electrónico y la aparición de bibliotecas digitales, además como resultado de este crecimiento, cada vez es más difícil el proceso de localizar la información. (Senso, 2003)

Por esta razón, resulta evidente la necesidad de establecer mecanismos que permitan una indización y descripción más exhaustiva de los recursos electrónicos, para solucionar el problema de recuperación. En la actualidad se cuenta con sistemas de metadatos que pueden ser las herramientas que permitan la realización de estas tareas. (Senso, 2003)

Entre las causas más comunes a las que se enfrenta el usuario, al tratar de recuperar documentos electrónicos en la red, están las siguientes:

- La falta de datos suficientes para su indización.
- Inadecuado uso de estructuras de metadatos.
- El uso de protocolos inadecuados.

2.6.4.1 Falta de datos suficientes para su indización

Para profundizar acerca de estos problemas es necesario conocer un poco más a detalle cómo se indexa la información principalmente en las bases de datos convencionales y las de los buscadores y algunas otras herramientas de la Web.

La indexación es el proceso por el que un buscador va rastreando los sitios que se encuentran en la Web y va incorporando a su base de datos el contenido de esas URLs. La indexación es la única forma como los sitios Web, pueden aparecer en los buscadores, toda página que el robot no haya rastreado ni indexado no figurará en los resultados.

Los buscadores poseen un robot que es una especie de programa que se dedica a rastrear las páginas Web e incorporar lo que cree interesante a la base de datos de su buscador, tanto para indexar nuevas páginas Web como para actualizar los contenidos de su base de datos, es importante mencionar que los buscadores no indexan todo lo que recorren, sólo indexan lo que creen oportuno.

Los buscadores poseen una importante capacidad para decidir la importancia de los temas que componen un documento, lo que podríamos asimilar a la indización tradicional, y tecnológicamente es factible obtener la materia de los mismos documentos mediante técnicas de clasificación automática. Sin

embargo existen muchas otras formas de ofrecer información sobre un documento (lo que en el ámbito de la Bibliotecología entenderíamos por catalogación) que ningún sistema parece hoy día ser capaz de ofrecer. (Del Valle Gaztamiza, 2005)

Eva Méndez menciona que a diferencia de los indizadores humanos, estas aplicaciones, en general, no identifican características de un documento como la materia de la que trata, el autor, la fecha de publicación, el tipo de documento o las condiciones de acceso (no pueden distinguir, por ejemplo, si un documento es un poema o un informe científico). (Eva Méndez, 2002)

Esta “información catalográfica” se podría ofrecer como parámetros de búsqueda decisivos a la hora de recuperar un documento. Esta información requiere de su inserción manual por parte de los autores, ya que el proceso en conjunto no parece ser automatizable hasta el día de hoy.

Los buscadores de Internet, como los sistemas tradicionales de recuperación de información, representan las colecciones de documentos por medio de índices inversos. Estos índices los construyen con las palabras extraídas de cada documento (página Web, documento pdf...etc) por un sistema de indización automática que aplica a cada sitio unos determinados algoritmos:

- Analiza la frecuencia de aparición de las palabras
- Elimina palabras vacías
- Agrupa palabras con la misma raíz
- Analiza y valora el lugar de aparición de cada palabra
 - Título
 - Resumen
 - Metadatos

Finalmente, propone unos términos ponderados, es decir que tienen, cada uno de ellos, unos valores positivos.

Estas ponderaciones determinarán en parte el orden de recuperación de los documentos, ya que la idea final es la de proporcionar al usuario un listado ordenado de recursos pertinentes, un *ranking*. También tendrán en cuenta la importancia relacional del documento y los documentos más citados, o más enlazados, serán considerados más importantes.

2.6.4.2 Inadecuado uso de estructuras de metadatos

Otro problema, al que se enfrentan los usuarios al querer acceder a la información que necesitan para resolver sus tareas o actividades académicas, es que las personas que integran sitios nuevos en la red utilizan estructuras inadecuadas de metadatos y esto hace difícil su recuperación.

Esto se demuestra en el estudio que se realizó en la Universitat Pompeu Fabra en España, por el grupo DigiDoc, en 61 revistas digitales de libre acceso del ámbito de la información y la documentación. Este estudio se realizó en los meses de diciembre de 2005 y enero de 2006. (Cristòfol Rovira y Mari-Carmen Marcos, 2006)

Al tratarse de revistas que tienen que ver con nuestra área de trabajo, el grupo planteo la siguiente hipótesis: “si los profesionales de la documentación son quienes más interés demuestran en que las páginas Web cuenten con metadatos, las revistas digitales que sirven como bases a su disciplina usarán mayoritariamente metadatos y lo harán correctamente”. (Cristòfol Rovira y Mari-Carmen Marcos, 2006)

Los resultados que se encontraron en este estudio, en primera instancia eran alentadores debido a que se encontró en la revisión que el 97% de los sitios cuentan con algún tipo de metadato, esto no significaba que los metadatos se encontraran en todas la páginas; por ejemplo, el porcentaje de homepages con metadatos fue del 85%. Además, un 21% de los sitios Web contenían, solamente metadatos de tipo técnico generados automáticamente por el editor de html, y un 31% presentan los artículos en formato pdf. (Cristòfol Rovira y Mari-Carmen Marcos, 2006)

Tomando en cuenta lo anterior los autores del estudio estiman que no se cumplió la hipótesis planteada al inicio de la investigación, esto debido a que el 59% de estos sitios no usan metadatos útiles para la recuperación de la información, esto puede ser por tres razones:

1. Por qué no existen
2. Son solo de tipo técnico, generados automáticamente
3. Los artículos están en formato PDF

Solamente el 13% usa metadatos específicos en las páginas Web de los artículos, es decir, metadatos adecuados para la descripción de esas páginas. (Cristófol Rovira y Mari-Carmen Marcos, 2006)

Por supuesto, que esto definitivamente, va en contra de los usuarios que buscan y tratan de recuperar información en la red, puesto que muchos de los sitios que contienen información académica no cuentan con metadatos acordes o útiles para la recuperación de la información, y esto hace que los usuarios al desconocer esta situación, pues no saben por qué razón no encuentran los documentos que necesitan y solo piensan que definitivamente no existe la información que ellos buscan, porque al realizar la búsqueda en un buscador estos sitios no pueden ser encontrados fácilmente, porque los sitios carecen de metadatos adecuados para su descripción y por lo tanto el motor de búsqueda no puede recuperar adecuadamente la información solicitada.

2.6.4.3 El uso de protocolos inadecuados.

Otro problema que existe en la recuperación de información y que repercute principalmente en los usuarios que tratan de encontrar y recuperar información en los buscadores de la red, es sin lugar a dudas la utilización en los sitios Web de protocolos de intercambio y recuperación de información inadecuados, con la finalidad de establecer entre los Sistemas de Información la

interoperabilidad, para posibilitar la transferencia e intercambio de información entre sistemas. (Gómez, 2007)

Gómez menciona que la interoperabilidad es “ la capacidad de un sistema de información de comunicarse y compartir datos, información, documentos y objetos digitales de forma efectiva (con una mínima o nula pérdida de su valor y funcionalidad), con uno o varios sistemas de información (siendo generalmente estos sistemas completamente heterogéneos, distribuidos y geográficamente distantes), mediante una interconexión libre, automática y transparente, sin dejar de utilizar en ningún momento la interfaz del sistema propio”. (Gómez, 2007)

En los Sistemas de Información, esta capacidad de interoperar entre ellos está dada por el uso de protocolos de intercambio, transferencia y recuperación de información bibliográfica.

En este sentido, se puede mencionar que son numerosos los protocolos desarrollados a través de los años con el fin de permitir la interoperabilidad entre Sistemas de Información, entre ellos: (Guajardo Salinas, 2010)

- Z39.50
- Simple Digital Library Interoperability Protocol (SDLIP)
- OAI (Open Archives Protocol)
- Guildford protocol
- Dienst protocol
- ZING SRU/W
- Starts (Stanford Protocol Proposal for Internet Retrieval and Search)
- Dienst protocol
- Lyceum Protocol
- Harvest: A Distributed Search System

Sin embargo, sólo algunos de éstos se han consolidado internacionalmente como estándares aceptados por los profesionales de la información, y son

usados de manera masiva y su desarrollo y mantenimiento se encuentra avalado por iniciativas tanto públicas como privadas, estos son:

- El Protocolo Z39.50 y
- El Protocolo OAI-PMH.

El Protocolo OAI-PMH

Para efectos de este trabajo solamente nos interesa el protocolo OAI-PMH (Open Archives Initiative - Protocol Metadata Harvesting o Iniciativa Abierta de Archivos – Protocolo de Recolección de Metadatos) es una herramienta de interoperabilidad que posibilita el intercambio de metadatos sobre cualquier material almacenado en soporte electrónico. Esta transferencia puede realizarse desde puntos proveedores de servicio a través de búsquedas que abarquen la información recopilada en distintos repositorios de archivos asociados (proveedores de datos).

La Open Archives Initiative se creó con el fin de desarrollar y promover estándares de interoperabilidad para facilitar la difusión de contenidos en Internet. Surgió como un esfuerzo para mejorar el acceso a archivos de publicaciones electrónicas científicas (eprints), pero luego se hizo extensiva a la comunicación e intercambio de metadatos de materiales digitales.

Básicamente, OAI-PMH utiliza transacciones HTTP (Hypertext Transfer Protocol o Protocolo de Transferencia de Hipertexto, que es el protocolo usado en cada transacción de la World Wide Web) para emitir preguntas y obtener respuestas entre un servidor o archivo y un cliente o servicio recolector de metadatos.

Esta estructura de interoperabilidad corresponde al enfoque de harvesting (recopilación), en la que los datos y metadatos son transferidos desde la fuente remota al destino en el cuál se realizarán los servicios de búsqueda.

“El protocolo soporta múltiples formatos para expresar los metadatos, no obstante requiere que todos los servidores ofrezcan los registros utilizando Dublin Core codificado en XML. Además de este formato cada servidor es libre de ofrecer los registros en otros formatos adicionales (MARC por ejemplo). Un cliente puede pedir que los registros se le sirvan en cualquiera de los formatos soportados por el servidor, esto con el fin de que en el futuro las diferentes comunidades que utilicen el protocolo definan sus propios formatos que sean más precisos que el Dublin Core. Por ejemplo la comunidad de archivos de e-prints está trabajando en un formato denominado AMF (Academic Metadata Format) que describirá todos los elementos que intervienen en el proceso de comunicación científica: documentos, autores, instituciones y canales de distribución de documentos.”(Barrueco.J.M., 2003: 99-106)

La arquitectura de OAI-PMH se basa en clientes y servidores; los primeros son los archivos que proporcionan la información y los segundos son los recolectores o servicios que toman los datos, y los presentan a los usuarios finales.

Desde 2001 se ha mantenido constante el crecimiento del número de archivos que han implementado el protocolo OAI-PMH, hasta llegar a más de 45 archivos de diferentes disciplinas. Se pueden destacar el CERN (informes y publicaciones en Física), Citebase, y los archivos del área de Bibliotecología y Documentación, tales como @rchiveSIC, DLIST (Digital Library of Information Science and Technology) y E-LIS (Eprints in Library and Information Science).

Uso del Protocolo OAI-PMH

OAI-PMH es una plataforma simple para acceder a la información bibliográfica disponible en un archivo o repositorio. De esta forma, cualquier institución puede realizar una implementación del protocolo para poner a disposición su propia base de datos o catálogo, y a la vez consultar otras.

Para su implementación el protocolo requiere ciertos requisitos, ya sea como proveedor de datos o como proveedor de servicios, que pueden dificultar la utilización de este protocolo:

- Metadatos (almacenados en BD)
- Servidor Web
- Programación de interfaz
- Identificador de archivo/URL Base
- Formato de metadatos
- Jerarquía lógica de sets
- Control del flujo
- Generador de errores (Respuestas XML)
- Acceso a BD para extraer los metadatos de acuerdo al formato de metadatos requerido
- Generador XML para crear las respuestas

Actualmente existen algunos sistemas integrados de gestión de bibliotecas que entre sus funcionalidades cuentan con un cliente OAI-PMH ya instalado, solución más práctica y expedita.

Aplicaciones del Protocolo OAI-PMH

De la creación de contenidos accesibles en la Web a partir de la información conservada en archivos, se derivan las limitaciones que desde el punto de vista funcional presentan los buscadores para encontrar recursos digitales en general y documentales en particular. Hasta hace algún tiempo, la utilidad de los buscadores se circunscribía a páginas estáticas HTML aunque últimamente se ha incorporado la búsqueda en el contenido de documentos en PDF, Word y otro tipo de formatos. “Desde luego, las bases de datos permanecen inaccesibles para ellos [los buscadores], lo cual representa un enorme reto a la hora de aumentar la visibilidad los recursos digitales que produzcan los archivos.”

La solución a este inconveniente es precisamente el uso de la recolección de metadatos (metadata harvesting) a través del protocolo OAI-PMH.

Según la opinión de algunos autores, el OAI-PMH debiese ser un protocolo a implantar en aquellas instituciones que quieran no sólo crear contenidos para la Web sino también implantar aplicaciones que favorezcan su localización.

Si bien OAI-PMH surgió en el seno de la comunidad académica y científica para la búsqueda y recuperación de textos electrónicos, es perfectamente aplicable en cualquier contexto documental, ya que es una alternativa a la visión del mecanismo de búsqueda y recuperación propuesto por el modelo distribuido de Z39.50.

Además, las características de flexibilidad del OAI-PMH para adaptarse a cualquier ámbito, pudiendo ofrecer información sobre todo tipo de recurso, tanto físico como digital, permiten la diseminación de información a través de la generación de metabuscadores mediante su uso y por tanto la formación de repositorios de información y de bibliotecas digitales.

Como conclusiones a este capítulo, se puede mencionar que definitivamente el incremento de la información contenida en la Web, es de consideraciones impactantes, haciendo que se incrementa el tamaño de la red día con día. Esto debido a las facilidades que ofrecen algunos servicios, que permiten que los usuarios de la red puedan colocar sus propias páginas Web en los servidores de estos servicios, dando paso a sitios que no cuentan con un soporte académico y por lo tanto al ser indizados por las bases de datos de los motores de búsqueda, hacen que estos se llenen de sitios que al recuperarlos en una búsqueda ofrezcan una gran cantidad de documentos que no tienen valor académico y sean considerados como basura.

Por otro lado, no toda la información que se encuentra en la red carece de soporte académico, sino al contrario existe una gran cantidad de información que está respaldada por instituciones académicas, el problema es que las herramientas de búsqueda que existen en la Web no las pueden recuperar eficientemente debido a que la información no está organizada o está de

manera inadecuada, por lo tanto es importante, que al crear las bases de datos de los motores de búsqueda, estas se diseñen de forma tal que permitan que la información sea organizada de tal manera que cuente con elementos suficientes para poder recuperarla eficientemente y así los usuarios no pierdan tiempo al revisar cantidades impresionantes de documentos, que en ocasiones no les son útiles para sus actividades académicas.

CAPITULO 3

EVALUACIÓN DE HERRAMIENTAS DE BUSQUEDA EN LA WEB

Dentro del ámbito de las herramientas de búsqueda en la Web, además de crearlas y utilizarlas, es necesario conocer cuál es la forma en la que están trabajando, para saber si cumplen con su objetivo principal, el de permitir que los usuarios recuperen la información que necesitan, para satisfacer las necesidades de ellos. Para conocer si estas herramientas recuperan adecuadamente la información académica, es necesario realizar una evaluación, que nos permita conocer el grado de efectividad con que las herramientas de búsqueda recuperan la información académica.

3.1 Selección

Para llevar a cabo, esta evaluación, fue necesario realizar una selección de los buscadores y metabuscadores que se encuentran operando en la red. Lo cual resultó difícil, debido a que existe una gran cantidad de ellos, de diferentes tipos y tamaños, se estima que existen actualmente en la red más de 5300 buscadores. Por esta razón y para efectos de esta investigación, se utilizó solamente los motores de búsqueda, metabuscadores y los buscadores semánticos, más utilizados en la Web.

Para seleccionar las herramientas de búsqueda antes mencionadas, se utilizó un sitio que es considerado un directorio de buscadores, llamado Beaucoup, que contiene más de 2500 herramientas. (www.beaucoup.com, 2015). Este directorio sirvió como un apoyo de inicio para realizar la selección de las herramientas más utilizadas por los usuarios de la red.

Para la selección de los motores de búsqueda y metabuscadores que mayormente se utilizan por parte de los usuarios que acuden a ellos en la búsqueda de información en la red, se tomaron como referencia tres sitios Web

que han hecho investigación sobre cuáles son los buscadores más utilizados, siendo los siguientes:

1. eBizMBA
2. ListofSearchEngines
3. Reliabelsoft

Con respecto a los metabuscadores, se tomó como referencia para su selección a los que aparecen en la página de Listofsearch en el rubro de metabuscadores.

En el caso de los buscadores semánticos, se utilizaron dos sitios Web, estos son:

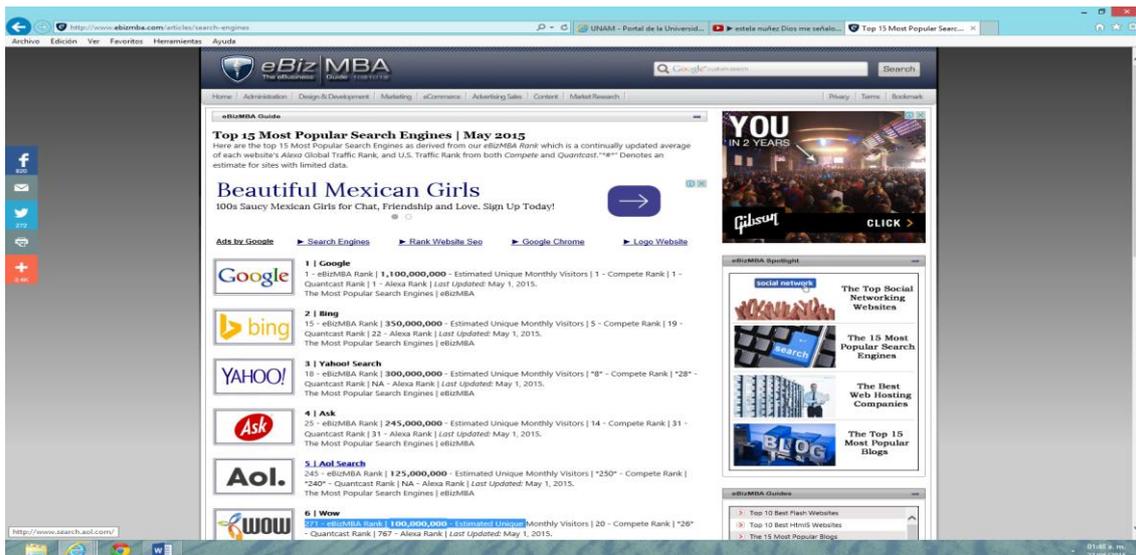
- 1.- Search engine journal
- 2.- Makeuseof

3.2 Fuentes utilizadas en la selección de las herramientas de búsqueda

En este apartado se describe de manera sencilla cada una de las fuentes que se utilizaron para la selección de los motores de búsqueda, metabuscadores y buscadores semánticos, así como el lugar que ocupan las herramientas de búsqueda en el ranking de estos sitios.

eBizMBA

En el sitio de eBizMBA se incluyen 15 motores de búsqueda más utilizados, estos buscadores están respaldados por Rango eBizMBA que es una media actualizada continuamente de Alexa Global, por el ranking de tráfico de cada sitio Web, su última actualización fue el primero de mayo de 2015.



Página principal de eBiz MBA <http://www.ebizmba.com/articles/search-engines>

En la tabla # 1 eBizMBA, se muestra el rango sobre los buscadores más consultados de acuerdo al uso de cada uno de ellos de forma mensual, con base al ranking de tráfico global.

Tabla #3.1 eBizMBA visitantes al mes

Buscador	Rango eBizMBA visitantes al mes
1. Google	1,100,000,000
2. Bing	350,000,000
3. Yahoo	300,000,000
4. Ask	245,000,000
5. AOL	125,000,000
6. Wow	100,000,000
7. WebCrawler	65,000,000
8. MyWebSearch	60,000,000
9. Infospace	24,000,000
10. Info	13,500,000
11. DuckDuckGo	13,000,000
12. Blekko	12,500,000
13. Contenko	11,000,000
14. Dogpile	10,500,000
15. Alhea	7,500,000

Se puede observar en la tabla anterior, eBizMBA que contempla los 15 buscadores más utilizados de acuerdo a rango de visitantes por mes, que definitivamente sorprende el número de consultas mensuales que van del número 1 al número 5 desde 1, 100, 000, 000 de consultas mensuales de

Google, hasta las 125, 000, 000 de AOL. Otro aspecto importante de esta lista es que integra en ella a un metabuscador como lo es Dogpile.

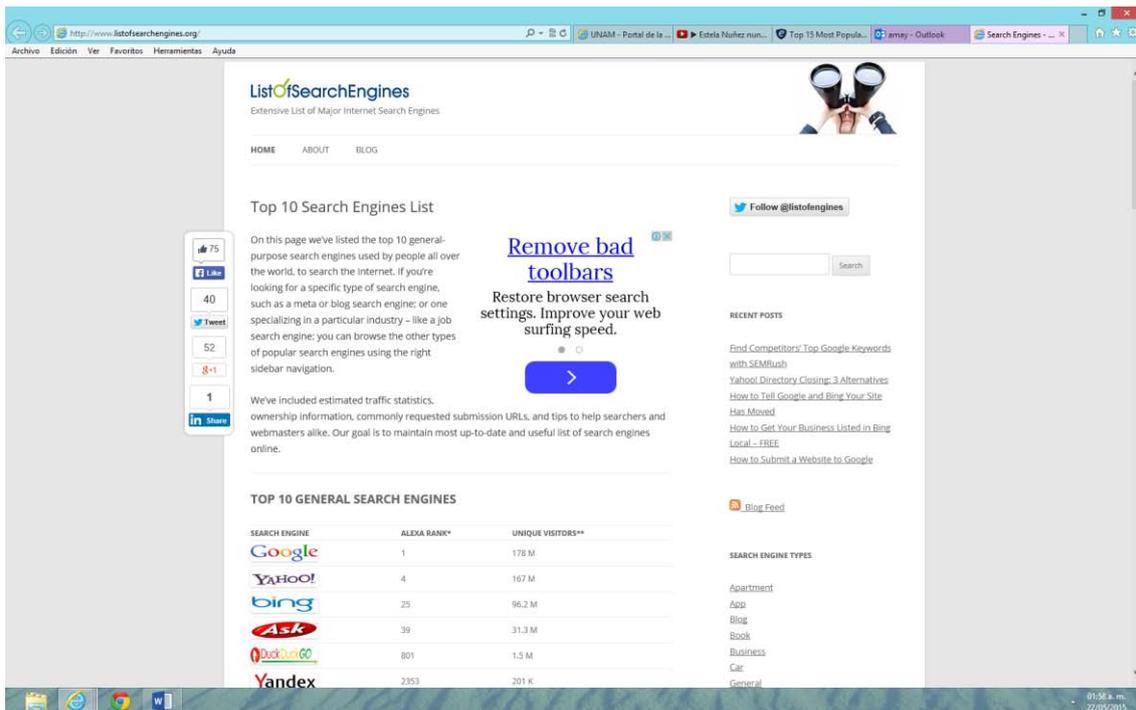
ListofSearchEngines

Otro sitio que hace una valoración de los buscadores más utilizados es ListofSearchEngines, el cual presenta una lista de los 10 buscadores más utilizados, en Estados Unidos de acuerdo al ranking de Alexa.

Alexa, es una compañía que depende de Amazon.com. Es el medidor internacional por excelencia. Su función es realizar un ranking con las páginas más visitadas generando gráficos y dando estadísticas del tráfico de la Web proporcionando información del uso de estas herramientas por los usuarios, para hacer este ranking, Alexa, se basa en una estimación de visitas y no en visitas reales

La empresa recopila datos del tráfico mediante la navegación de millones de usuarios que usan una barra de navegación especial. A través de esta barra Alexa, pueden recopilar información como el número de usuarios que entran en la página Web y el número de visitas.

El sitio principal de ListofSearchEngines, muestra los 10 buscadores más utilizados en los Estados unidos.



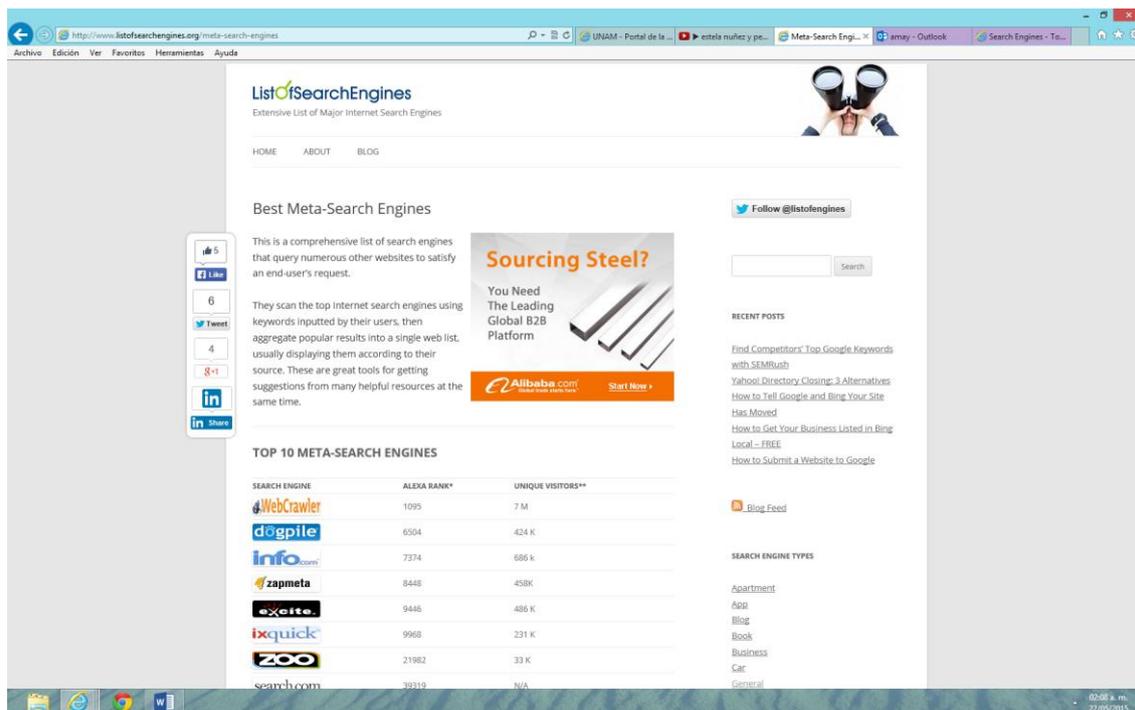
Página principal de ListofSearchEngines <http://www.listofsearchengines.org/>

Tabla # 3.2 ListofSearchEngines visitantes al mes

Buscador	Ranking Alexa visitantes al mes
1. Google	178, 000, 000
2. Yahoo	167, 000, 000
3. Bing	96, 200, 000
4. Ask	31, 300, 000
5. DuckDuckGo	1, 500, 000
6. Yandex	201, 000
7. Blekko	3, 400, 000
8. Scrub La Web	30, 000
9. Entireweb	49, 000
10. Gigablast	No aplica

Se puede apreciar en la tabla #2, la lista que es muy parecida a la de eBiz MBA, en cuanto a los primeros cinco buscadores, aunque no es muy clara con respecto al número de usuarios o consultas que tiene cada buscador por mes.

Por otro lado este mismo sitio también contempla un apartado sobre metabuscadores, para el caso de estas herramientas de búsqueda, se tomó como referencia solamente a los que aparecen en el sitio de Listofsearch en el rubro de metabuscadores.



Página de Metabuscadore <http://www.listofsearchengines.org/meta-search-engines>

Esta es una lista completa de metabuscadore que consultan otros numerosos sitios Web para satisfacer la petición de un usuario final. Ellos buscan en diferentes herramientas, principalmente en los motores de búsqueda de Internet utilizando palabras clave introducidas por sus usuarios, a continuación, se agregan los mejores resultados en una sola lista de la Web, por lo general se presentan en función de su origen. Estos son una gran herramienta para obtener sugerencias de muchos recursos útiles al mismo tiempo.

Tabla # 3.3 ListofSearchEngines metabuscadore visitantes al mes

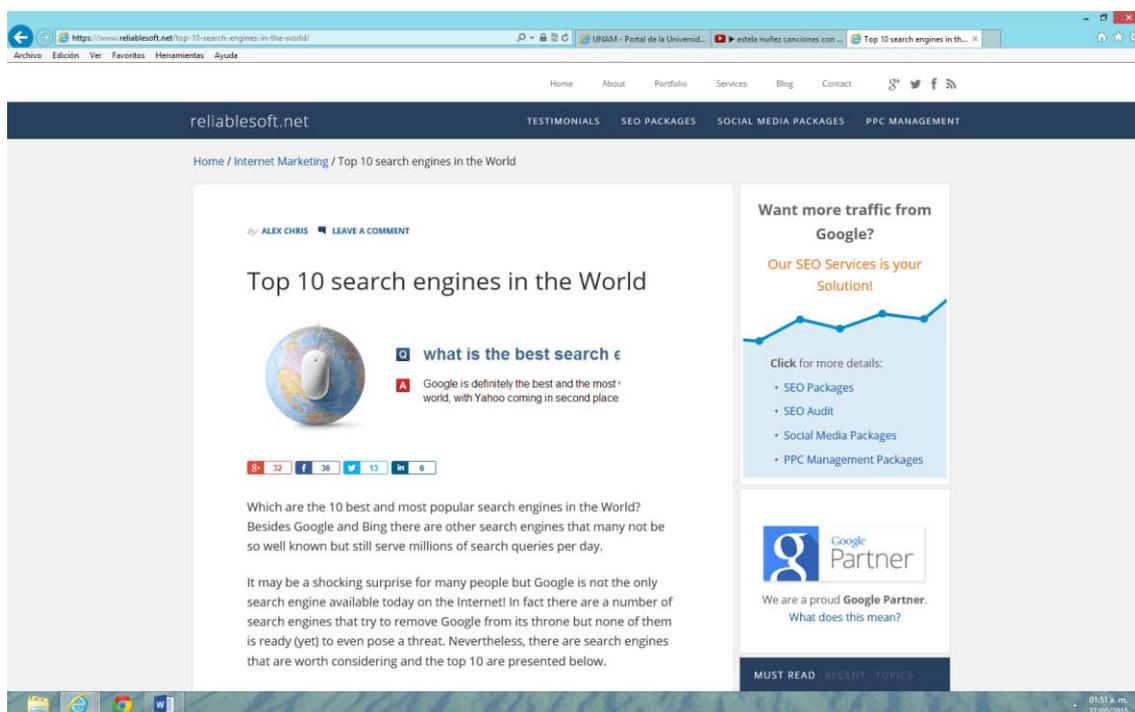
Metabuscadore	Ranking Alexa visitantes al mes
1. WebCrawler	7, 000, 000
2. Dogpile	424, 000
3. Info.com	686, 000
4. Zapmeta	458, 000
5. Exite	486, 000
6. Ixquick	231, 000
7. Search.com	N/A
8. Zoo	N/A
9. Mamma	N/A
10. Yippy	No aplica

Alexa ranking de tráfico (global). Estima Mensual único visitante de Compete.com (sólo en EE.UU.). Actualización para el 2014.

En la tabla #3 se puede observar que solamente seis de los 10 metabuscadores, cuentan con un ranking de Alexa, en los últimos cuatro no se establece un número de visitantes por mes, esto tal vez podría deberse a que no se tienen datos confiables de su uso por parte de los usuarios de la red en Estados Unidos, o que son tan pocas las consultas a ellos que probablemente no alcanzan el estándar para que Alexa los pueda contabilizar.

Reliablesoft.

Otro sitio que rankea a los mejores 10 buscadores, es Reliablesoft.



Página principal de <https://www.reliablesoft.net/top-10-search-engines-in-the-world/>

Tabla # 3.4 Reliablesoft.

Buscador
1. Google
2. Bing
3. Yahoo
4. Ask
5. AOL
6. Blekko
7. WolframAlpha
8. DuckDuckGo
9. WayBackMachine
10. ChaCha.com

En la tabla #4, se puede observar que este sitio presenta otra variante, debido a que no informa como se llevó a cabo el ranking de las herramientas de búsqueda, es decir no menciona cuantos usuarios o búsquedas al día, semana o mes tienen los buscadores que presenta, solamente da el ranking del 1 al 10, según su percepción, pero para efectos de esta investigación, nos sirve debido a que nos permite darnos cuenta cuales son los buscadores más populares de internet.

Como ya se mencionó anteriormente, para el caso de los buscadores semánticos, se utilizaron solamente dos sitios Web. El sitio **Search engine journal** y el sitio **Makeuseof**

Search engine journal, este sitio presenta una lista de los nueve buscadores semánticos, más utilizados por los internautas.



<http://www.searchenginejournal.com/semantic-search-engines/9832/>

En la tabla # 5 se presentan los nueve buscadores semánticos que rankea este sitio, pero solamente funcionan cuatro de ellos, los otros cinco están fuera de servicio o tardan mucho al realizar una búsqueda y al final aparece la leyenda

de que no se puede acceder a esa página. Los buscadores semánticos que presenta son:

Tabla #3.5 Search engine journal buscadores semánticos

Buscadores semánticos	Funciona	No funciona
Hakia		X
Kosmix		X
Exalead		X
SenseBot		
Cognición Buscar		
Lexxe		
Swoogle		
Factbites		
Powerset		

Makeuseof

The screenshot shows a web browser window displaying the article 'Top 7 Semantic Search Engines As An Alternative To Google' on the Makeuseof website. The article is dated May 1, 2010, and is written by Nancy Messiah. The page layout includes a navigation menu, a search bar, and a newsletter sign-up form. There are also several advertisements and sponsored links, including 'Search Engine', 'Webcam Recording Software', and 'Mapas y Direcciones'. The article text discusses the power and popularity of Google search engine and compares it to other similar search engines like Bing, which are based on page rankings and algorithms. It mentions that semantic search engines are able to understand the context in which the words are being used, resulting in smart, relevant results.

<http://www.makeuseof.com/tag/top-7-semantic-search-engines-alternative-google-search/>

Con respecto a este sitio, en la tabla #6 se aprecia que ofrece siete buscadores semánticos y presenta casi los mismos problemas que el sitio anterior, entre los buscadores semánticos que lista se encuentran los siguientes:

Tabla # 3.6 Makeuseof buscadores semánticos

Buscadores semánticos	Funciona	No funciona
Hakia		X
Kngine		X
Kosmix		X
DuckDuckGo		
Evri		
Powerset		
Truevert		

Al revisar estos dos sitios nos pudimos percatar que presentan en su conjunto 11 buscadores semánticos, pero al tratar de utilizarlos solamente se pudo tener acceso a cuatro de ellos, por lo tanto y para efectos de esta evaluación solo se consideraran cuatro buscadores semánticos.

Después de haber revisado, cuales son los Motores de búsqueda, metabuscadores y buscadores semánticos más utilizados, se determinó, utilizar solo cinco motores, cinco metabuscadores y cuatro buscadores semánticos, los cuales se presentan en la tabla #7:

Tabla # 3.7 motores de búsqueda, metabuscadores y buscadores semánticos

Motores de búsqueda	Metabuscadores	Buscadores semánticos
1. Google	WebCrawler	Kngine
2. Yahoo	Dogpile	Exalead
3. Bing	Info.com	DuckDuckgo
4. Ask	Zapmeta	Swoogle
5. AOL	Exite	

Fuente: elaboración propia

3.3 Características de las herramientas de búsqueda seleccionadas

A continuación se presenta una breve descripción de cada una de las herramientas de recuperación de información seleccionadas para el presente estudio.

Motores de búsqueda

Google: es el motor de búsqueda más popular a nivel mundial. Originalmente llamado BackRub, en los Estados Unidos es considerado por la mayoría de las personas como uno de los mejores buscadores de la red, obteniendo el 65% de las consultas de la Web. El 03 de octubre 2010, Alexa lo reportó como el sitio Web más visitado en todo el mundo.

Yahoo: fue originalmente un directorio de sitios Web, se convirtió en su propio motor a base de orugas en 2002. A finales de 2009, Yahoo! fue el segundo mayor motor de búsqueda en la Web. Sin embargo, a finales de agosto de 2010, los resultados de Yahoo comenzaron a utilizar el índice de Bing. A partir del 01 de febrero de 2012, Alexa lo reporta como el cuarto sitio Web más visitado en todo el mundo.

Bing: Lanzado en 2009 por Microsoft, Bing es un servicio de búsqueda basado en la Web. A finales de agosto, Bing comenzó oficialmente la entrega de resultados hoo! 'S. Los antiguos motores de búsqueda de Microsoft incluyen Microsoft Live y MSN. A partir del 1 de febrero 2012, Alexa posiciona a Bing como el sitio Web más visitado en todo el mundo.

Ask: formalmente conocido como Ask Jeeves, Ask fue fundada en 1996 con la idea de permitir a los usuarios obtener respuestas a las preguntas formuladas en la vida cotidiana, en lenguaje natural, así como la búsqueda de palabras clave tradicionales. Hoy Ask.com incluye soporte para las preguntas de matemáticas, diccionario, y conversión.

AOL: Inició como un directorio en 1998, los contenidos formaban parte del directorio Open Directory Project (DMOZ). En octubre de 1999 obtuvo el dominio aolsearch.com,. En 1999, integró el motor de búsqueda Inktomi para indexar el directorio, a partir de octubre del año 2003, AOL Search llegó a un acuerdo con Google para utilizar sus resultados.

Metabuscadores

MetaCrawler: uno de los metabuscadores más antiguos de la red, MetaCrawler inicio en julio de 1995 en la Universidad de Washington. Fue comprado por InfoSpace, un proveedor de contenido en línea, en febrero de 1997

Dogpile: metabuscador que envía las búsquedas a una lista personalizada de diferentes motores de búsqueda, directorios y sitios especializados, a continuación, muestra los resultados de cada motor de búsqueda de forma individual, es propiedad de InfoSpace.

Zapmeta: una herramienta de búsqueda que proporciona a los usuarios la capacidad de buscar simultáneamente en varios motores de búsqueda bajo una misma interfaz. Junto con la búsqueda Web, ZapMeta ofrece actualmente un directorio basado en datos de El Open Directory Project y Búsqueda de productos impulsados por PriceGrabber.

Info.com: es un metabuscador que funciona como una plataforma de búsqueda de gran alcance que reúne lo mejor de la Web, los buscadores en los que busca información, se encuentran:  Google, Yahoo!, Bing.com, Ask, LookSmart, About y Open Directory.

Excite: anteriormente un motor de búsqueda rastreador basado con su propio índice, se transformó en metabuscador cuando es adquirido por InfoSpace en 2001.

Buscadores semánticos

Ngine: es un buscador semántico, que da significado a las palabras y permite ubicar de manera inteligente los contenidos, además hace comparaciones de temas o resultados, contiene en sus base de datos, información actualizada del tiempo, cambio de diferentes monedas, el precio de acciones, resultados deportivos, asimismo vincula todos los datos relacionados con la temática solicitada.

Exalead: es un buscador de Internet que destaca por sus innovaciones y experiencia visual. Se desarrolló en Francia y su interfaz está disponible en inglés, francés, italiano y alemán.

Al ingresar al sitio Web el usuario encontrará en la parte de arriba una presentación similar a la de Google, sin embargo en la parte de abajo están disponibles 8 atajos (distribuidos en cuatro columnas, con dos filas cada uno), los cuales incorporan una imagen miniatura.

Cuando el usuario realiza una búsqueda se encuentra con una lista de resultados, con el nombre de la página Web, una pequeña descripción y unos sub-directorios. Pero lo más agradable son las imágenes miniatura que se sitúan a la izquierda. También destaca la columna derecha, llamada "Refine su búsqueda", mediante la cual se pueden filtrar los resultados con las siguientes opciones: Temas relacionados, Tipo de sitio (Blog, Foro, etc.), Multimedia (Audio, Video o RSS), Idiomas, Directorios y Tipo de archivo.

DuckDuckgo es un motor de búsqueda cuyo objetivo es el respeto a la privacidad del usuario al no recoger información sobre éste que lo identifique. Detrás del proyecto está Gabriel Weinberg, master en Tecnología en el MIT (Massachusetts Institute of Technology). El proyecto es de código abierto y está escrito en Perl, el motor de búsqueda DuckDuckGo es rápido y ofrece resultados similares en una búsqueda que cualquier otro. Únicamente recupera documentos en idioma inglés como única alternativa. Dispone de un plugin para FireFox, así como para iPhone e iPad. Técnicamente está basado en APIS de otros motores de búsqueda, como Yahoo, por lo que se le considera un motor híbrido, las búsquedas las efectúa en sitios que contienen información pública.

Swoogle

Después de haber presentado un breve resumen de las herramientas de búsqueda seleccionadas para realizar el presente estudio, fue necesario que se vieran también los criterios e indicadores que se utilizaron para la evaluación de cada uno de los motores de búsqueda, metabuscadores y buscadores semánticos.

3.4 Estrategia de búsqueda

Como resultado de este proceso de selección de herramientas de búsqueda de información, se consideraron cinco motores de búsqueda, cinco metabuscadores y cuatro buscadores semánticos.

Asimismo, para la evaluación de las herramientas de búsqueda seleccionadas se utilizaron diferentes criterios, indicadores y variables, así como otros aspectos, entre los que se encuentran, los criterios de Levene, las variables de Lara y Martínez, así como identificar si las herramientas de búsqueda utilizan metadatos, RDF y ontologías. Finalmente fue necesario estructurar una

estrategia de búsqueda, que nos permitiera conocer si las herramientas de búsqueda recuperaban la información eficientemente.

Otro aspecto importante, en la investigación fue la selección del tema para comparar las búsquedas en las diferentes herramientas, el tema seleccionado, fue:

“Las relaciones diplomáticas entre Estados Unidos y Cuba” en español

“Diplomatic relations between the United States and Cuba” en inglés

La decisión de incluir este tema obedeció a que es una temática muy relevante en estos momentos, tanto para los dos países involucrados, Estados Unidos y Cuba, así como a nivel internacional. Por otro lado, se decidió que la búsqueda debería realizarse en idioma Inglés y español, debido a que son los idiomas oficiales de los países protagonistas.

Como ya se mencionó, la búsqueda se llevó a cabo en los dos idiomas, por lo tanto fue necesario estructurar una estrategia de búsqueda con la temática seleccionada.

Para dar inicio a la búsqueda se decidió utilizar los siguientes términos que se presentan en la tabla #8:

Tabla #3.8 términos de estrategia de búsqueda

Términos en español	Términos en inglés
cuba	between
diplomaticas	cuba
entre	diplomatic
estados	relations
relaciones	states
unidos	united

Fuente: elaboración propia

Se puede apreciar que tanto en español como en inglés se eliminaron las palabras vacías, es decir palabras que no tienen significado en la búsqueda,

como son: los artículos, de igual manera todos los términos se escribieron en minúsculas y en el caso de la búsqueda en español no se acentuaron los términos, es importante aclarar que las estrategias se utilizaron para buscar información en todas las herramientas seleccionadas.

Búsqueda libre español inglés

Se buscó en todas las herramientas de manera libre, es decir utilizando todos los términos seleccionados y se colocó en las casillas de búsqueda de la siguiente manera:

- **relaciones diplomaticas entre estados unidos cuba**
- **diplomatic relations between united states cuba**

Búsqueda con operadores lógicos boléanos español inglés

Para esta estrategia de búsqueda se utilizaron dos opciones, el operador lógico **AND** y se combinó con las comillas "" para unir dos palabras, es necesario aclarar que el operador **AND** se colocó en mayúsculas, quedando de la siguiente manera:

- relaciones **AND** diplomaticas **AND** entre **AND** "estados unidos" **AND** cuba
- diplomatic **AND** relations **AND** between **AND** " united states " **AND** cuba

Búsqueda por campos español inglés

Para obtener búsquedas más precisas se decidió buscar en el campo de título, combinando también operadores lógicos boléanos y las comillas, quedando de la siguiente manera:

intitle:relaciones AND intitle:diplomaticas AND intitle:entre AND intitle:"estados unidos" AND intitle:cuba

intitle:diplomatic AND intitle:relations AND intitle:"united states" AND intitle:cuba

3.5 Criterios de evaluación para las herramientas de búsqueda

Para dar inicio a esta evaluación, sobre la búsqueda y recuperación de información eficiente en buscadores que se encuentran en la red, fue necesario contar con una serie de variables que nos permitan realizar un análisis, de cada una de las herramientas de búsqueda que se seleccionaron previamente.

En este sentido, Lara y Martínez (2006:14-15) mencionan que para realizar una consulta es necesario tener en cuenta un conjunto de variables, como son:

1. Lenguaje de interrogación, que debe ofrecer diferentes tipos de operadores: lógicos, de comparación, de truncamiento, de proximidad, de especificación de campo.
2. Posibilidad de refinar (refine) una búsqueda inicial.
3. Campos limitadores que nos permitan reducir la búsqueda: dominios, lenguas, países, fecha de creación del recurso.
4. Búsquedas alternativas: búsqueda simple, búsqueda avanzada, búsquedas combinando operadores e índices temáticos, etc
5. Opciones avanzadas: buscar diferentes recursos (texto, sonido, imagen), guardar y reutilizar búsquedas, diferentes formatos en los resultados de búsqueda (estándar, detallado, compacto, etc.), búsqueda de

conceptos relacionados (related topics), consulta directa en bases de datos (infranets), etc.

Por otro lado, también se utilizaron los criterios de Levene (2010), estos son:

Localización: Este indicador se refiere a que en una búsqueda, en el resultado aparezcan primero los documentos en donde se encuentren la mayoría de las palabras utilizadas en la búsqueda. Asimismo, se habla de una mayor relevancia de los documentos, cuando las palabras que se buscan aparecen principalmente en el título.

Frecuencia de aparición: Este indicador se refiere a que mientras más veces aparezca un término de la búsqueda en un documento, tendrá más relevancia para el resultado.

Popularidad: Este indicador mide el número de enlaces que apuntan a una página. Por lo tanto si una página cuenta con muchas referencias suele considerarse que es más popular y por lo tanto mejor que otra que tiene menos referencias.

Precio: Este indicador se ve principalmente en los buscadores comerciales, en los cuales se cobra una cuota para que una página aparezca antes que otras, derivado del pago realizado.

Interfaz

El motor de consultas proporciona la interfaz entre el índice de búsqueda, el usuario, y la Web. El motor de consultas procesa una consulta del usuario en dos pasos. En el primer paso, el motor de consulta recupera de la información del índice de búsqueda de páginas Web potencialmente relevantes que coinciden con las palabras clave en la consulta del usuario, y en la segunda

etapa se produce un ranking de los resultados, a partir de las más relevantes a las menos relevantes.

Asimismo, el indicador de interfaz se basa en páginas Web con formularios, entre ellos podemos mencionar:

Formularios: El mecanismo de entrada de datos de las páginas Web son formularios normalmente basados en una caja de texto (en donde el usuario introduce la palabra o frase buscada) y un botón de envío (al pinchar sobre él se envía la consulta). Existen otras soluciones que permiten búsquedas más avanzadas con formularios más complejos que permiten, por ejemplo, introducir varias palabras, añadir expresiones booleanas, buscar en un idioma concreto, buscar por proximidad, etc.

Páginas Web de resultados: los resultados se muestran en una página Web en grupos de ítems. Cada ítem contiene una pequeña descripción, el contexto en el que se ha encontrado y el enlace. Existen también soluciones más avanzadas que permiten la traducción automática, etc

Posterior a la búsqueda de información en cada buscador y metabuscador, se llevará a cabo un análisis de los resultados para saber si las herramientas de búsqueda seleccionadas cumplen con los aspectos antes mencionados y si recuperan eficientemente la información.

Otros aspectos, que también se analizarán, tienen que ver con lo verdaderamente importante de este estudio, estos aspectos son: si los motores de búsqueda, metabuscadores y buscadores semánticos, usan estructuras de metadatos en la organización de los índices que contienen la información en la Web.

Por otro lado también es importante conocer si en la organización de sus índices, utilizan RDF para describir el contenido y las relaciones de contenido disponibles en un sitio Web o en una colección de documentos, así como para

describir los derechos de propiedad intelectual o las políticas de privacidad de un sitio Web, asimismo es necesario conocer si utilizan una descripción RDF y si crean tripletas.

Otro aspecto importante es el uso de ontologías en la creación de los índices para darle significado a los contenidos de información, creando vocabularios controlados, dándole un mayor nivel de profundización semántica, además de proporcionar una descripción lógica y formal que pueda ser interpretada tanto por las personas, como por las máquinas.

3.6 Resultados de la evaluación

3.6.1 Resultados de búsqueda en los motores

Antes de iniciar con la presentación de los resultados de la evaluación que se llevó a cabo en los motores de búsqueda, en la tabla #9 se muestran los resultados que se obtuvieron al revisar los motores previamente seleccionados, con respecto a las principales características de búsqueda de información sugeridas por Levene.

Tabla #3.9 características de los motores de búsqueda

Buscador	Lenguaje de interrogación	Refinar búsqueda	Campos limitadores	Búsquedas alternativas	Opciones avanzadas
Google	Si	Si	Si	Si	Si
Bing	Si	Si	Si	Si	Si
Yahoo	Si	Si	Si	Si	Si
Ask	Si	Si	Si	Si	Si
AOL	Si	Si	Si	Si	Si

Fuente: elaboración propia

Como se puede apreciar, los cinco motores seleccionados, cumplen con todos los aspectos de las variables de una búsqueda.

Con respecto a la variable de lenguaje de interrogación, permiten el uso en la búsqueda de diferentes operadores, como los lógicos AND, OR y NOT. Además permiten utilizar otros operadores como el caso de las comillas “ “ que se utiliza para buscar por frase, permiten también buscar por algunos campos: título, url, cache, tipo de archivo o formato, etc.

En el caso de la segunda variable, posibilidad de refinar la búsqueda. Los cinco permiten refinar la búsqueda sin salir de la misma, esto se da desde el cuadro de búsqueda, cuando un usuario coloca una o varias palabras, se desprende una pantalla donde aparecen otros términos que pueden apoyar al usuario en su búsqueda o a decidir a utilizar otros términos.

La variable de campos limitadores, se presenta principalmente en las búsquedas avanzadas, y permite reducir la búsqueda, con el uso de diferentes campos, como son: el idioma, país, fecha de actualización de un recurso o por dominio.

En la variable de búsquedas alternativas, los cinco permiten realizar búsquedas simples, es decir utilizando el cuadro de búsqueda del buscador, asimismo permite realizar búsquedas avanzadas y también combinando operadores e índices temáticos.

Finalmente en la variable de opciones avanzadas, los cinco dan la oportunidad de buscar diferentes recursos, como son imágenes, sonido, texto, documentos en el Messenger, mapas, video, etc.

Con respecto a los metabuscadores, se realizó también una revisión, de las principales características de búsqueda de información sugeridas por Levene, obteniéndose los siguientes resultados, que se presentan en la tabla #10:

Tabla #3.10 características de búsqueda en los metabuscadores,

Metabuscador	Lenguaje de interrogación	Refinar búsqueda	Campos limitadores	Búsquedas alternativas	Opciones avanzadas
WebCrawler	Si	Si	Si	Si	Si
Dogpile	Si	Si	Si	Si	Si
Info.com	Si	Si	Si	Si	Si
Zapmeta	Si	Si	Si	Si	Si
Exite	Si	Si	Si	Si	Si

Fuente: elaboración propia

Al revisar las características de búsqueda de cada uno de los metabuscadores seleccionados, se encontró lo siguiente:

Como los metabuscadores buscan en los principales motores de búsqueda, permiten el uso del lenguaje de interrogación, todos permiten el uso en la búsqueda de diferentes operadores, entre los principales se encuentran los lógicos AND, OR y NOT.

Los buscadores en los que pueden buscar información cada uno de los metabuscadores.

- **Metacrawler:** Open Text, Lycos, Web Crawler, Hobbot, InfoSeek, Excite, Deja News, Inktomi, Yahoo y Galaxy.
- **Dogpile:** Actualmente muestra resultados de Google, Yandex y Yahoo!
- **Info.com:** Google, Yahoo!, Bing.com, Ask, LookSmart, About y Open Directory.
- **Zapmeta:** Yahoo, Microsoft, Gigablast, Altavista, Entireweb etc...
- **Exite:** FAST, About, Ask Jeeves, Inktomi, Overture y muchos otros.

Como se puede apreciar, en líneas anteriores, los motores de búsqueda y metabuscadores seleccionados, cumplen cabalmente con las características de búsqueda que algunos autores mencionan que debe de contar un buscador, para recuperar eficientemente la información que se encuentra en la Web.

Ahora, bien, para comprobar lo antes mencionado, es necesario realizar una búsqueda de información real, para cubrir este aspecto, por lo tanto se utilizó

la estrategia anteriormente estructurada, la cual se someterá a una búsqueda utilizando las características de búsqueda de los buscadores y metabuscadores, ya antes presentada.

3.6.2 Resultados de la recuperación de información en Google

Búsquedas en español

Búsqueda libre

A continuación se presentan los resultados de la búsqueda de información mencionada anteriormente, en el buscador Google.

Es importante aclarar que esta fue una búsqueda libre, se introdujo la frase completa en español **relaciones diplomáticas entre estados unidos cuba** y se encontró que el buscador entregó una gran cantidad de resultados, cerca de 520,000, como se puede apreciar en la imagen #3.1 búsqueda en Google.

Imagen #3.1 Búsqueda libre en español en Google



Pero al tratar de revisar página por página, para comprobar si la mayoría de los resultados que ofreció el buscador tenían que ver con los términos que se introdujeron y se encontró que en la página 34 de resultados solamente

aparecieron 331 resultados de los 520,000 iniciales, esto se puede apreciar en la siguiente imagen #3.2 Búsqueda libre en español en Google.

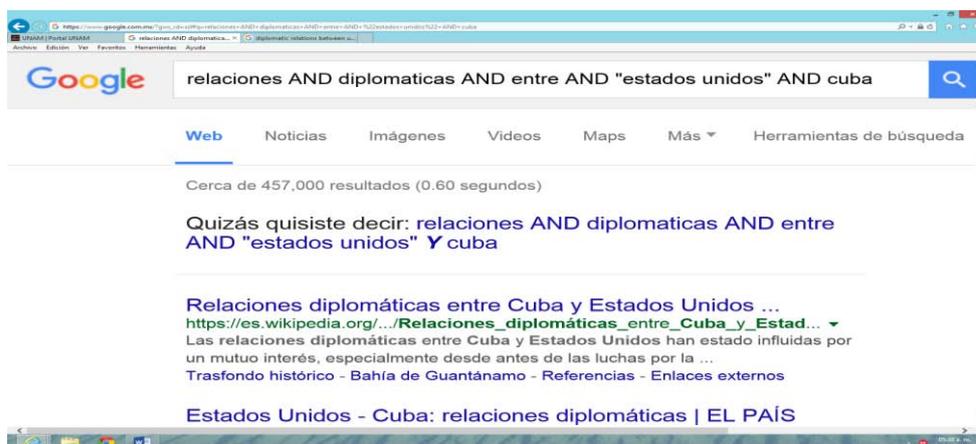
Imagen #3.2 búsqueda libre en español en Google



Búsqueda con operadores lógicos boléanos en español en Google

Al realizar la búsqueda en español, utilizando los operadores lógicos boléanos, los resultados fueron los siguientes, cerca de 457,000 documentos recuperados, tal como se muestra en la imagen #3.3 búsqueda con operadores lógicos boléanos

Imagen #3.3 búsqueda con operadores lógicos boléanos en español en Google



Si comparamos la primera búsqueda en español con esta última, podemos ver que en la búsqueda libre los resultados que ofreció el buscador fue de 520,000

y utilizando los operadores lógicos y las comillas disminuyo el número de resultados a cerca 457, 000 resultados, es decir 63, 000 resultados menos, esto nos indica que cuando se utilizan los operadores lógicos booleanos, en especial el operador lógico AND, se recuperan menos resultados.

Búsqueda por campos

Finalmente al combinar operadores lógicos, comillas y realizar la búsqueda en un campo específico como es el título, los resultados que ofreció el buscador, fueron cerca de 14 resultados, que al revisarlos se encontró que la mayoría contaban con todos los términos en el título del documento, como se puede apreciar en la imagen #3.4.

Imagen #3.4 búsqueda por campos en español en Google



Otro aspecto interesante es que al tratar de ver los 14 resultados solamente el buscador ofreció 13 resultados.

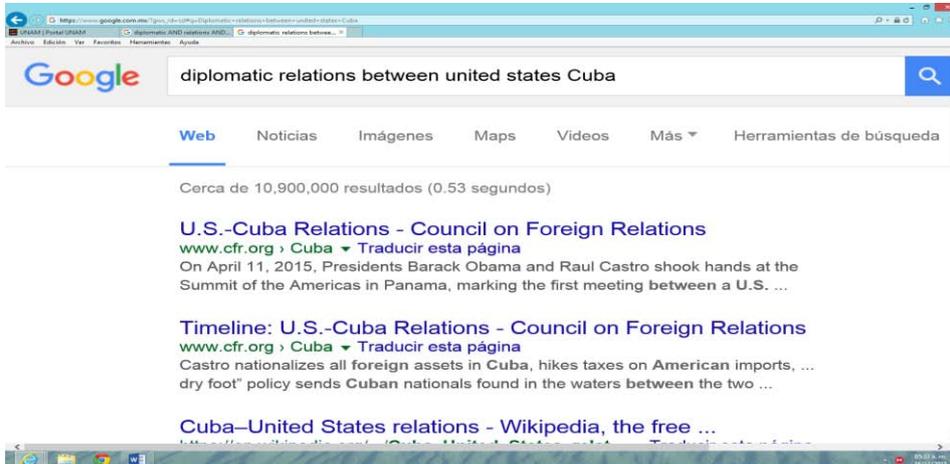
Búsquedas en inglés

Búsqueda libre

Con respecto a la búsqueda de información en inglés, diplomatic relations between united states cuba, se encontró que el buscador proporcionó una cantidad impresionante de resultados, cerca de 10,900,000, lo cual resulta

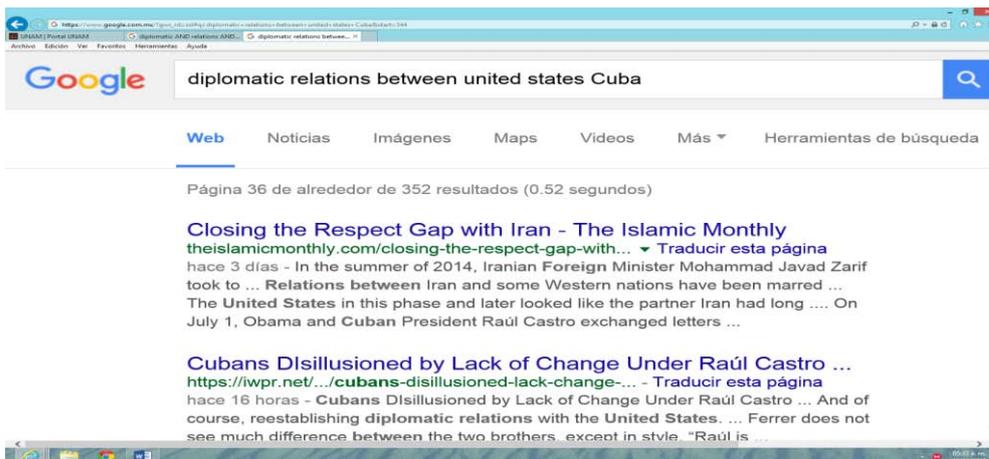
lógico si tomamos en consideración que este tema es muy relevante en este momento en los Estados Unidos, así lo demuestra la imagen #3.5.

Imagen #3.5 búsqueda libre en inglés en Google



Por otro lado, al realizar el mismo procedimiento que en la búsqueda en español, sucedió algo parecido, al revisar en la página 36 de resultados, solamente aparecieron 352 resultados de cerca de 10,900,000, como se puede apreciar en la imagen #3.6.

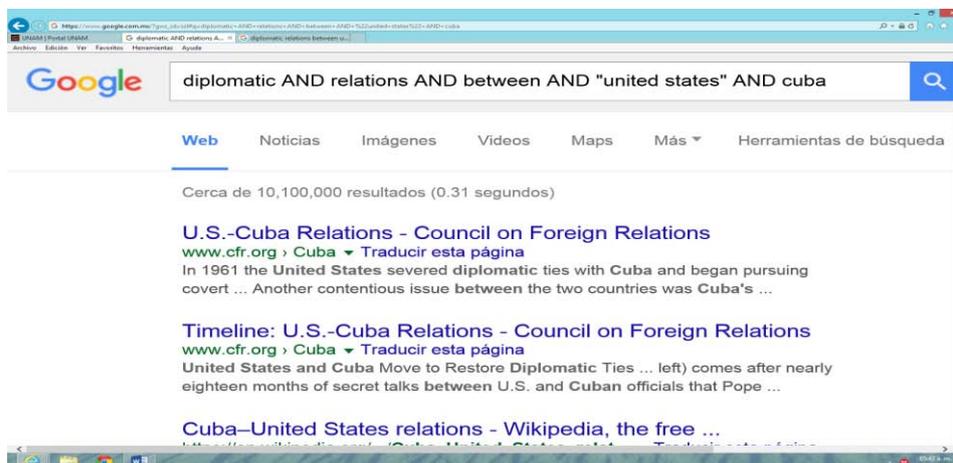
Imagen #3.6 resultados visibles en búsqueda libre en inglés en Google



Búsqueda con operadores lógicos booleanos

Con respecto a la búsqueda en idioma inglés, utilizando también los operadores lógicos booleanos, El resultado fue sorprendente, porque se encontraron cerca de 10,100,000 resultados, 800, 000 resultados menos que la primera búsqueda en inglés que fue de 10,900,000 resultados, algo diferente a la búsqueda en español que solamente elimino 63, 000 resultados, cuando se utilizaron operadores lógicos y comillas, como se puede apreciar en la imagen #3.7.

Imagen #3.7 búsqueda operadores lógicos booleanos en inglés en Google



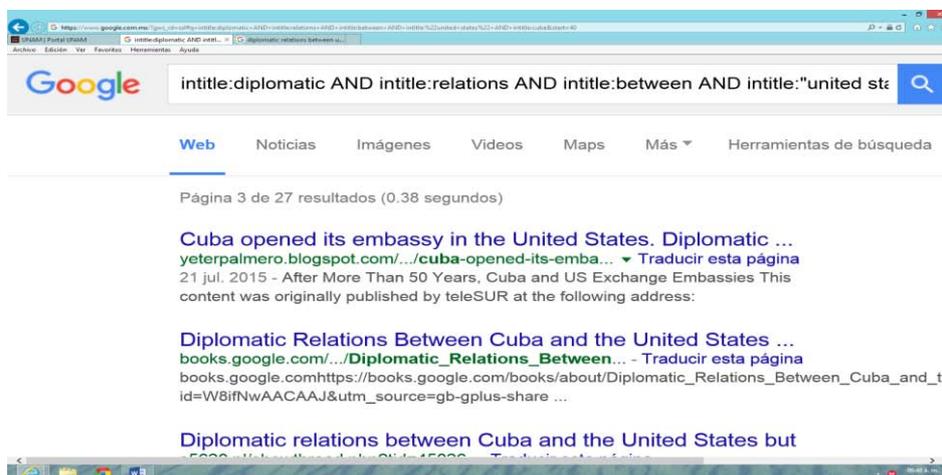
Búsqueda por campos

Finalmente se realizó la misma estrategia en inglés y los resultados que ofreció el buscador, fueron cerca de 63 resultados, que al revisarlos solamente aparecieron 27 y todos contaban con los términos en el título del documento, que es lo que se le había solicitado al buscador, como se puede apreciar en las imágenes # 3.8 y #3.9.

Imagen #3.8 búsqueda por campos en inglés en Google



Imagen #3.9 resultados visibles en búsqueda por campos en inglés en Google



3.6.3 Resultados de la recuperación de información en Yahoo

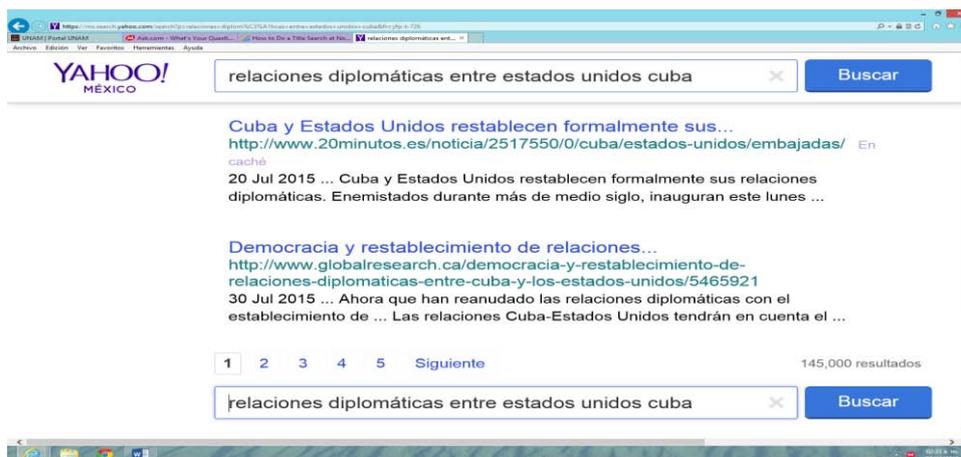
Para el análisis de la recuperación de información en el motor Yahoo, se utilizó la misma búsqueda y estrategia, a continuación se presentan los resultados.

Búsquedas en español

Búsqueda libre

Se introdujo la frase completa y el buscador entregó una gran cantidad de resultados, cerca de 145,000, como se puede apreciar en la imagen #3.10.

Imagen #3.10 Búsqueda libre en español en Yahoo



Al revisar página por página, para comprobar si la mayoría de los resultados que aparecieron en el buscador tenían que ver con los términos que se introdujeron, nos percatamos que en la página de resultados número 19 solamente aparecieron 183 resultados de los 145,000 iniciales, como se puede ver en la imagen #3.11.

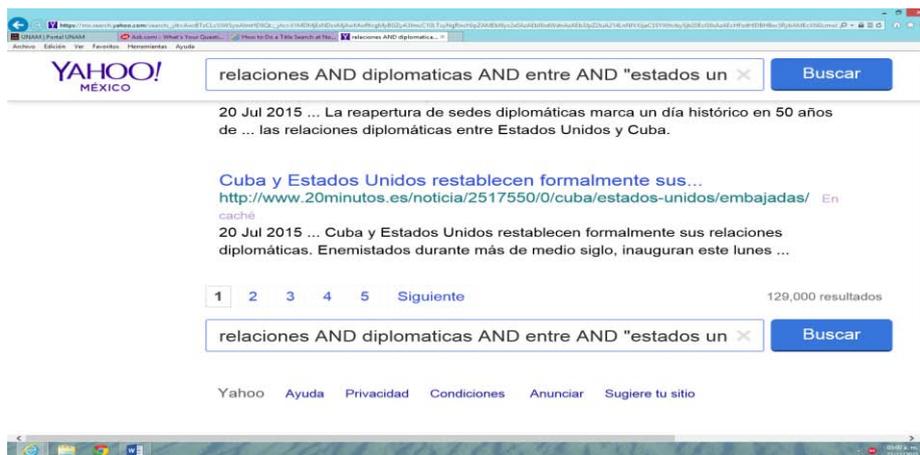
Imagen #3.11 Resultados visibles de búsqueda libre en español en Yahoo



Búsqueda con operadores lógicos boléanos

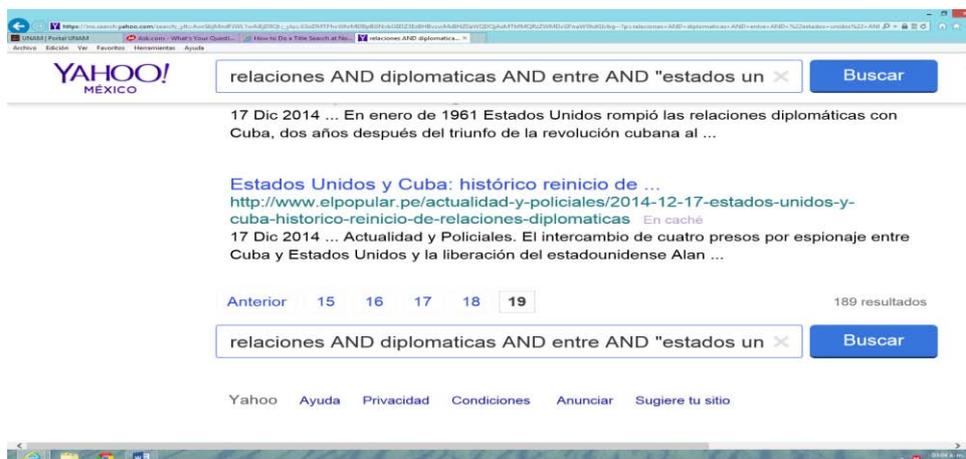
Al realizar la búsqueda en español, utilizando los operadores lógicos boléanos, los resultados fueron los siguientes, cerca de 129,000 recuperados, tal como se muestra en la imagen #3.12.

Imagen #3.12 Búsqueda con operadores booleanos en español en Yahoo



Al revisar los resultados de la búsqueda, en la página 19 encontramos que solo existen 189, esto lo podemos apreciar en la imagen #3.13.

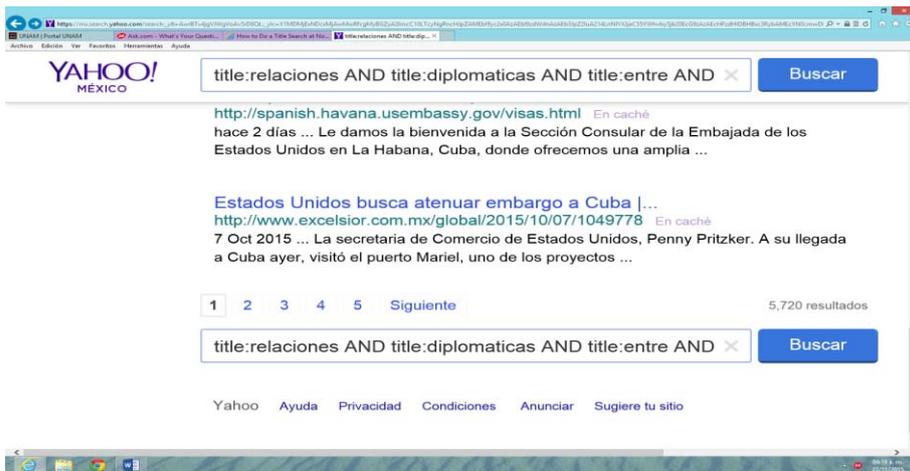
Imagen #3.13 Resultados visibles de búsqueda con operadores booleanos



Búsqueda por campos

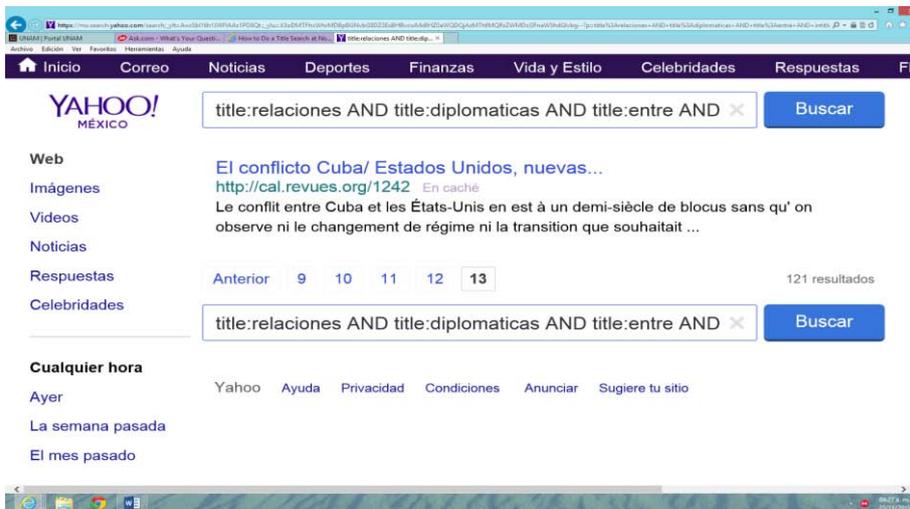
Al combinar operadores lógicos, comillas y realizar la búsqueda en un campo específico como es el título, los resultados que ofreció el buscador, fueron cerca de 5,720, como se puede apreciar en la imagen #14.

Imagen #3.14 Búsqueda por campos en español en Yahoo



Al revisar los resultados de la búsqueda, en la página 13 solo existen 121 enlaces, como se aprecia en la imagen #3.15.

Imagen #3.15 Resultados visibles búsqueda por campos en español en Yahoo



Búsquedas en inglés

Búsqueda libre

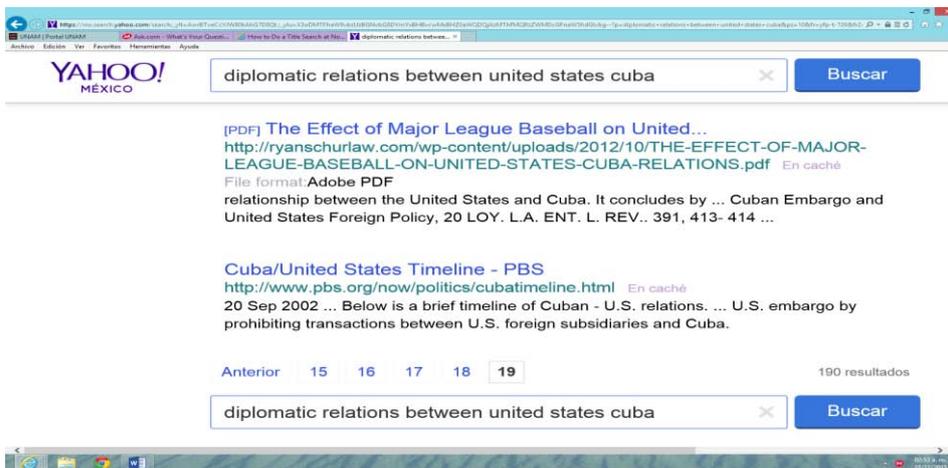
En la búsqueda libre en inglés la recuperación de documentos fue cerca de 7,270,000, así se puede observar en la imagen #3.16

Imagen #3.16 Búsqueda libre en inglés en Yahoo.



Al revisar los resultados de la búsqueda, en la página 19 encontramos que solo existen 190 resultados, esto se puede apreciar en la imagen #17.

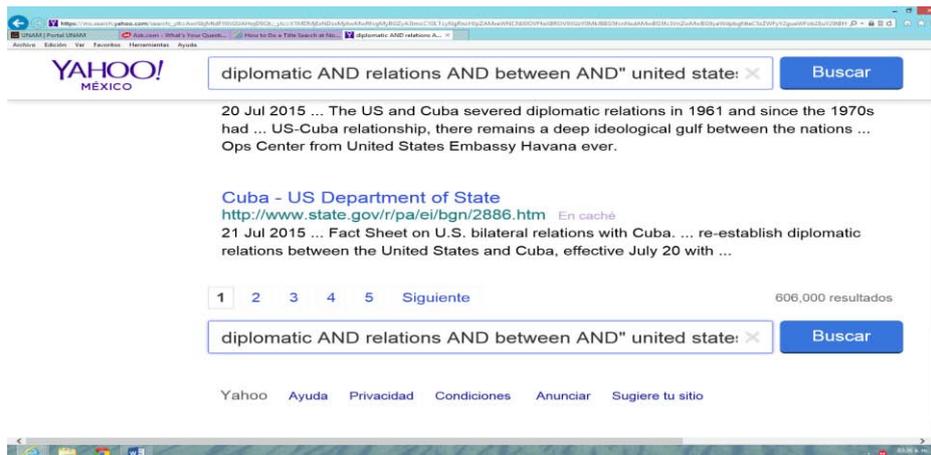
Imagen #3.17 Resultados visibles búsqueda libre en inglés en Yahoo



Búsqueda con operadores lógicos boléanos

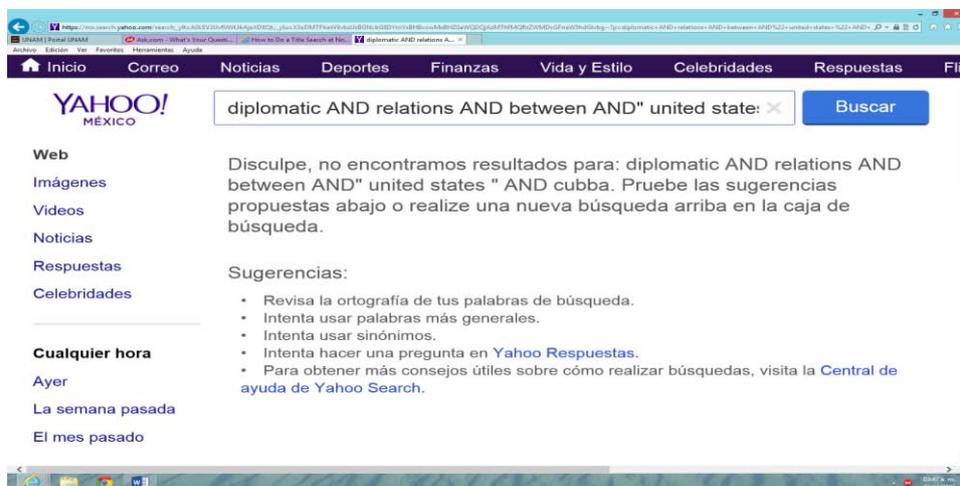
Con respecto a la búsqueda en idioma inglés, utilizando también los operadores lógicos boléanos, los resultados fueron los siguientes, cerca de 606,000 recuperados, tal como se muestra en la imagen #3.18.

Imagen #3,18 Búsqueda con operadores lógicos en inglés en Yahoo



Al revisar los resultados de la búsqueda, en la página 19 se encontró que solo existen 200, asimismo en la página 20 aparece la leyenda no se encontraron resultados para nuestra búsqueda, como se aprecia en la imagen #3.19 .

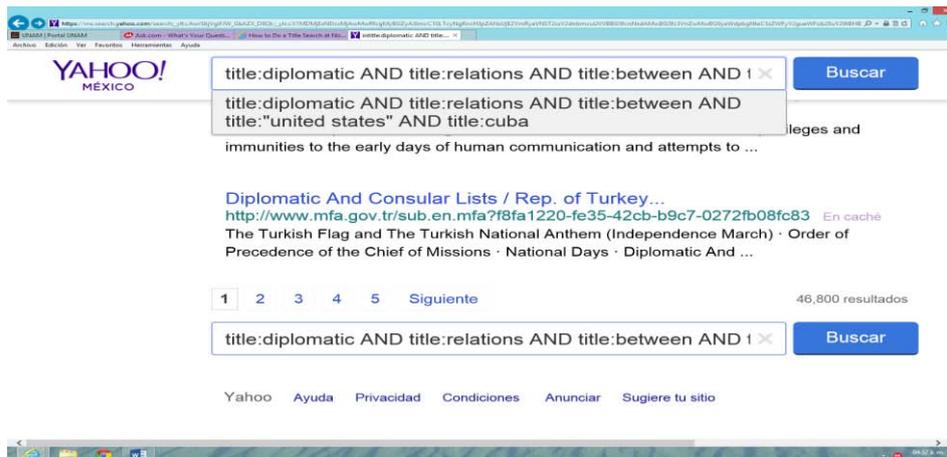
Imagen #3.19 Resultados visibles búsqueda con operadores en inglés en Yahoo



Búsqueda por campos

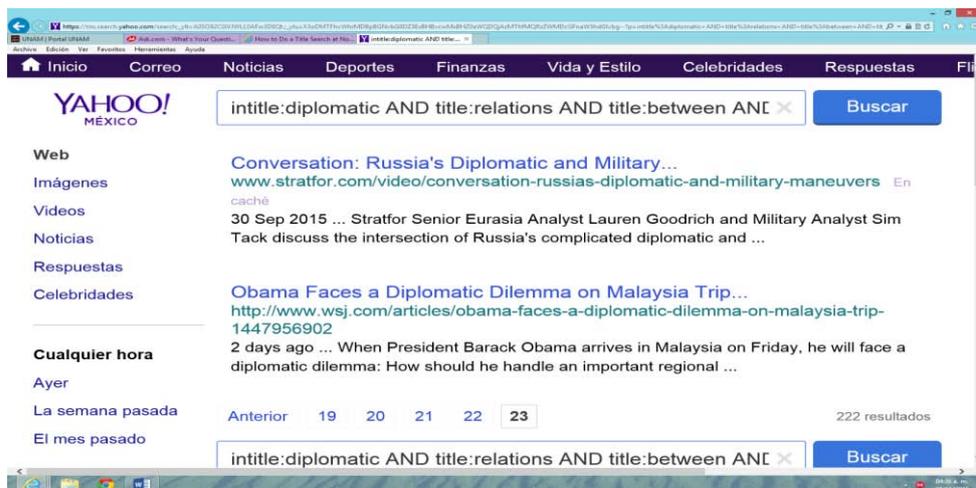
Finalmente se realizó la misma estrategia en inglés y los resultados que ofreció el buscador, fueron cerca de 46,800, como se puede apreciar en la imagen #3.20.

Imagen #3.20 Búsqueda por campos en inglés en Yahoo



Al revisar los resultados de la búsqueda, en la página 23 encontramos que solo existen 222 enlaces. Así lo muestra la imagen #3.21.

Imagen #3.21 Resultados visibles búsqueda por campos en inglés en Yahoo



3.6.4 Resultados de la recuperación de información en Bing

Al igual que en los anteriores motores de búsqueda, se utilizó para Bing la misma búsqueda y estrategia, a continuación se presentan los resultados.

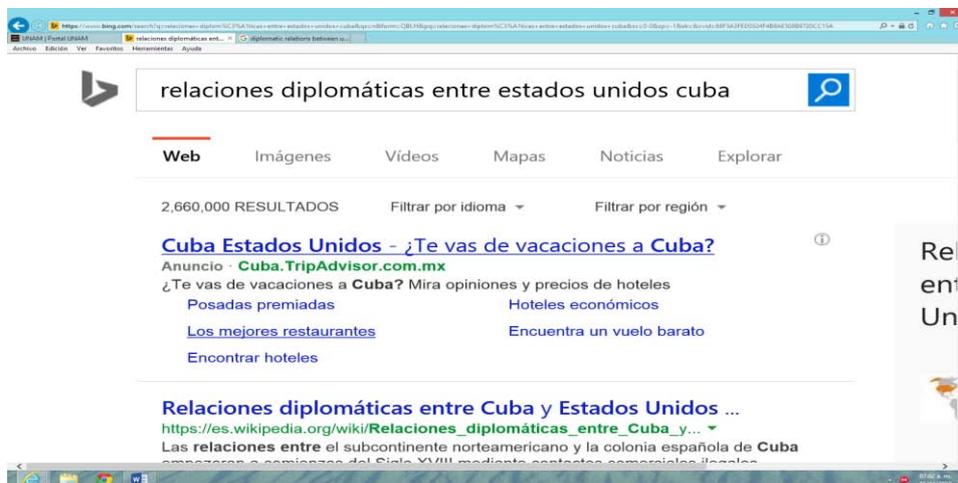
Búsquedas en español

Búsqueda libre

La primera búsqueda fue libre, en ella se utilizaron todos los términos en español.

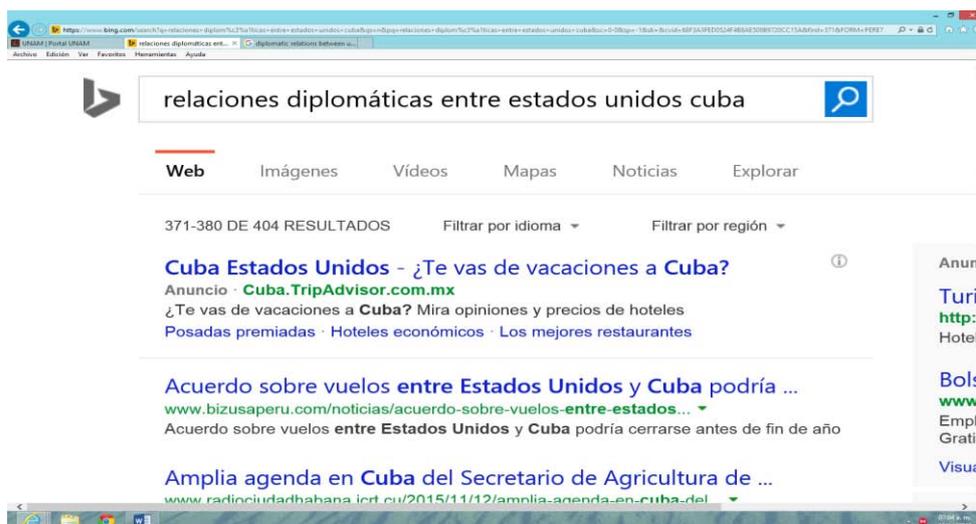
Se introdujo la frase completa en español y el buscador entregó una gran cantidad de resultados, cerca de 2, 660,000, como se puede apreciar en la imagen #3.22.

Imagen #3.22 Búsqueda libre en español en Bing



Al revisar página por página, para comprobar si la mayoría de los resultados que aparecieron en el buscador tenían que ver con los términos que se introdujeron, en la página de resultados número 37 solamente aparecieron 380 resultados de los 2,660,000 iniciales, esto se puede apreciar en la imagen #3.23.

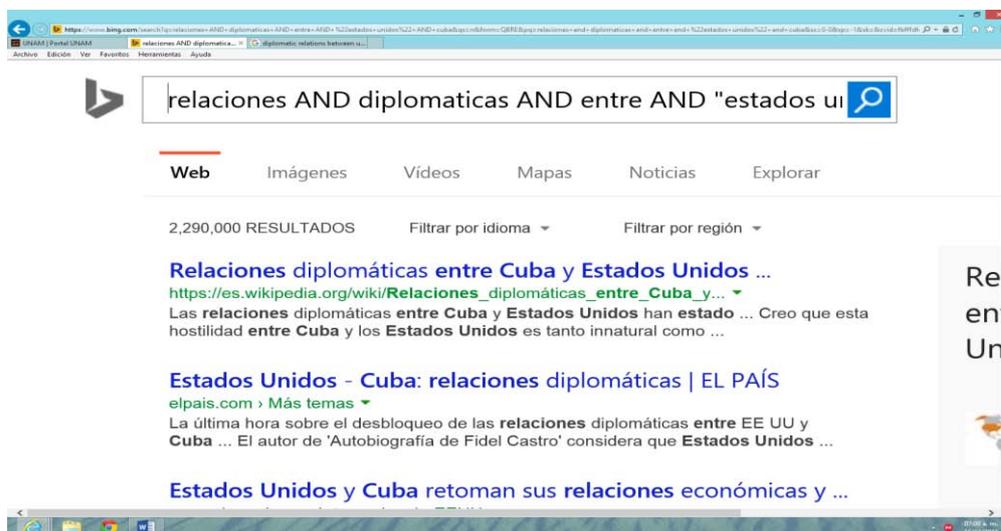
Imagen #3.23 Resultados visibles búsqueda libre en español en Bing



Búsqueda con operadores lógicos boléanos

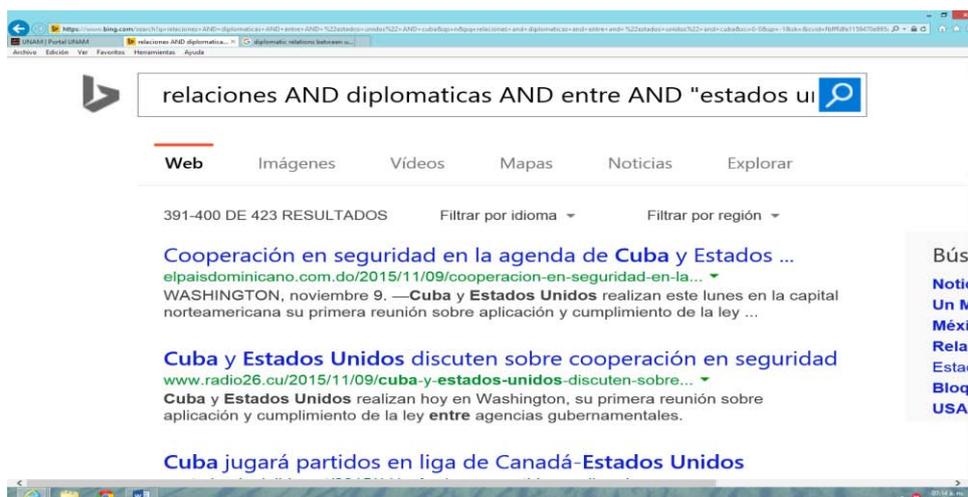
Al realizar la búsqueda en español, utilizando los operadores lógicos boléanos, los resultados fueron los siguientes, cerca de 2,290,000 resultados recuperados, tal como se muestra en la imagen #3.24.

Imagen #3.24 Búsqueda con operadores lógicos en español en Bing



Al revisar los resultados que se obtuvieron, en la página 39 solo existen 423 resultados de cerca de 2,290,000, así lo demuestra la imagen #3.25.

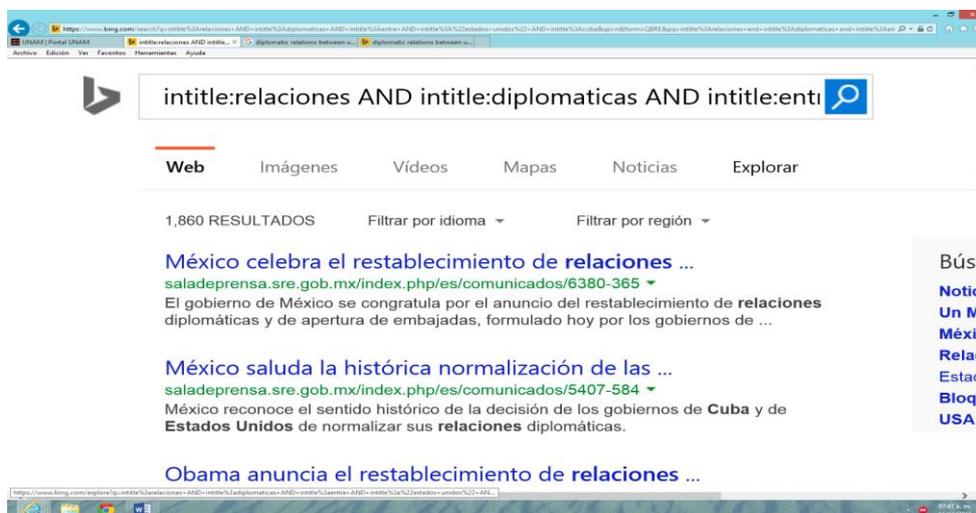
Imagen #3.25 Resultados visibles búsqueda con operadores en español en Bing



Búsqueda por campos

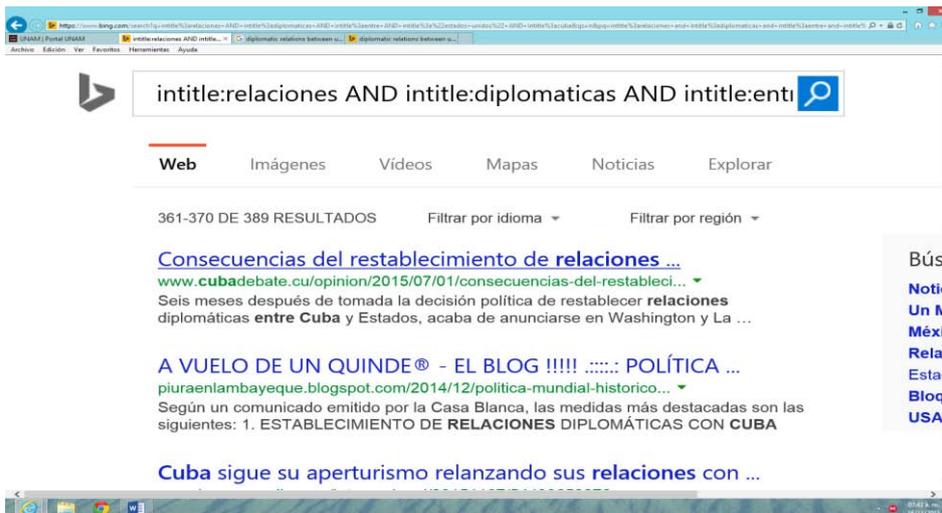
Al combinar operadores lógicos, comillas y realizar la búsqueda en el campo de título, los resultados en el buscador, fueron cerca de 1,860 resultados. Como se puede apreciar en la imagen #3.26.

Imagen #3.26 Búsqueda por campos en español en Bing



Al revisar los resultados se encontró que el buscador solo presenta 389 y la mayoría contaban con todos los términos en el título del documento, como se muestra en la imagen #3.27.

Imagen #3.27 Resultados visibles búsqueda por campos en español en Bing

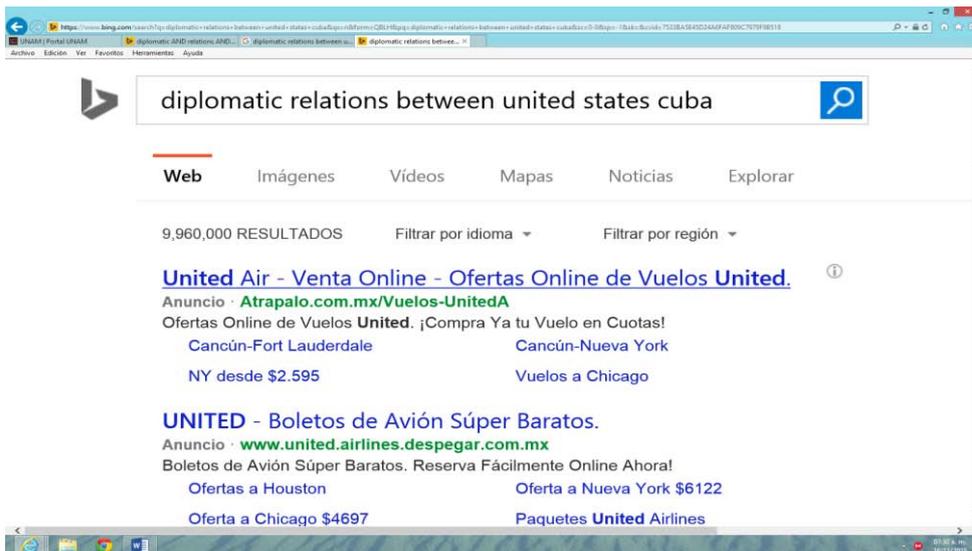


Búsquedas en inglés

Búsqueda libre

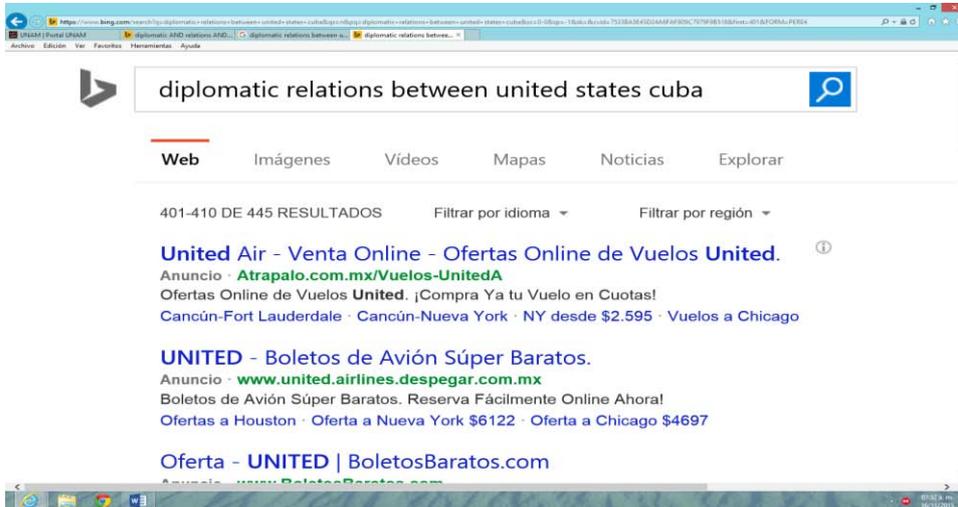
En la búsqueda libre en inglés, la recuperación de documentos fue cerca de 9,960,000, así lo vemos en la imagen #3.28.

Imagen #3.28 Búsqueda libre en inglés en Bing



Al revisar los resultados de la búsqueda, en la página 40 solo existen 445 resultados de cerca de 9,960,000 esto se puede apreciar en la imagen #3.29.

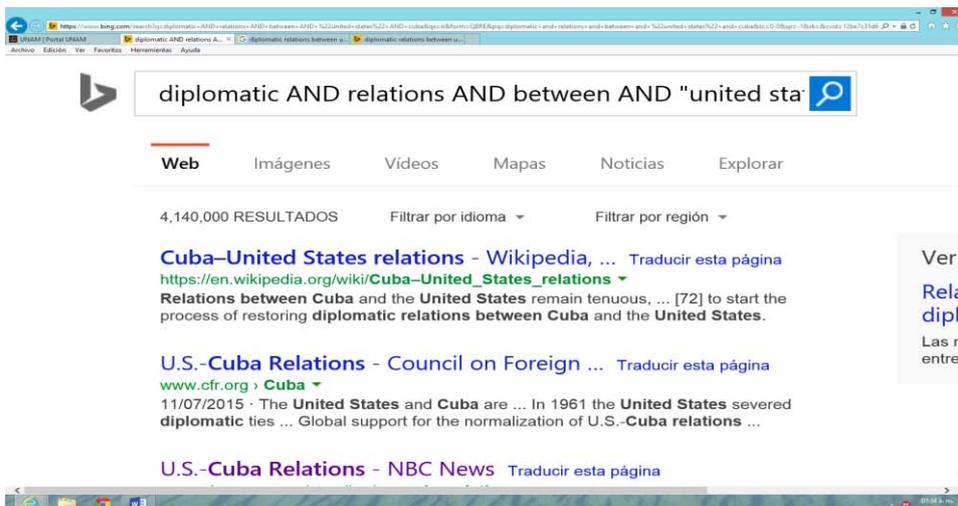
Imagen #3.29 Resultados visibles búsqueda libre inglés en Bing



Búsqueda con operadores lógicos boléanos

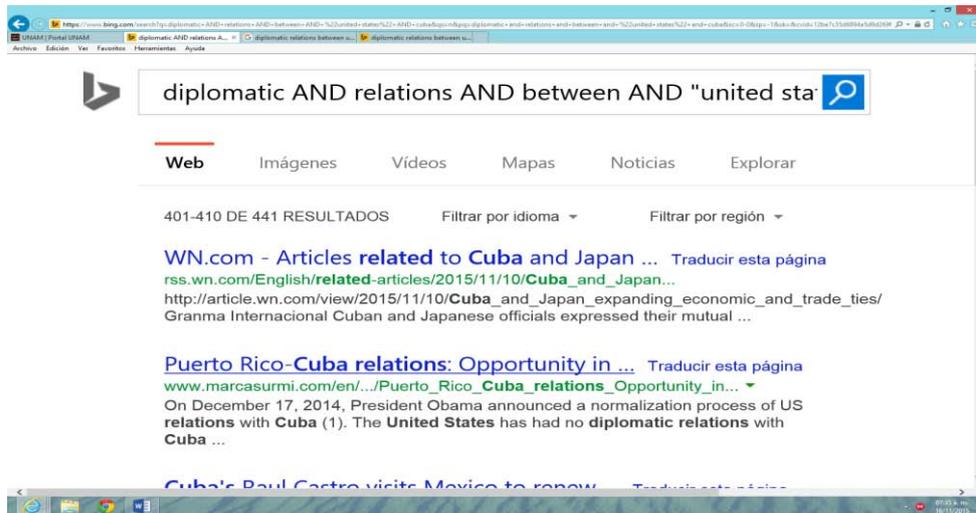
Con respecto a la búsqueda en idioma inglés, utilizando también los operadores lógicos boléanos, se obtuvieron cerca de 4,140,000 resultados, así lo muestra la imagen #3.30.

Imagen #3.30 Búsqueda con operadores en inglés en Bing



Por otro lado, al revisar los resultados, en la página 40 solo aparecen 441 resultados, como lo demuestra la imagen #3.31.

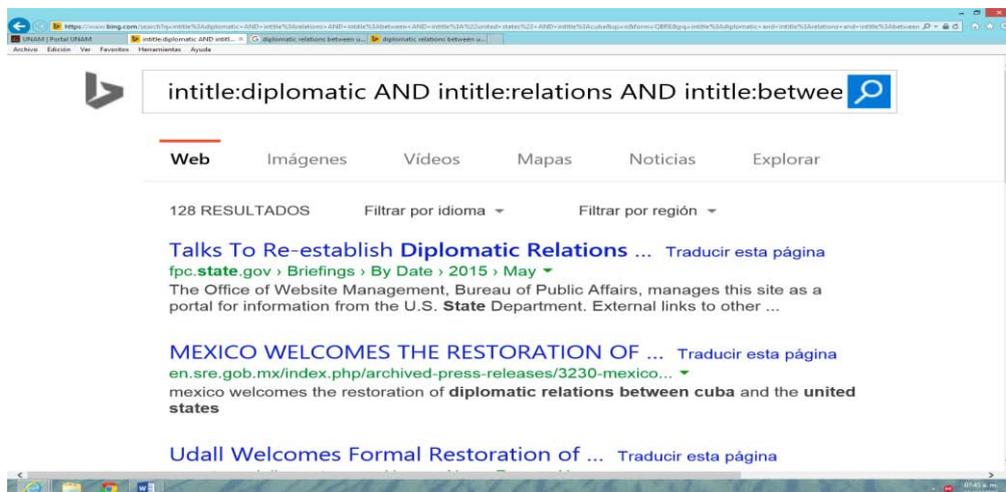
Imagen #3.31 Resultados visibles búsqueda con operadores en inglés en Bing



Búsqueda por campos

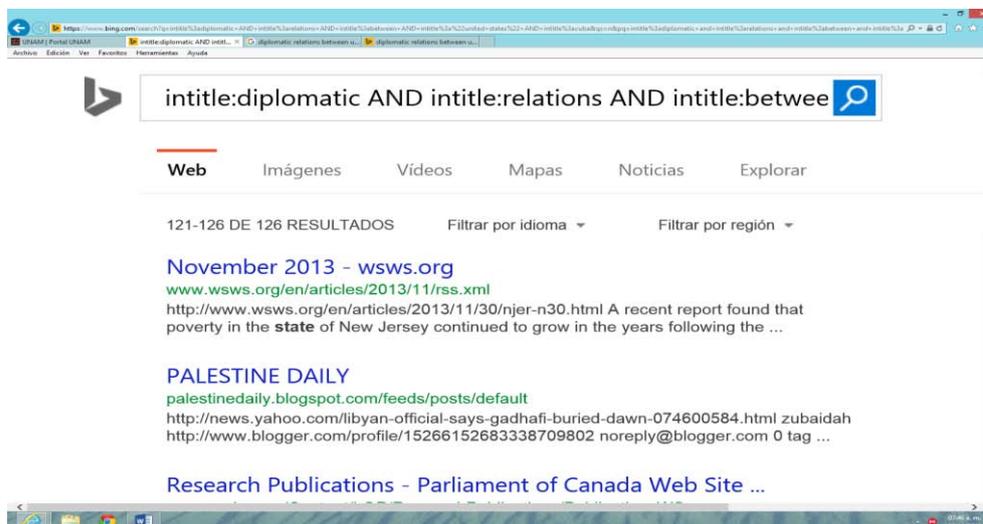
Finalmente se realizó la estrategia en inglés, el buscador recuperó cerca de 128 resultados, algo que llama mucho la atención es que el buscador Bing recupero más documentos en español que en inglés, esto es raro por tratarse de un tema que tiene gran relevancia en Estados Unidos, en la imagen #3.32 se puede observar.

Imagen #3.32 Búsqueda por campos en inglés en Bing



De la misma forma al revisarlos, se encontró que solamente aparecieron 126 resultados, menos que los que se pueden revisar en español. Así lo demuestra la imagen #3.33.

Imagen #3.33 Resultados visibles búsqueda por campos en inglés en Bing



3.6.5 Resultados de la recuperación de información en Ask

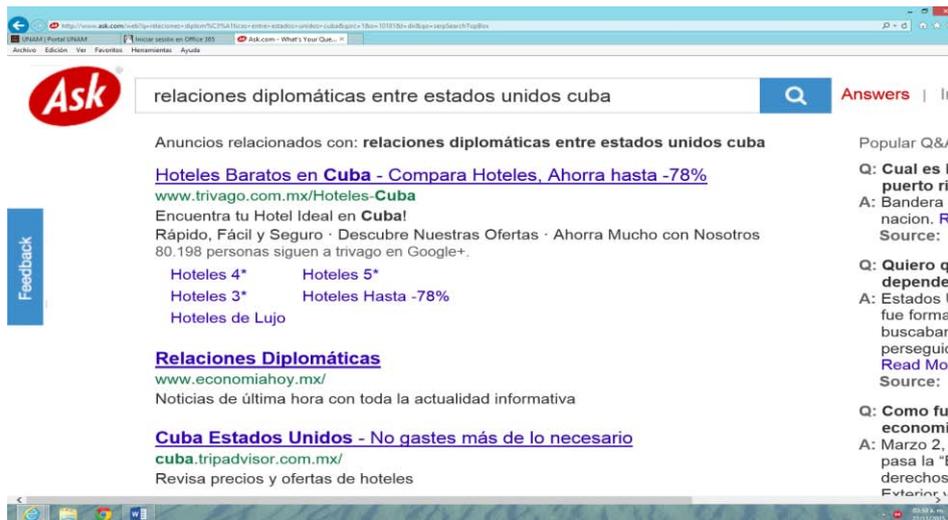
Al igual que en los otros motores de búsqueda, se utilizó la misma búsqueda y estrategia, a continuación se presentan los resultados.

Búsquedas en español

Búsqueda libre

Se introdujo la frase completa y se encontró que el buscador entregó muy pocos resultados, en comparación con los otros dos motores, solamente presenta un aproximado de 100 de resultados, todos ellos relacionados con los términos de la búsqueda, además en ningún lugar de la página de resultados presenta el número de documentos entregados, como se puede apreciar en la imagen #3.34.

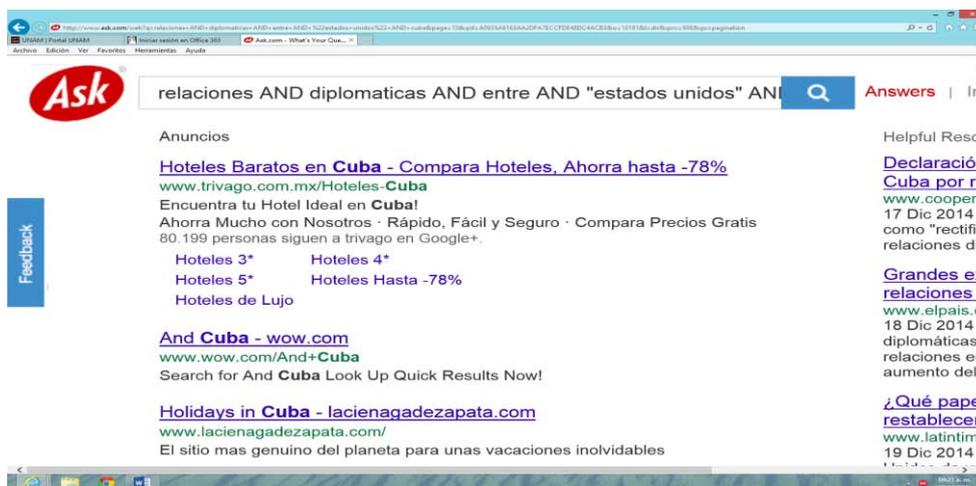
Imagen #3.34 Búsqueda libre en español en Ask



Búsqueda con operadores lógicos boléanos

Al realizar la búsqueda en español, utilizando los operadores lógicos boléanos, los resultados fueron los siguientes, cerca de 100 resultados recuperados, aunque aparecían más páginas de resultados a partir de la página 10 se repetían los resultados, tal como se muestra en la imagen #3.35.

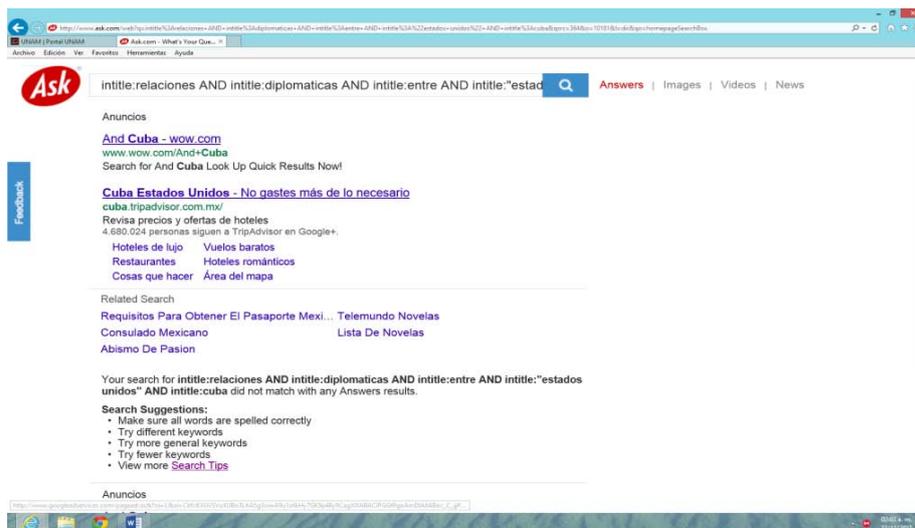
Imagen #3.35 Búsqueda con operadores lógicos en español en Ask



Búsqueda por campos

Al combinar operadores lógicos, comillas y realizar la búsqueda en el campo de título, no se obtuvieron resultados en el buscador, como se puede apreciar en la imagen #3.36.

Imagen #3.36 Búsqueda por campos en español en Ask



Búsquedas en inglés

Búsqueda libre

En la búsqueda libre en inglés, la recuperación de documentos fue cerca 100 resultados, además el motor de búsqueda presenta los resultados de manera separada, primero presenta resultados de hoteles, después noticias, un recuadro de búsquedas relacionadas y finalmente los resultados Web. Por otro lado todos los resultados tienen que ver con la búsqueda que se realizó, así se visualiza en la imagen #3.37.

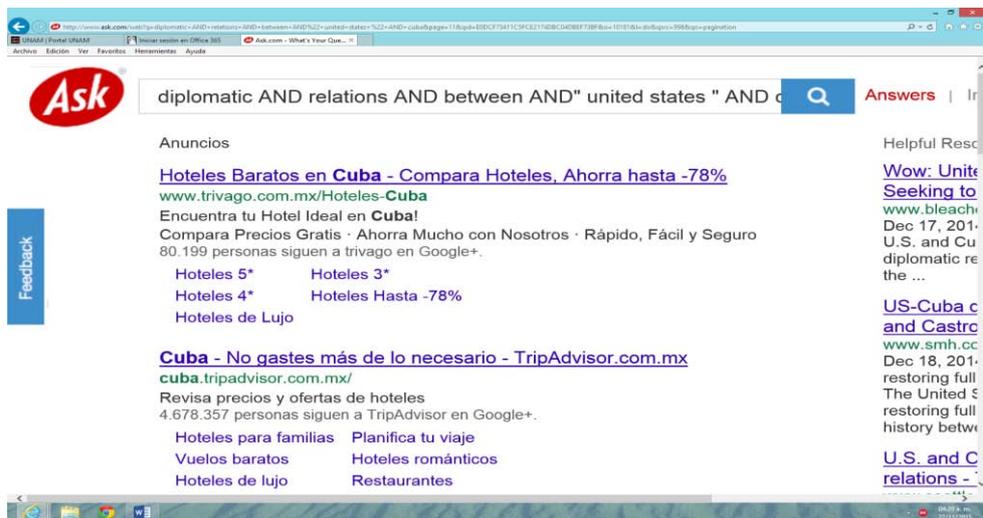
Imagen #3.37 Búsqueda libre en inglés en Ask



Búsqueda con operadores lógicos boléanos

Con respecto a la búsqueda en idioma inglés, los resultados fueron, aproximadamente cerca de 100 documentos recuperados, tal como se muestra en la imagen #3.38.

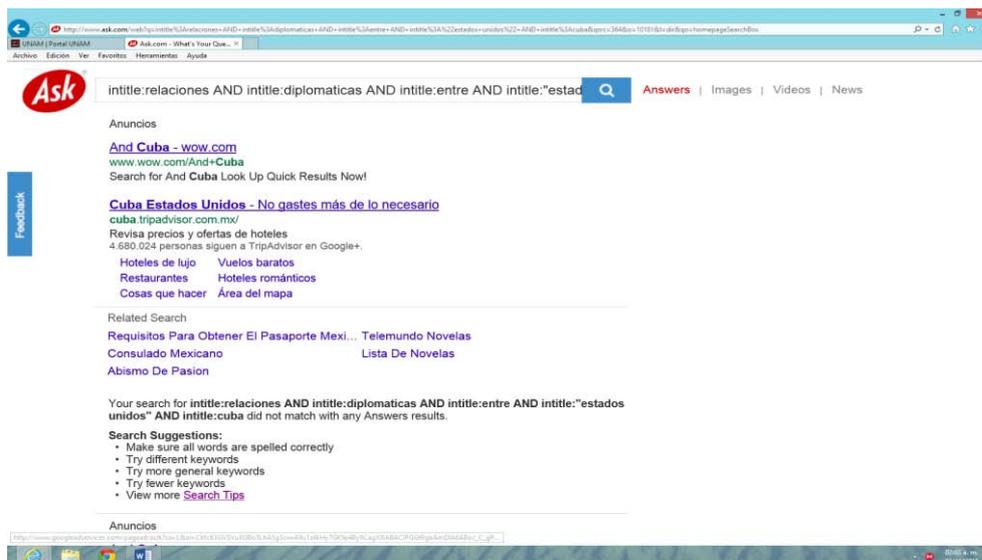
Imagen #3.38 Búsqueda con operadores lógicos en inglés en Ask



Búsqueda por campos

Al combinar operadores lógicos, comillas y realizar la búsqueda en el campo de título, no se obtuvieron resultados en el buscador, como se puede apreciar en la siguiente imagen #3.39.

Imagen #3.39 Búsqueda por campos en inglés en Ask



3.6.6 Resultados de la recuperación de información en AOL

Se utilizó la misma búsqueda y estrategia, a continuación se presentan los resultados.

Búsquedas en español

Búsqueda libre

Se introdujo la frase completa en español y el buscador entregó una gran cantidad de resultados, cerca de 608,000, como se puede apreciar en la imagen #3.40.

Imagen #3.40 Búsqueda libre en español en Aol



Al revisar los resultados que aparecieron en el buscador, en la página 21 de resultados solamente aparecieron 203 resultados de los 608,000 iniciales, esto se puede apreciar en la imagen #3.41.

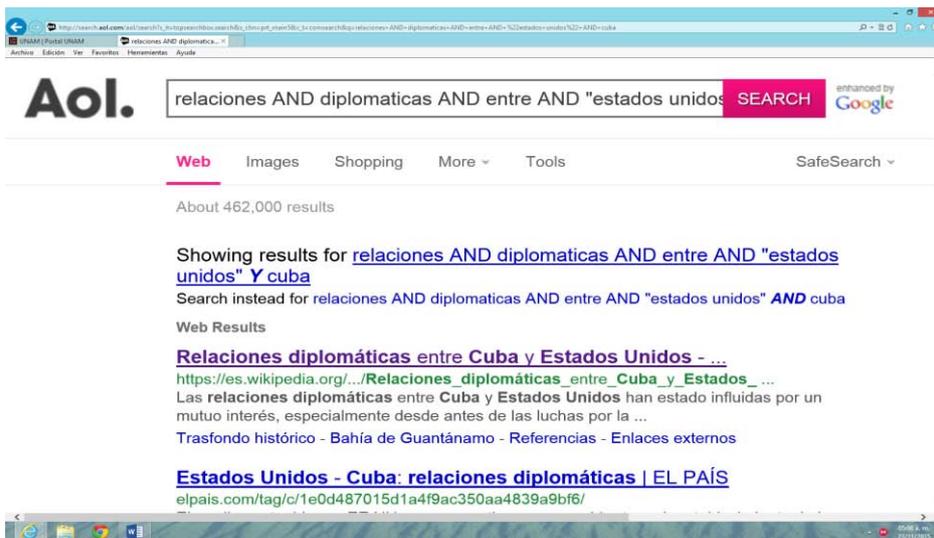
Imagen #3.41 Resultados visibles en búsqueda libre en español en Aol



Búsqueda con operadores lógicos boléanos

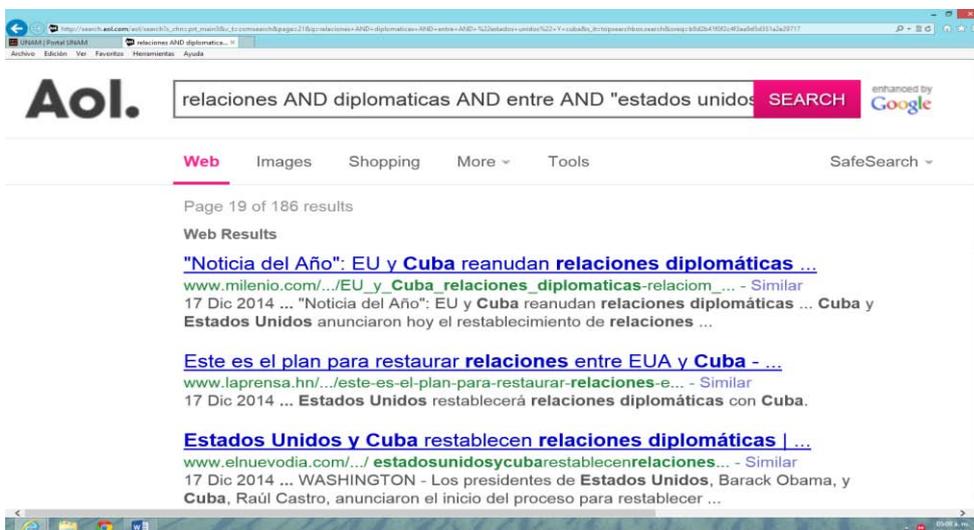
Al realizar la búsqueda en español, utilizando los operadores lógicos boléanos, los resultados fueron los siguientes, cerca de 462,000 resultados recuperados, tal como se muestra en la imagen #3.42.

Imagen #3.42 Búsqueda con operadores lógicos en inglés en Aol



Al revisar los resultados, en la página 19 se encontraron solo 186 resultados de cerca de 462,000 esto se puede apreciar en la imagen #3.43.

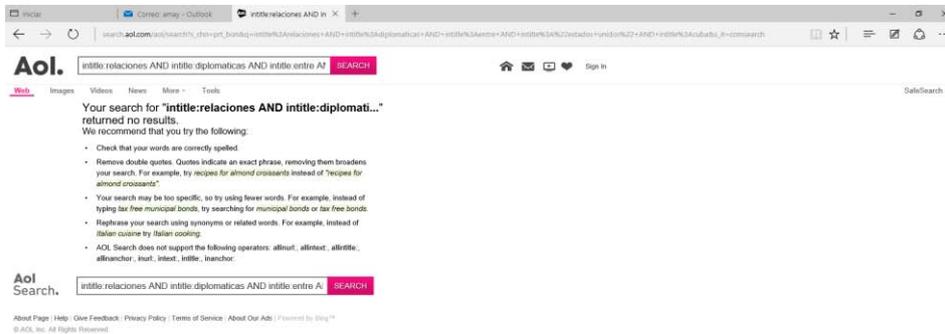
Imagen #3.43 Resultados visibles en búsqueda con operadores en español en Aol



Búsqueda por campos

Al realizar la búsqueda en español, utilizando el campo de título, no se obtuvieron resultados, tal como se muestra en la imagen #3.44.

Imagen #3.44 Búsqueda por campos en español en Aol

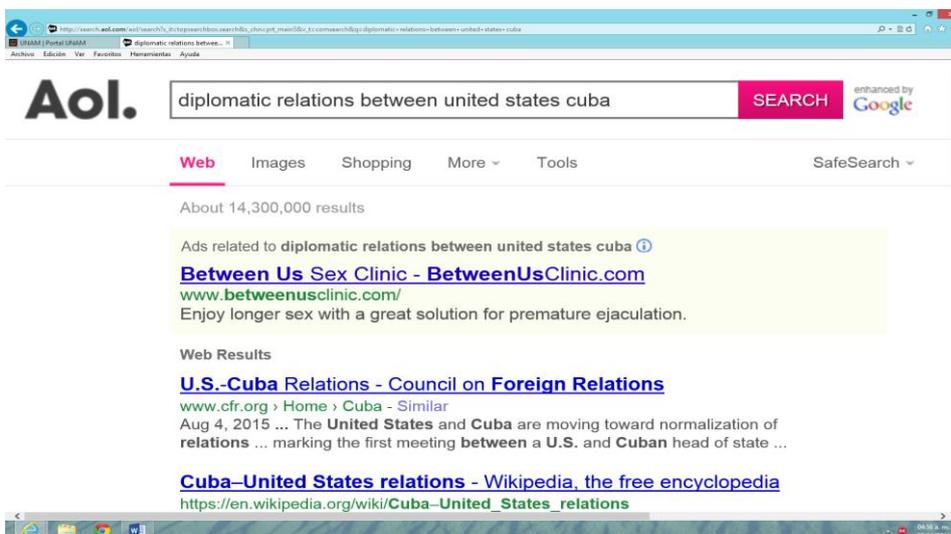


Búsquedas en inglés

Búsqueda libre

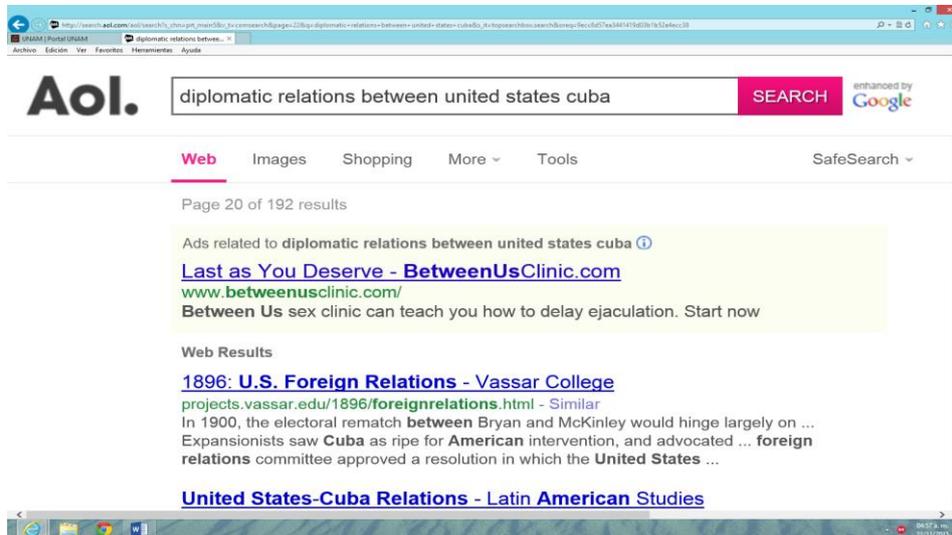
En la búsqueda libre en inglés, la recuperación de documentos fue cerca de 14,300,000, así se observa en siguiente imagen #3.45.

Imagen #3.45 Búsqueda libre en inglés en Aol



Al revisar los resultados, en la página 20 solo existen 192 resultados de cerca de 14,300,000 esto lo podemos apreciar en la imagen #3.46.

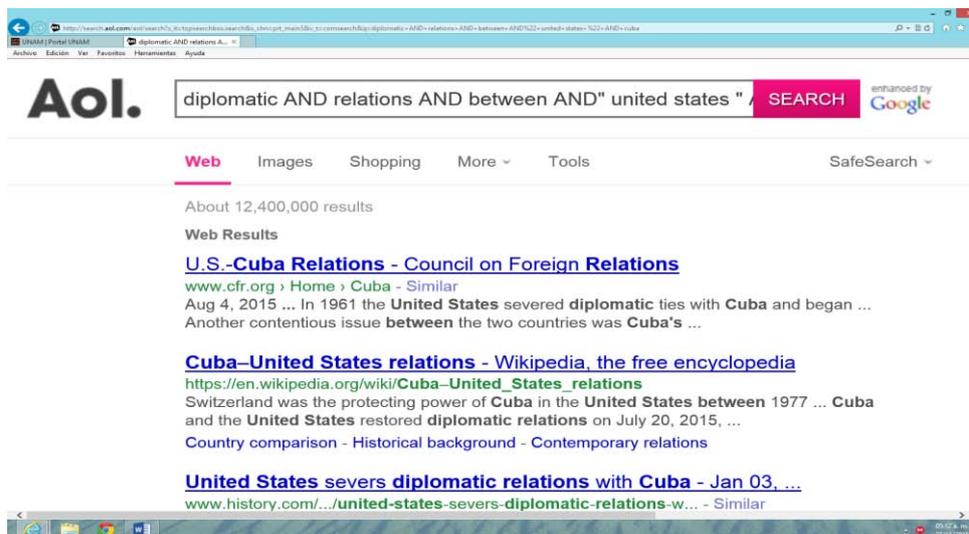
Imagen #3.46 Resultados visibles en Búsqueda libre en inglés en Aol



Búsqueda con operadores lógicos boléanos

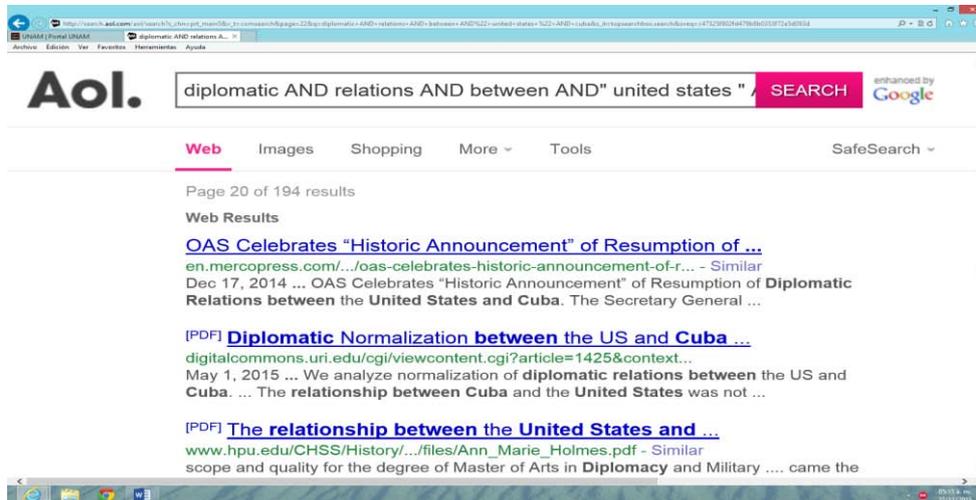
Con respecto a la búsqueda en idioma inglés, como resultado al utilizar los operadores lógicos, se obtuvieron cerca de 12,400,000 resultados, así lo demuestra la siguiente imagen #3.47.

Imagen #3.47 Búsqueda con operadores en inglés en Aol



Por otro lado, al revisar los resultados, en la página 20 solo aparecieron 194 resultados, como lo demuestra la imagen #3.48.

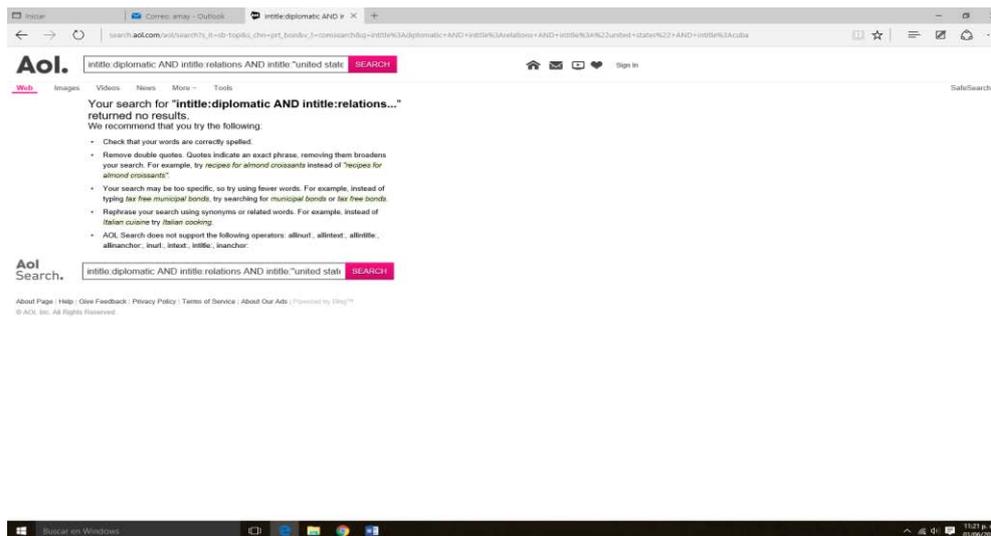
Imagen #3.48 Resultados visibles en búsqueda con operadores en inglés en Aol



Búsqueda por campos

Al realizar la búsqueda en inglés, utilizando el campo de título, no se obtuvieron resultados, tal como se muestra en la imagen #3.49.

Imagen #3.49 Búsqueda por campos en inglés en Aol



3.6.7 Resultados de la recuperación de información en WebCrawler

A continuación se presentan los resultados de la búsqueda de información en el metabuscador WebCrawler, con respecto a la búsqueda.

Búsquedas en español

Búsqueda libre

Se introdujo la frase completa en español en el metabuscador y nos entregó 73 páginas de resultados, cerca de 730, como se puede apreciar en las imágenes #3.50 y #3.51.

Imagen #3.50 Búsqueda libre en español en WebCrawler

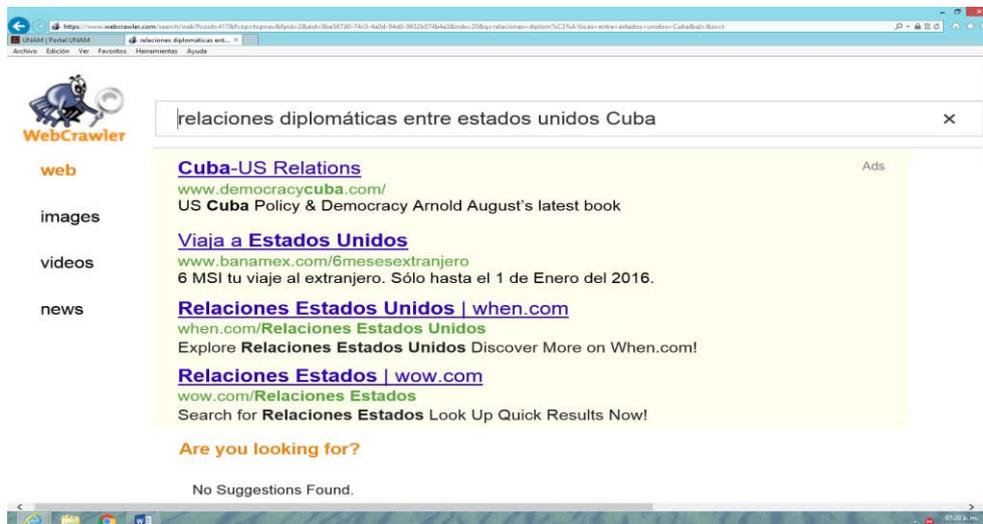


Imagen #3.51 Resultados visibles en búsqueda libre en español en WebCrawler



Búsqueda con operadores lógicos booleanos

Al realizar la búsqueda en español, utilizando los operadores lógicos booleanos, los resultados fueron los siguientes, el metabuscador presentó 67 páginas con cerca de 670 documentos recuperados, tal como se muestra en las imágenes #3.52 y #3.53.

Imagen #3.52 Búsqueda con operadores lógicos en español en WebCrawler

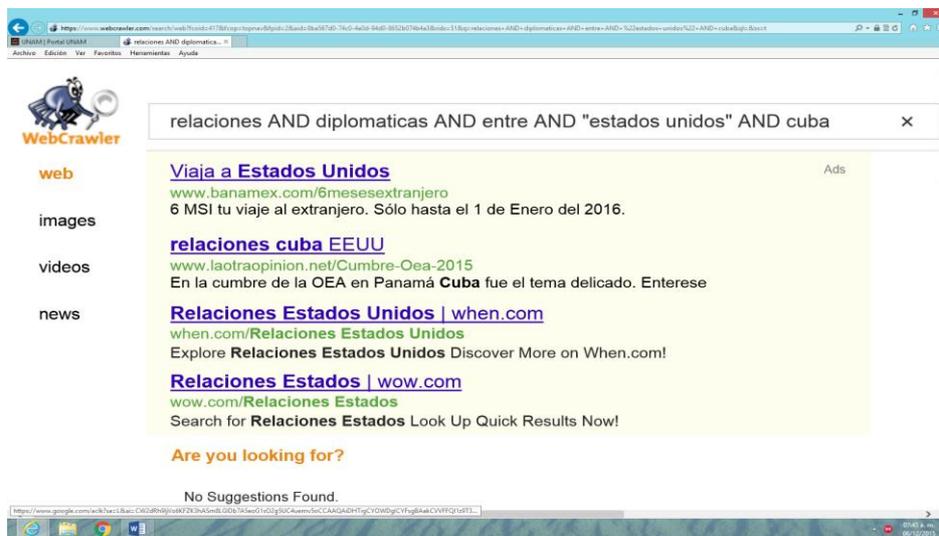


Imagen #3.53 Resultados visibles en búsqueda con operadores en español en WebCrawler

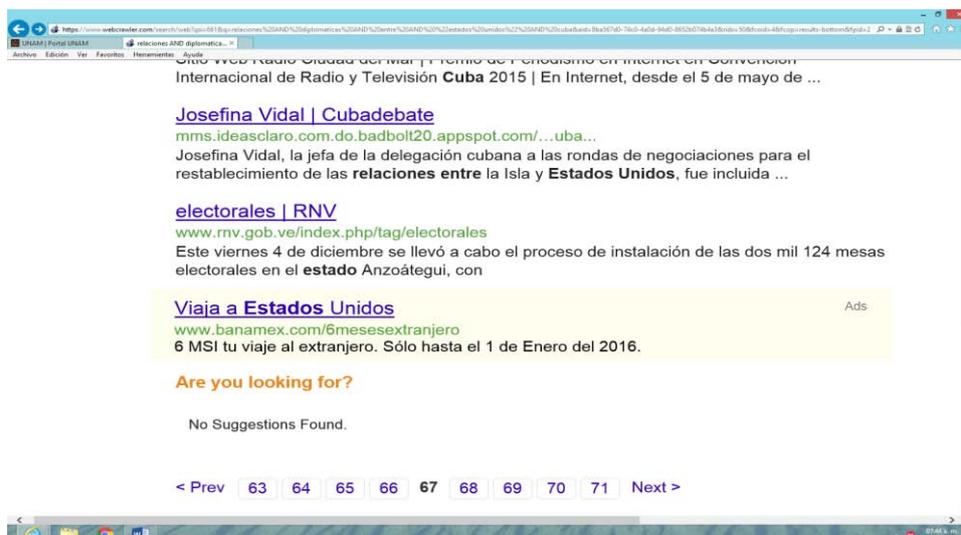


Imagen #3.55 Búsqueda libre en inglés en WebCrawler

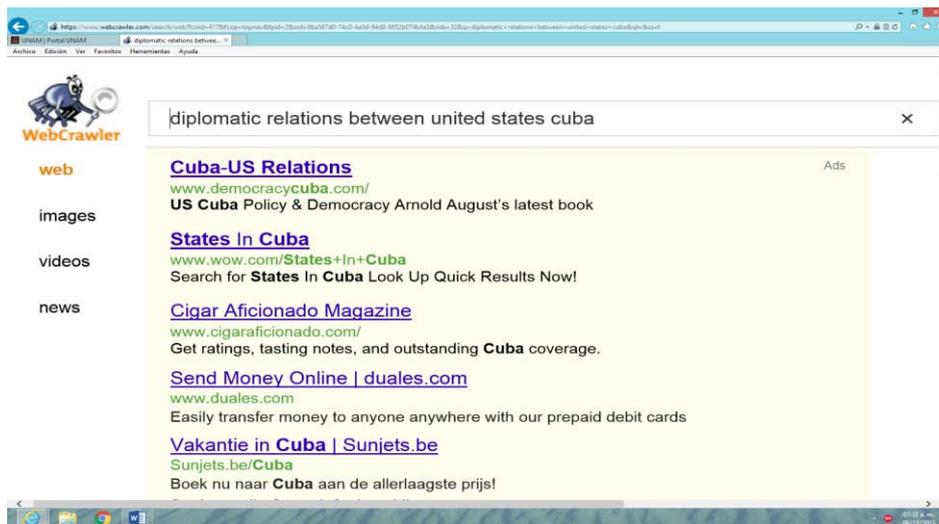
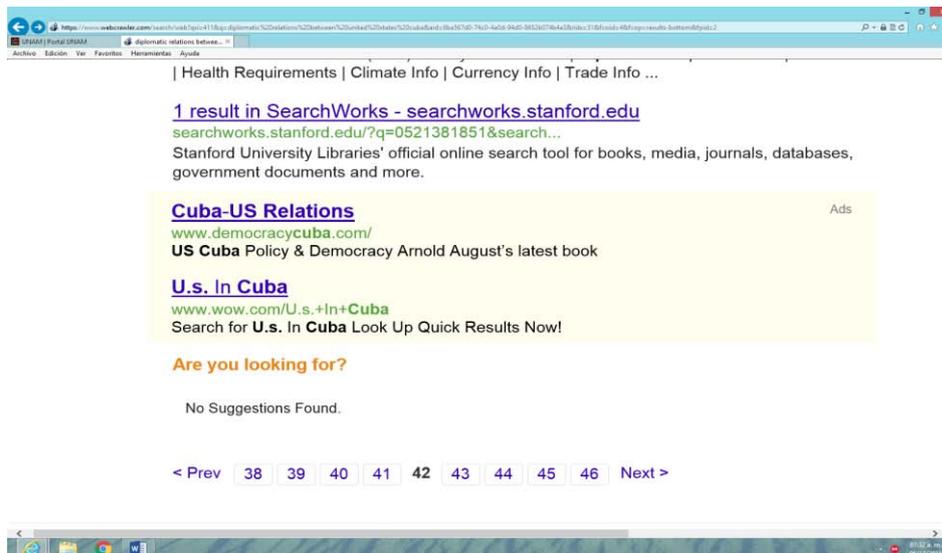


Imagen #3.56 Resultados visibles búsqueda libre en inglés en WebCrawler



Búsqueda con operadores lógicos booleanos

Al realizar la búsqueda en inglés, utilizando los operadores lógicos booleanos, los resultados fueron los siguientes, el metabuscador recuperó 30 páginas con cerca de 300 documentos recuperados, tal como se muestra en las imágenes #3.57 y #3.58.

Imagen #3.57 Búsqueda con operadores en inglés en WebCrawler

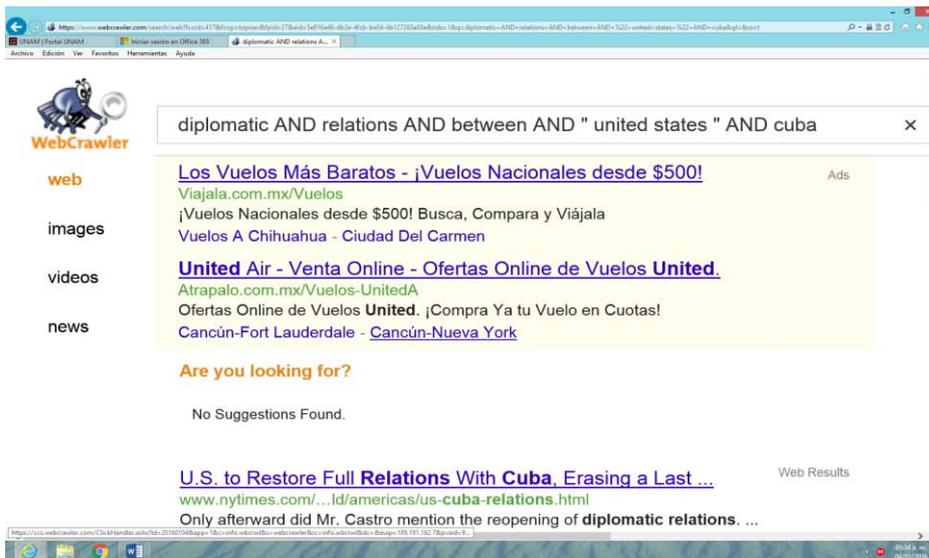
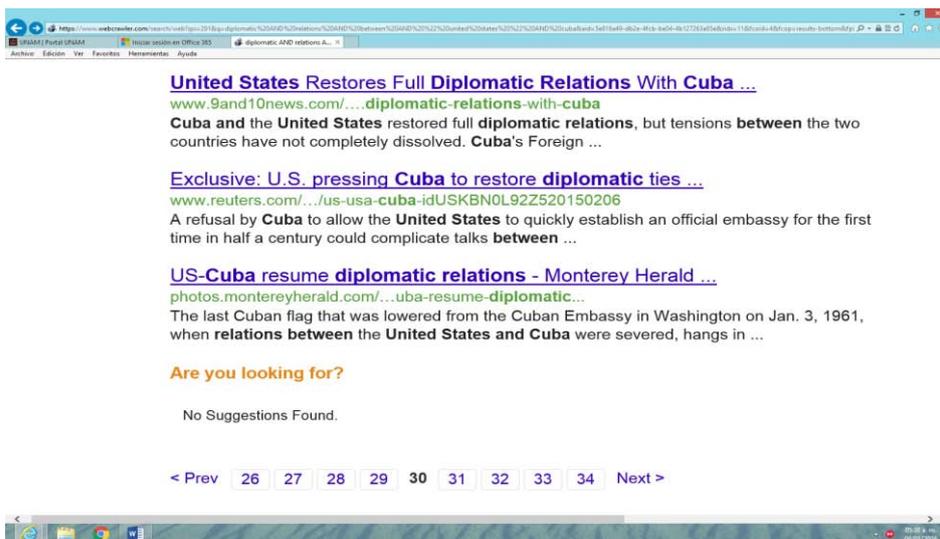


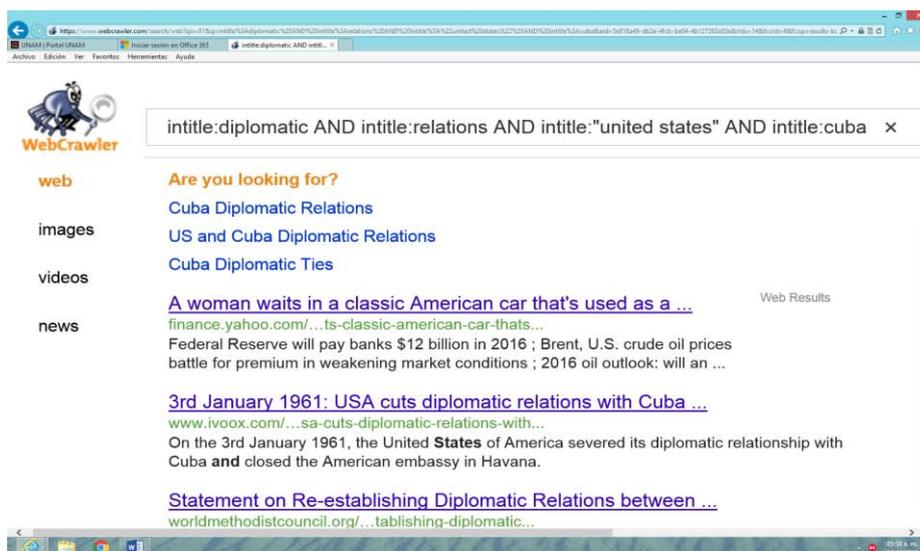
Imagen #3.58 Resultados visibles con operadores búsqueda con operadores en inglés en WebCrawler



Búsqueda por campos

Al combinar operadores lógicos, comillas y realizar la búsqueda en el campo de título, se obtuvieron resultados que no tenían que ver con la búsqueda, como se puede apreciar en la imagen #3.59.

Imagen #3.59 Búsqueda por campos en inglés en WebCrawler



3.6.8 Resultados de la recuperación de información en Dogpile

A continuación se presentan los resultados de la búsqueda de información en el metabuscador Dogpile, con respecto a la búsqueda.

Búsquedas en español

Búsqueda libre

Se introdujo la frase completa en español en el metabuscador y nos entregó 94 páginas de resultados, cerca de 940, como se puede apreciar en las imágenes #3.60 y #3.61.

Imagen #3.60 Búsqueda libre en español en Dogpile

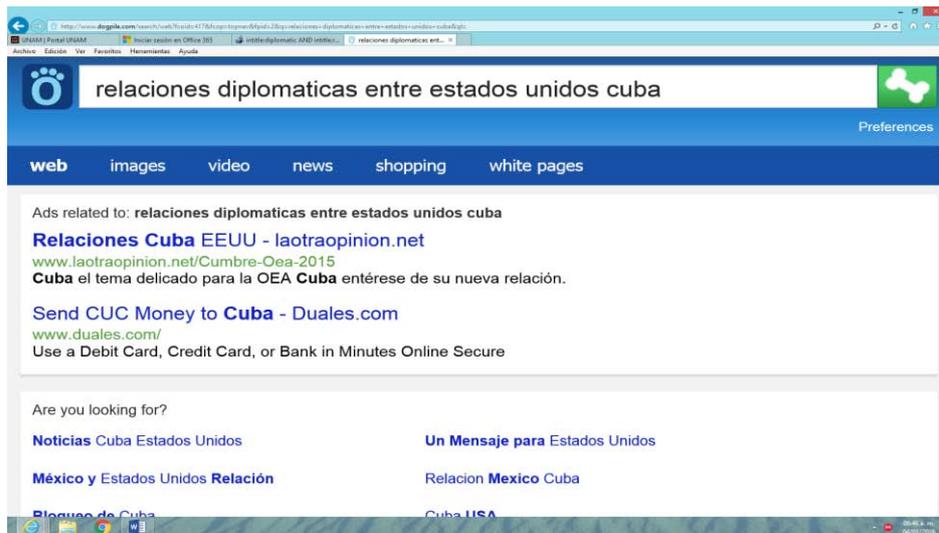
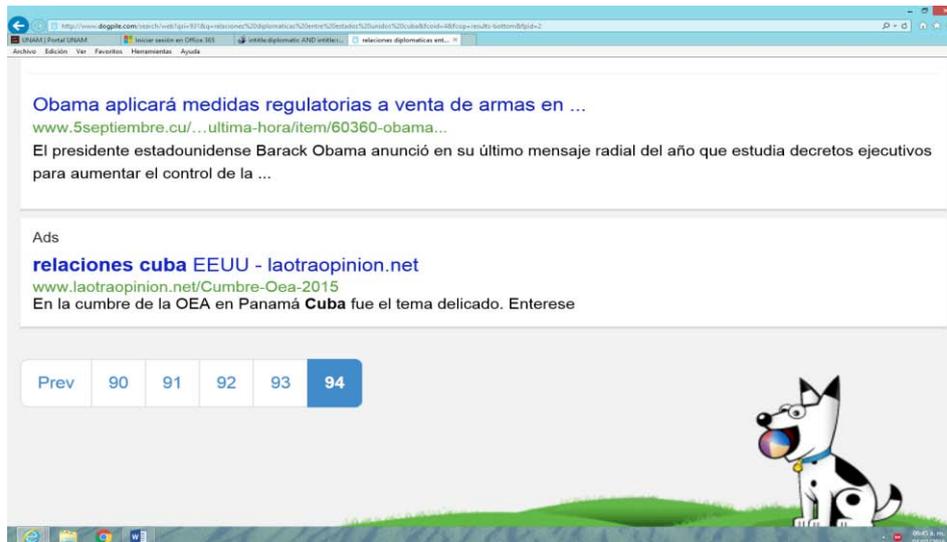


Imagen #3.61 Resultados visibles búsqueda libre en español en Dogpile



Búsqueda con operadores lógicos boléanos

Al realizar la búsqueda en español, utilizando los operadores lógicos boléanos, los resultados fueron los siguientes, el metabuscador recuperó 37 páginas con cerca de 370 documentos recuperados, tal como se muestra en las imágenes # 3.61 y #3.62.

Imagen #3.61 Búsqueda libre en español en Dogpile

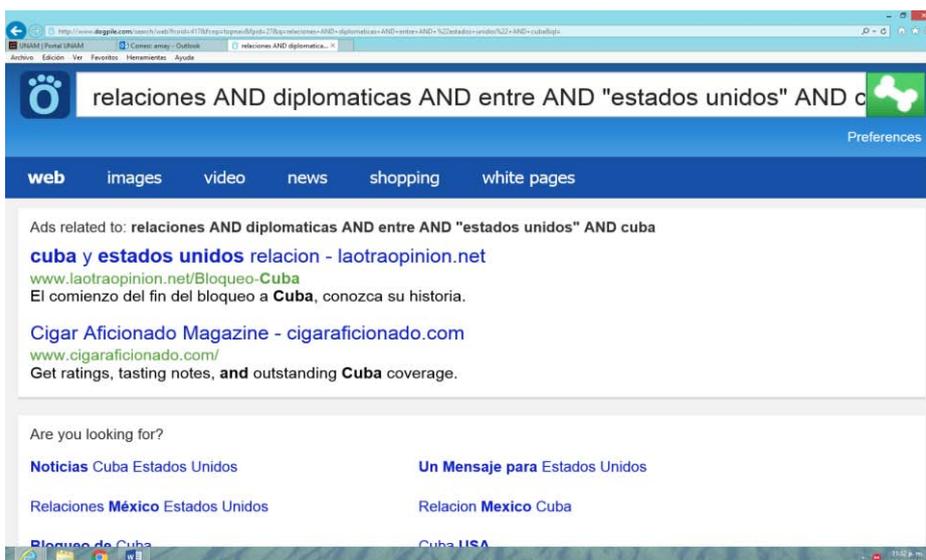


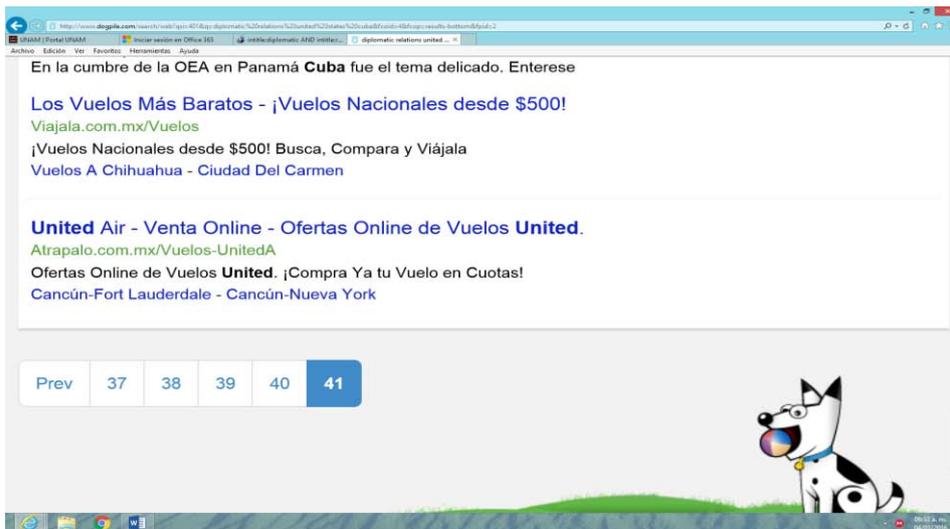
Imagen #3.62 Resultados visibles búsqueda libre en español en Dogpile



Búsqueda por campos

Al combinar operadores lógicos, comillas y realizar la búsqueda en el campo de título, no se obtuvieron resultados en el buscador, como se puede apreciar en la imagen #3.63.

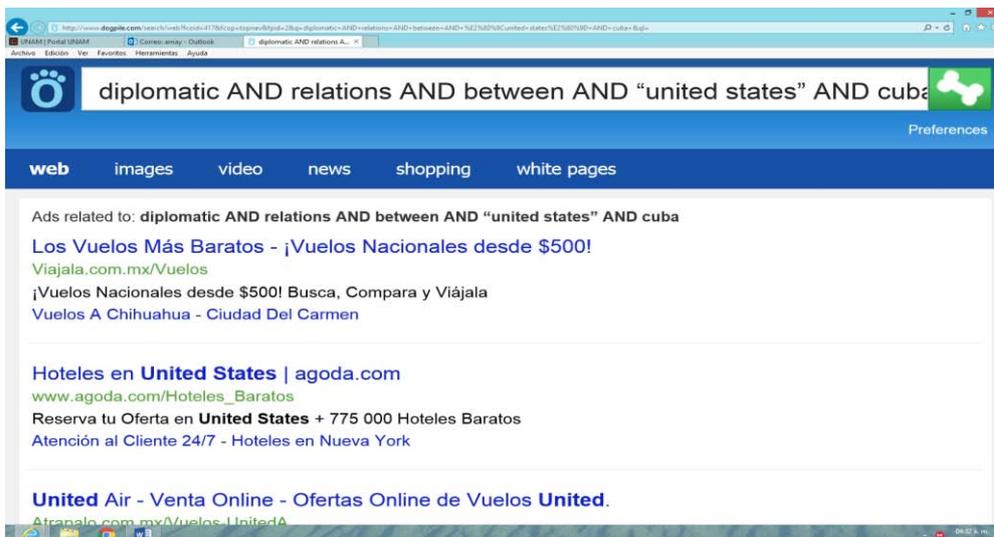
Imagen #3.65 Resultados visibles búsqueda libre en inglés en Dogpile



Búsqueda con operadores lógicos boléanos

Al realizar la búsqueda en inglés, utilizando los operadores lógicos boléanos, los resultados fueron los siguientes, el metabuscador recuperó 1 página con solo 10 documentos recuperados, tal como se muestra en la imagen #3.66

Imagen #3.66 Búsqueda con operadores lógicos en inglés en Dogpile



Búsqueda por campos

Al combinar operadores lógicos, comillas y realizar la búsqueda en el campo de título, se obtuvieron 13 páginas de resultados con aproximadamente 130 documentos, como se puede apreciar en las imágenes #3.67 y #3.68.

Imagen #3.67 Búsqueda por campos en inglés en Dogpile

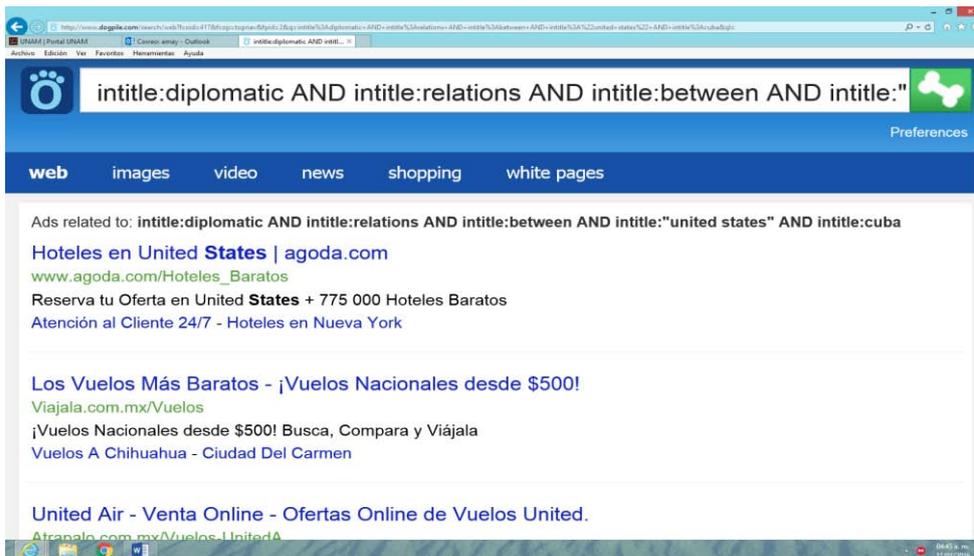
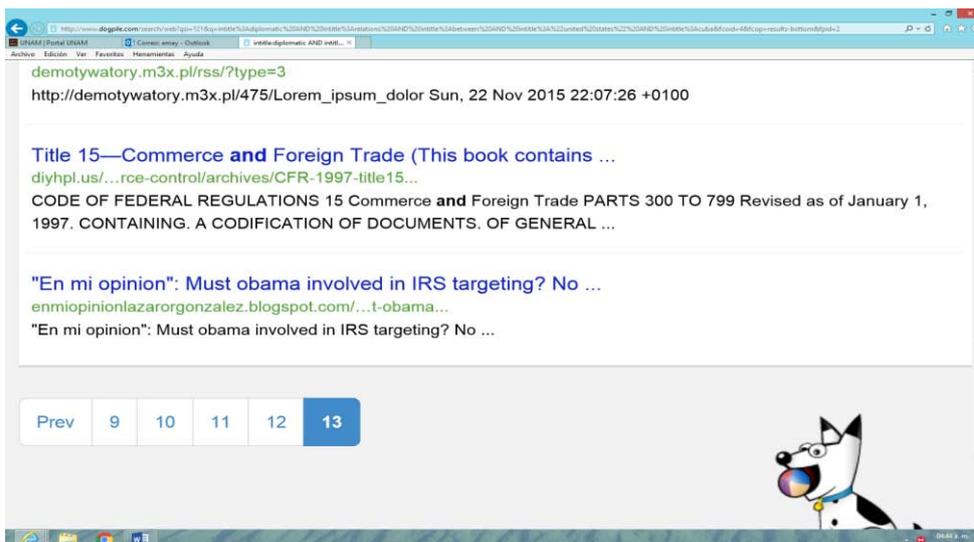


Imagen #3.68 Resultados visibles búsqueda por campos en inglés en Dogpile



Búsqueda con operadores lógicos boléanos

Al realizar la búsqueda en español, utilizando los operadores lógicos boléanos, los resultados fueron los siguientes, el metabuscador recuperó 33,800 resultados, pero solo se tiene acceso a 5 páginas con cerca de 50 documentos tal como se muestra en las imágenes #3.71 y #3.72.

Imagen #3.71 Búsqueda con operadores lógicos en español en Zapmeta

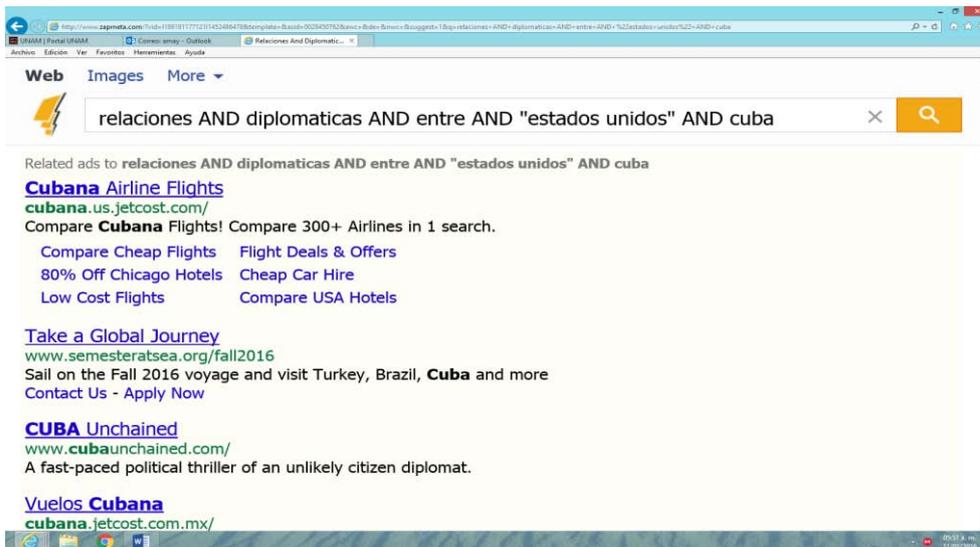
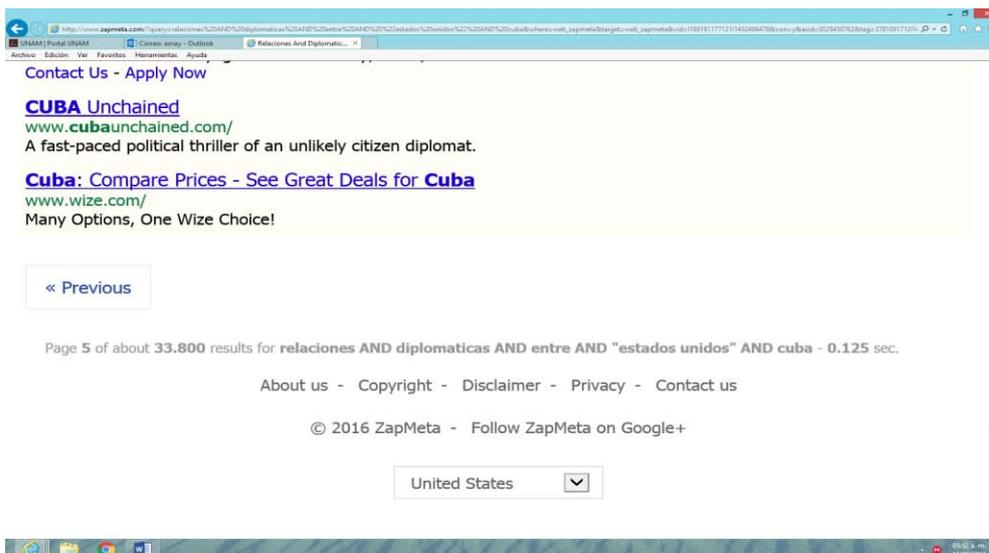


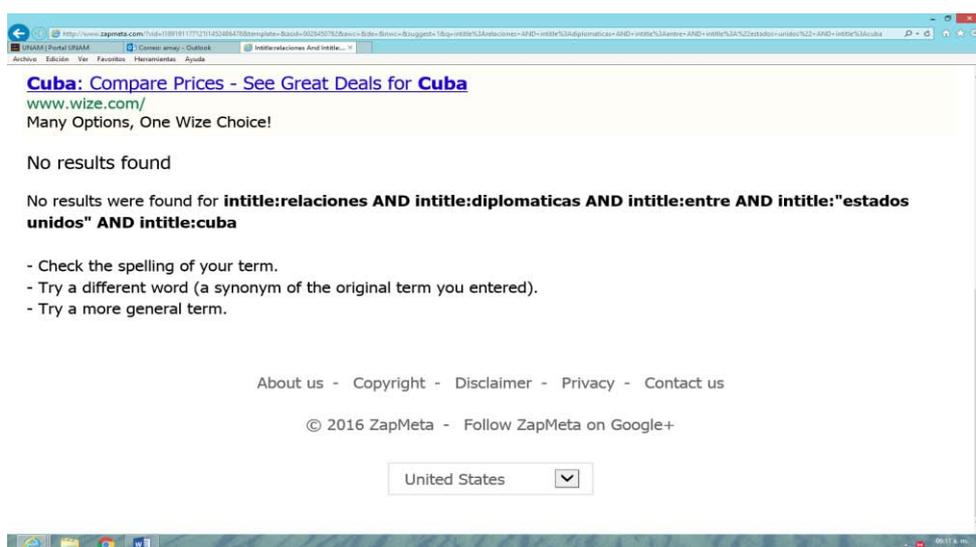
Imagen #3.72 Resultados visibles búsqueda con operadores lógicos en español en Zapmeta



Búsqueda por campos

Al combinar operadores lógicos, comillas y realizar la búsqueda en el campo de título, no se obtuvieron resultados en el metabuscador, como se puede apreciar en la siguiente imagen #3.73.

Imagen #3.73 Búsqueda por campos en español en Zapmeta



Búsquedas en inglés

Búsqueda libre

Con respecto a la búsqueda de información en inglés, el metabuscador, recupero 39, 600 000 resultados, pero solo se tiene acceso a 5 páginas con cerca de 50 documentos, como se puede apreciar en las imágenes #3.74 y #3.75.

Imagen #3.71 Búsqueda libre en inglés en Zapmeta

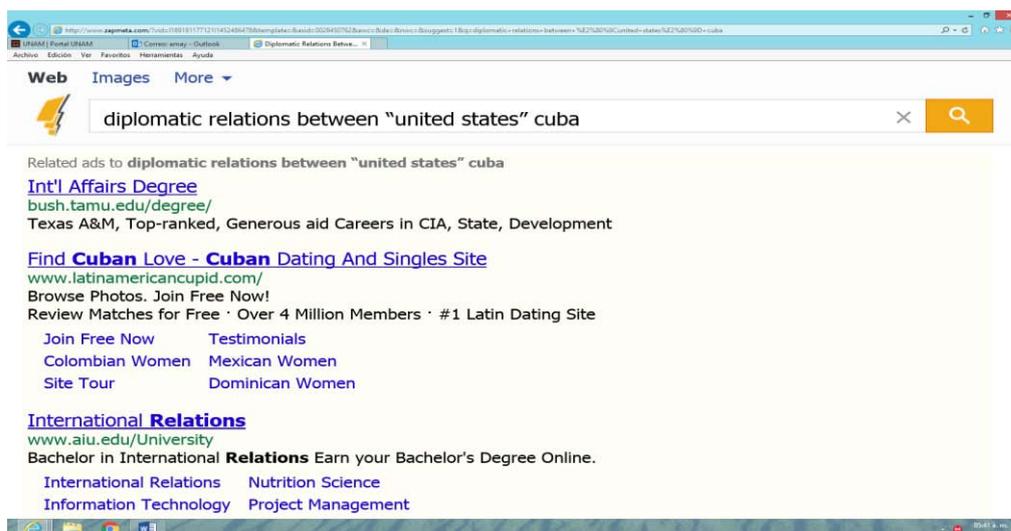
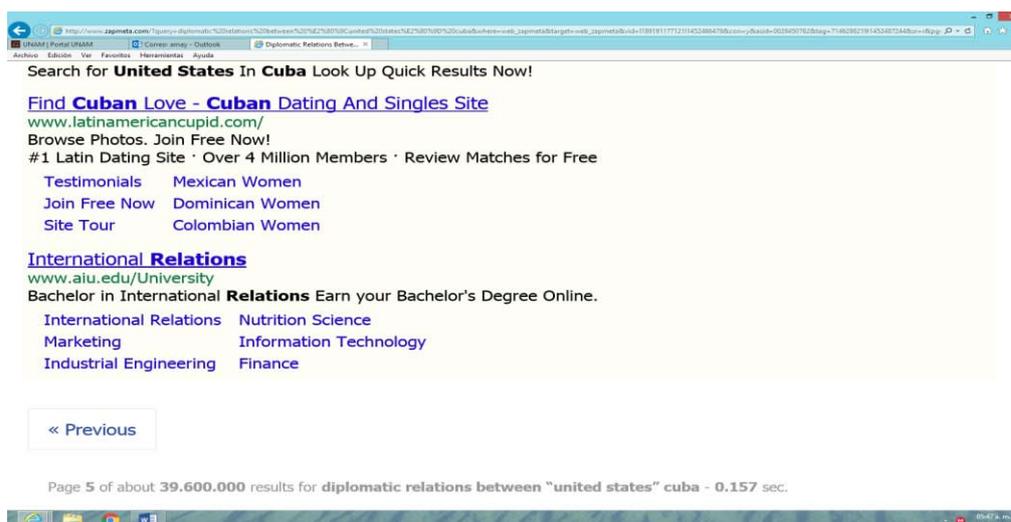


Imagen #3.72 Resultados visibles Búsqueda libre en español en Zapmeta



Búsqueda con operadores lógicos boléanos

Al realizar la búsqueda en inglés, los resultados fueron los siguientes, el metabuscador recuperó 1, 230 000 resultados, pero solo se tiene acceso a 5 páginas con cerca de 50 documentos, tal como se muestra en las imágenes #3.73 y #3.74.

Imagen #3.73 Búsqueda con operadores lógicos en inglés en Zapmeta

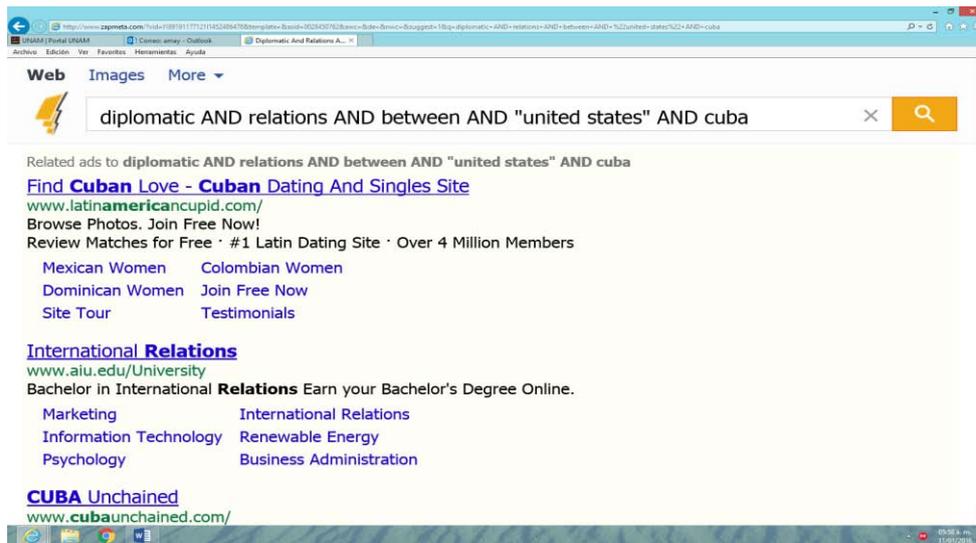
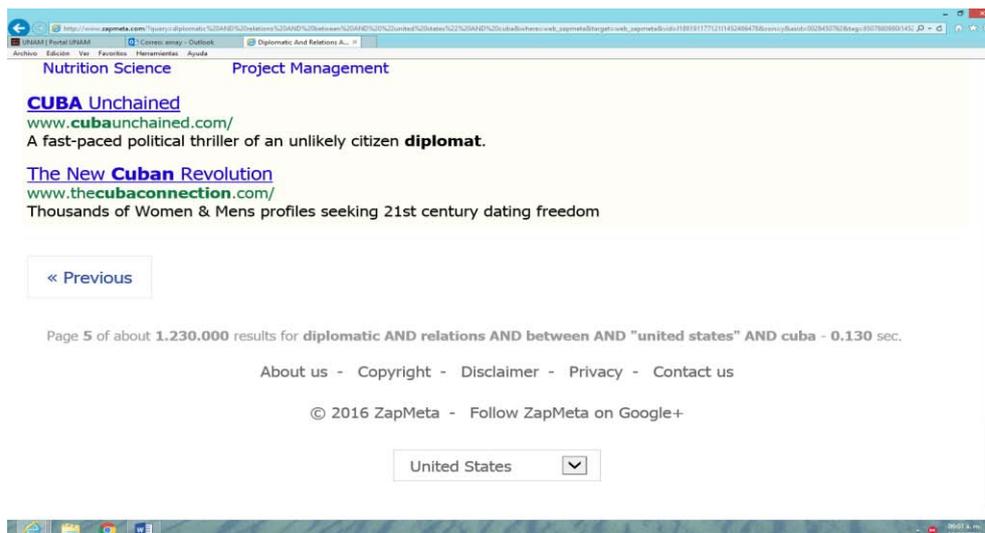


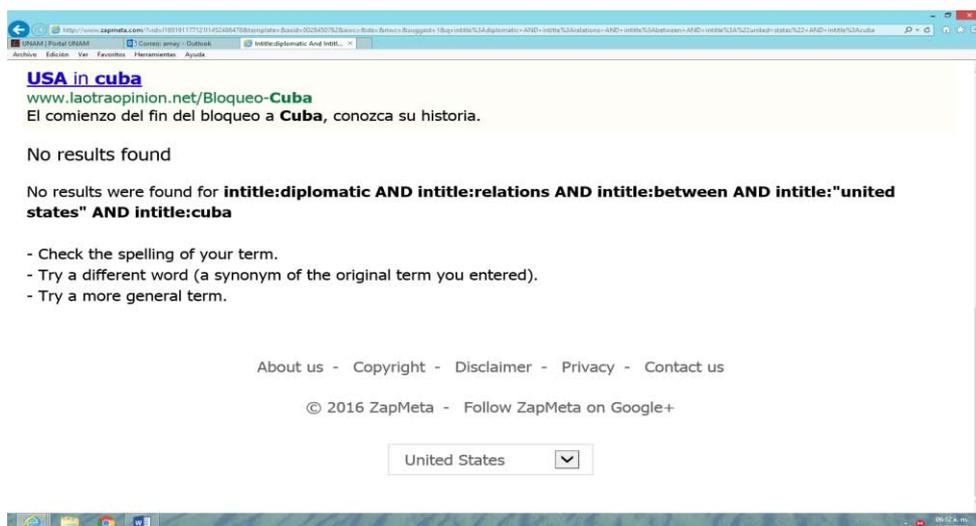
Imagen #3.74 Resultados visibles búsqueda con operadores lógicos en inglés en Zapmeta



Búsqueda por campos

Al combinar operadores lógicos, comillas y realizar la búsqueda en el campo de título, no se obtuvieron resultados, como se puede apreciar en la imagen #3.75.

Imagen #3.75 Búsqueda por campos en inglés en Zapmeta



3.6.10 Resultados de la recuperación de información en Info.com

A continuación se presentan los resultados de la búsqueda de información en el metabuscador Info.com con respecto a la búsqueda.

Búsquedas en español

Búsqueda libre

Se introdujo la frase completa en español en el metabuscador y nos entregó 30 páginas de resultados, cerca de 300 documentos, como se puede apreciar en las imágenes #3.76 y #3.77.

Imagen #3.76 Búsqueda libre en español en Info.com



Imagen #3.77 Resultados visibles búsqueda libre en español en Info.com



Búsqueda con operadores lógicos booleanos

Al realizar la búsqueda en español, utilizando los operadores lógicos booleanos, los resultados fueron los siguientes, el metabuscador recuperó 32 páginas con cerca de 320 documentos recuperados, tal como se muestra en las imágenes #3.78 y #3.79.

Imagen #3.78 Búsqueda con operadores lógicos en español en Info.com

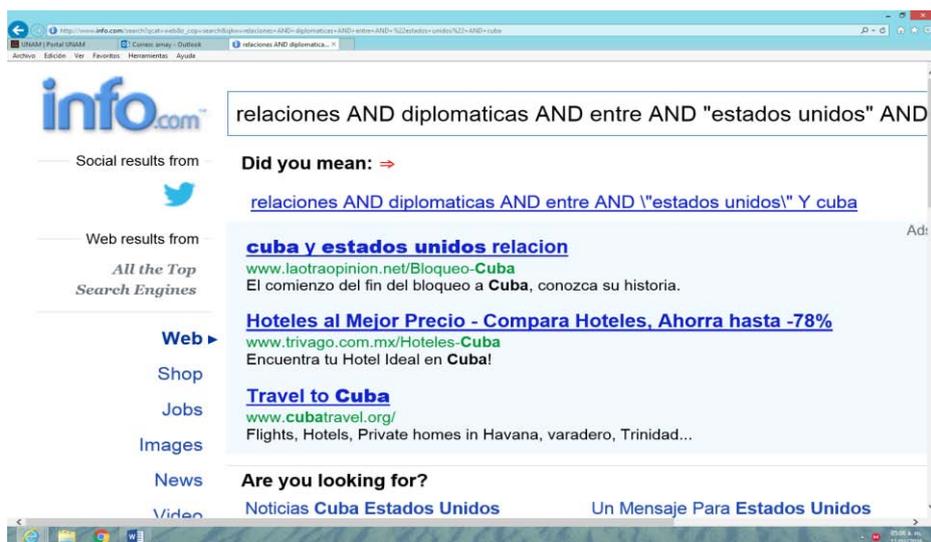


Imagen #3.79 Resultados visibles búsqueda con operadores lógicos en español en Info.com



Búsqueda por campos

Al combinar operadores lógicos, comillas y realizar la búsqueda en el campo de título, no se obtuvieron resultados en el metabuscador, como se puede apreciar en la imagen #3.80.

Imagen #3.80 Búsqueda por campos en español en Info.com



Búsquedas en inglés

Búsqueda libre

Con respecto a la búsqueda de información en inglés, el metabuscador, solamente recupero 40 páginas con cerca de 400 documentos, así lo demuestran las imágenes #3.81 y #3.82.

Imagen #3.81 Búsqueda libre en inglés en Info.com



Imagen #3.82 Resultados visibles búsqueda libre en inglés en Info.com



Búsqueda con operadores lógicos boléanos

Al realizar la búsqueda en inglés, los resultados fueron los siguientes, el metabuscador recuperó 38 páginas con cerca de 380 documentos recuperados, tal como se muestra en las imágenes #3.83 y #3.84.

Imagen #3.83 Búsqueda con operadores lógicos en inglés en Info.com

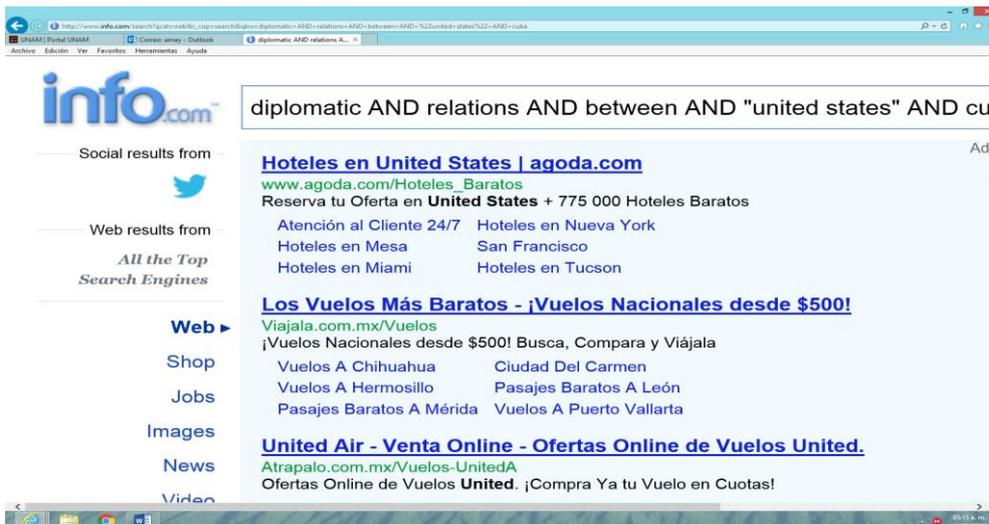


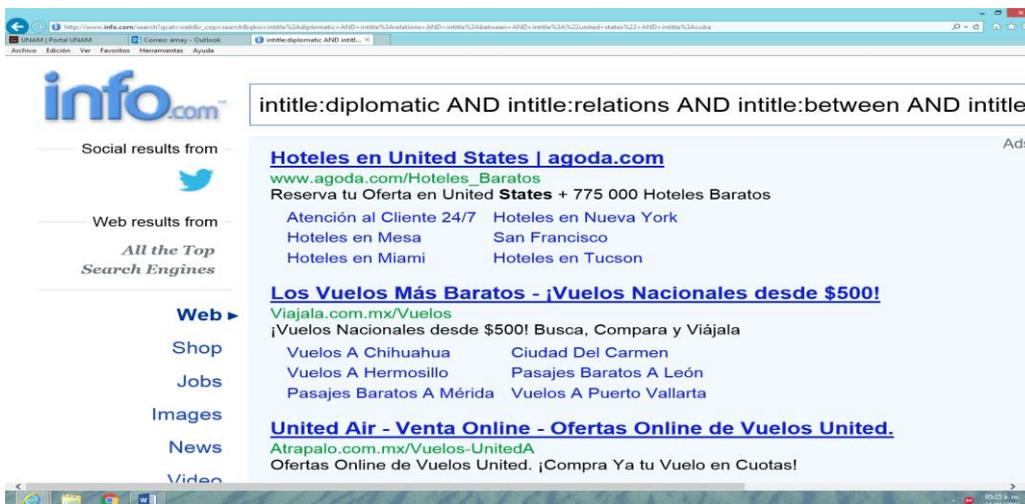
Imagen #3.84 Resultados visibles búsqueda con operadores en inglés en Info.com



Búsqueda por campos

Al combinar operadores lógicos, comillas y realizar la búsqueda en el campo de título, se obtuvo 1 página de resultados con aproximadamente 10 documentos, como se puede apreciar en la imagen #3.85.

Imagen #3.85 Búsqueda por campos en inglés en Info.com



3.6.11 Resultados de la recuperación de información en Exite

A continuación se presentan los resultados de la búsqueda de información en el metabuscador Exite con respecto a la búsqueda.

Búsqueda libre en español

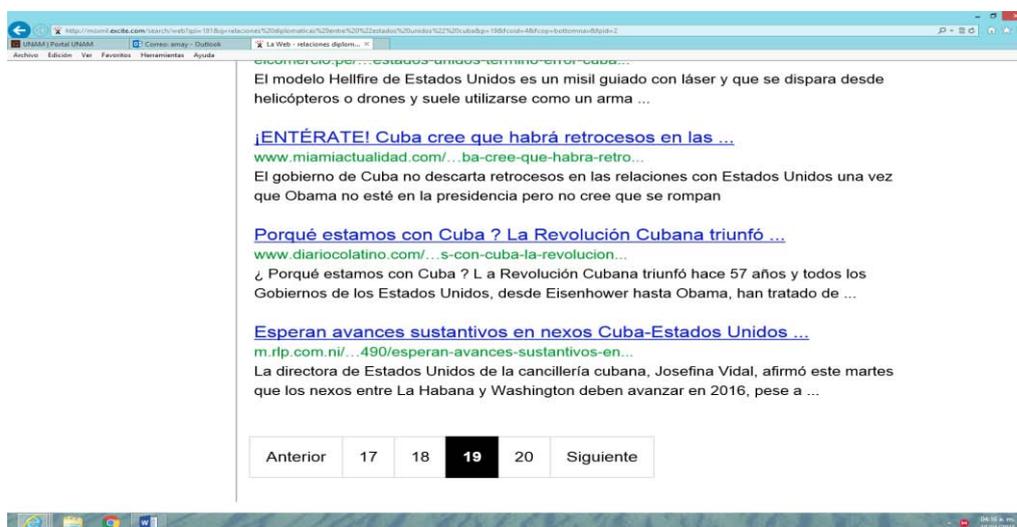
Búsqueda libre

Se introdujo la frase completa en español en el metabuscador y nos entregó 20 páginas de resultados, cerca de 400 documentos, como se puede apreciar en las imágenes #3.86 y #3.87.

Imagen #3.86 Búsqueda libre en español en Exite



Imagen #3.87 Resultados visibles búsqueda libre en español en Exite



Búsqueda con operadores lógicos booleanos

Al realizar la búsqueda en español, utilizando los operadores lógicos booleanos, los resultados fueron los siguientes, el metabuscador recuperó 20 páginas con cerca de 400 documentos recuperados, tal como se muestra en las imágenes #3.88 y #3.89

Imagen #3.88 Búsqueda con operadores lógicos en español en Exite



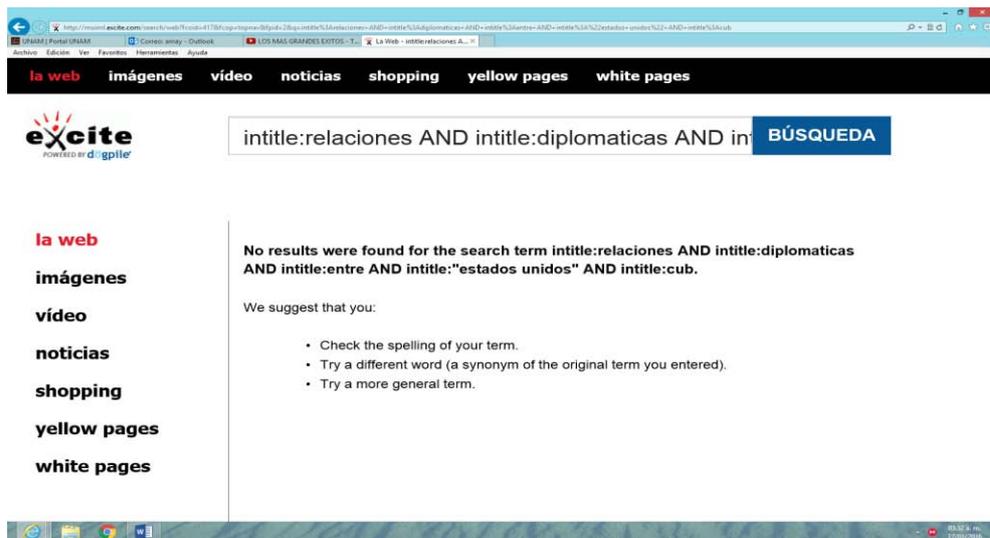
Imagen #3.89 Resultados visibles búsqueda con operadores lógicos en español en Exite



Búsqueda por campos

Al combinar operadores lógicos, comillas y realizar la búsqueda en el campo de título, no se obtuvieron resultados en el metabuscador, como se puede apreciar en la imagen #3.90.

Imagen #3.90 Búsqueda por campos en español en Exite



Búsquedas en inglés

Búsqueda libre

Con respecto a la búsqueda de información en inglés, el metabuscador, no recupero documentos, así lo demuestra la imagen #3.91.

Imagen #3.91 Búsqueda libre en inglés en Exite



Búsqueda con operadores lógicos boléanos

Al realizar la búsqueda en inglés, los resultados fueron los siguientes, el metabuscador recuperó 23 páginas con cerca de 460 documentos recuperados, tal como se muestra en las imágenes #3.92 y #3.93.

Imagen #3.92 Búsqueda con operadores lógicos en inglés en Exite

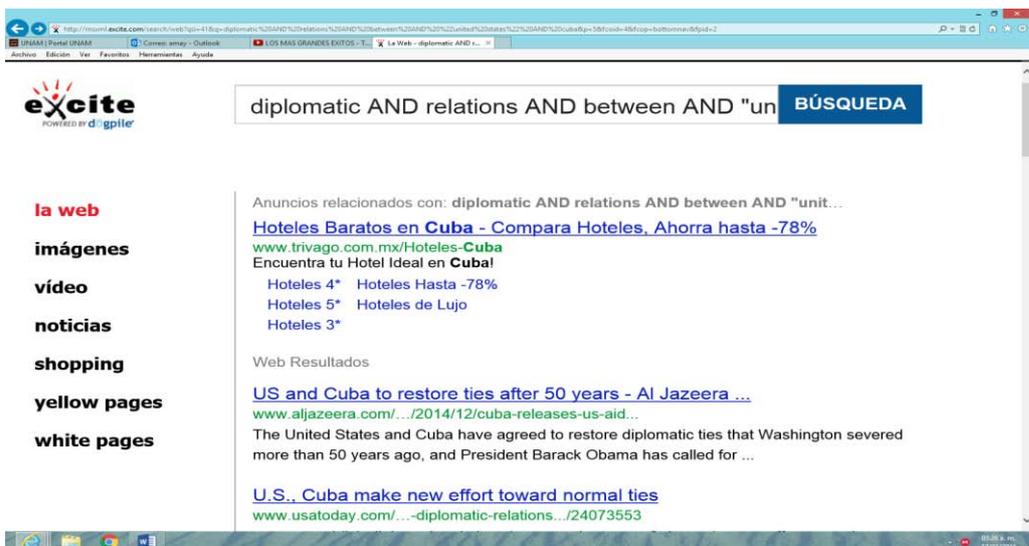
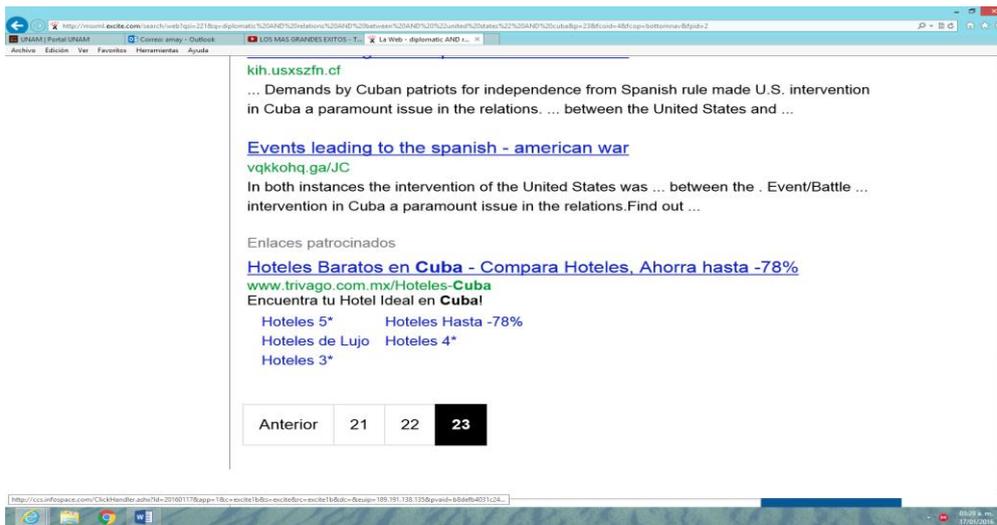


Imagen #3.93 Resultados visibles búsqueda con operadores lógicos en inglés en Exite



Búsqueda por campos

Al combinar operadores lógicos, comillas y realizar la búsqueda en el campo de título, se obtuvieron solamente 5 páginas de resultados con aproximadamente 100 documentos, como se puede apreciar en las imágenes #3.94 y #3.95.

Imagen #3.94 Búsqueda por campos en inglés en Exite

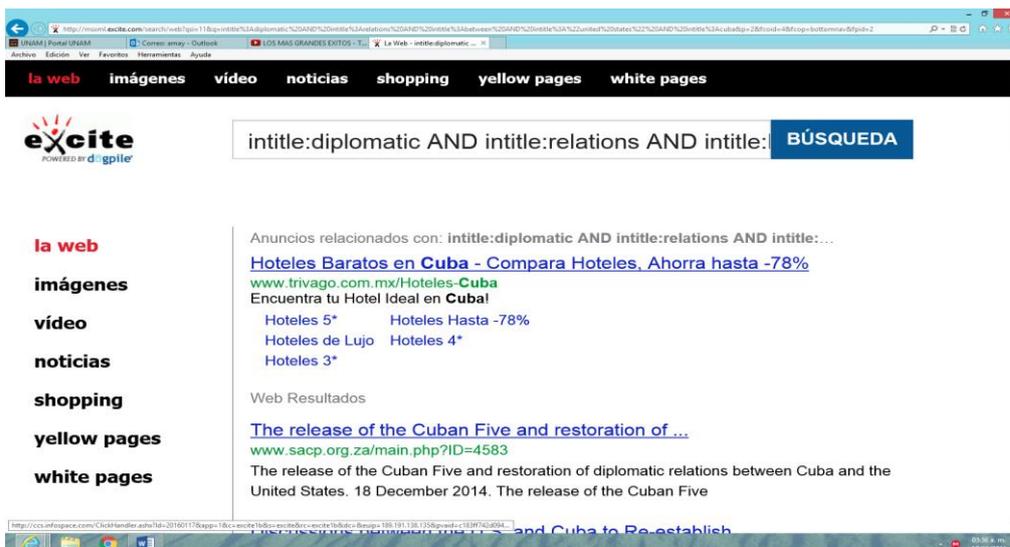
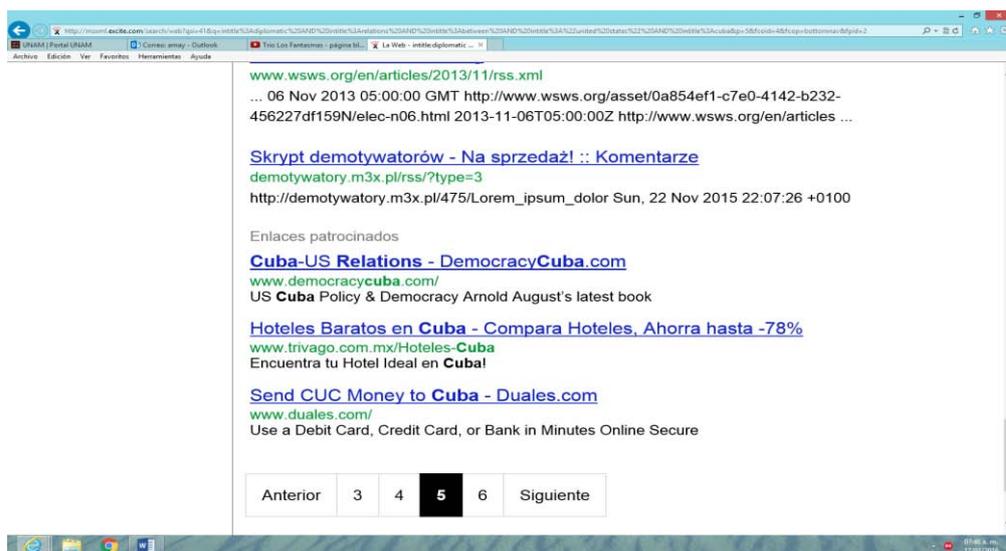


Imagen #3.95 Resultados visibles búsqueda por campos en inglés en Exite



3.6.12 Resultados de la recuperación de información en Kngine

A continuación se presentan los resultados de la búsqueda de información en el metabuscador Kngine con respecto a la búsqueda.

Búsquedas en español

Búsqueda libre

Se introdujo la frase completa en español en el buscador semántico nos entregó 1 página de resultados, cerca de 15 sitios, entre los cuales se puede observar que la mayoría son portales de noticias, periódicos y canales de televisión, como se puede apreciar en la imagen #3.96.

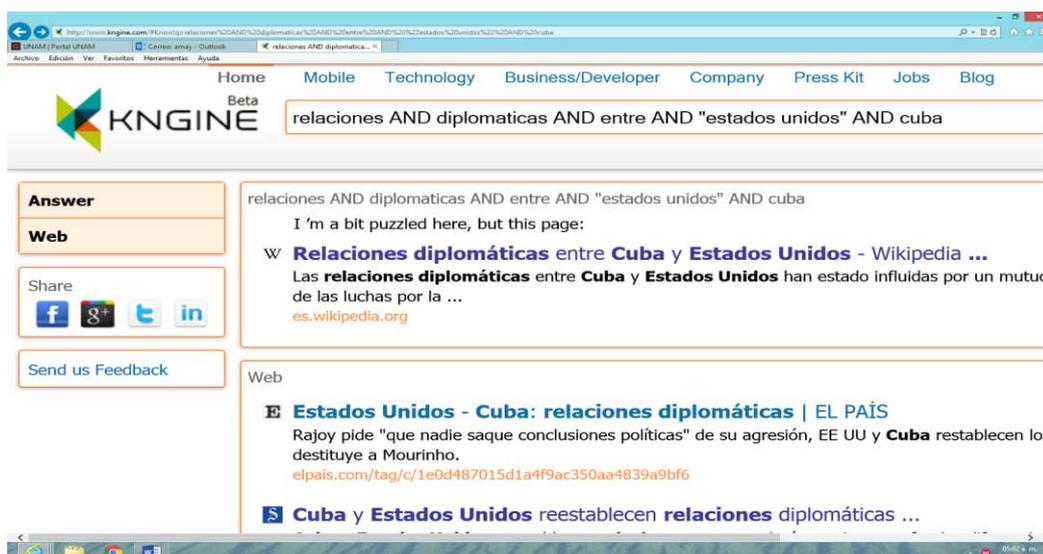
Imagen #3.96 Búsqueda libre en español en Kngine



Búsqueda con operadores lógicos boléanos

Con respecto a la búsqueda de información en inglés, el buscador semántico, recuperó 1 página de resultados, cerca de 15 sitios, la mayoría portales de noticias, diarios y cadenas de televisión, así lo demuestra la imagen #3.97.

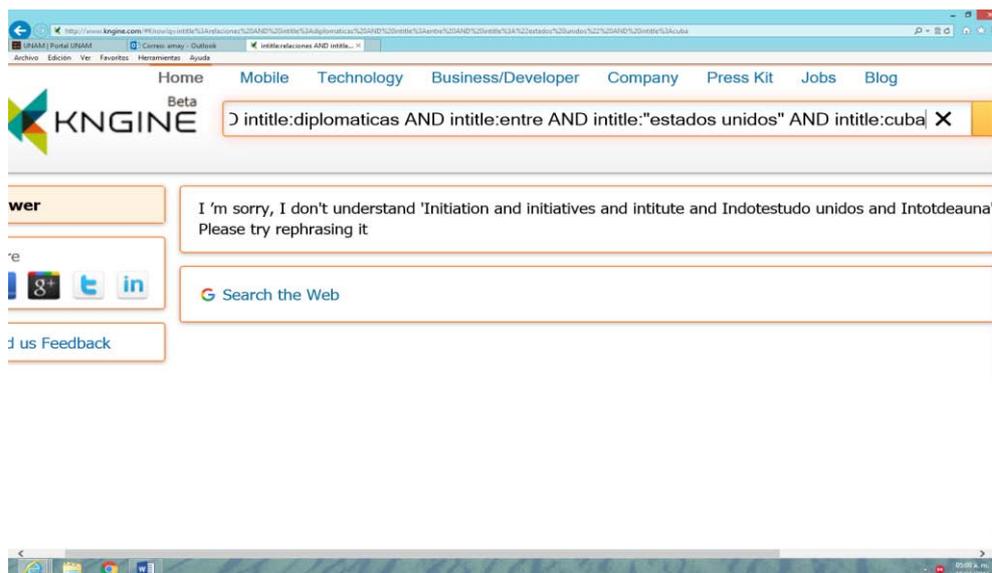
Imagen #3.97 Búsqueda con operadores lógicos en español en Kngine



Búsqueda por campos

Al realizar la búsqueda en el campo de título en español, no se obtuvieron resultados en el buscador semántico, como se puede apreciar en la imagen #3.98.

Imagen #3.98 Búsqueda por campos en español en Kngine

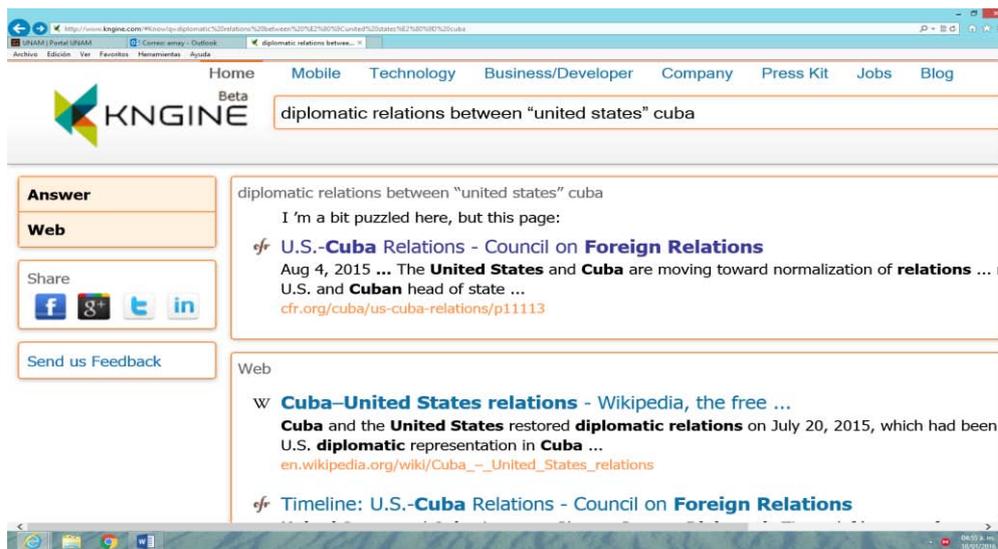


Búsquedas en inglés

Búsqueda libre

Con respecto a la búsqueda de información en inglés, el buscador semántico, recuperó 1 página de resultados, cerca de 15 sitios, la mayoría portales de noticias, diarios y cadenas de televisión, así lo demuestra la imagen #3.99.

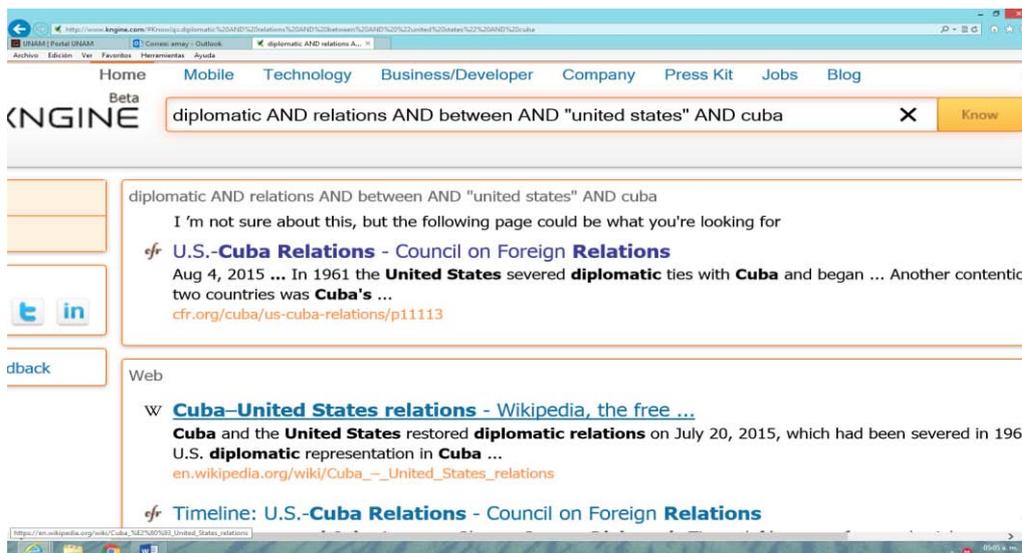
Imagen #3.99 Búsqueda libre en inglés en Kngine



Búsqueda con operadores lógicos boléanos

El buscador semántico, recuperó 1 página de resultados, cerca de 15 sitios, la mayoría portales de noticias, diarios y cadenas de televisión, así lo demuestra la imagen #3.100.

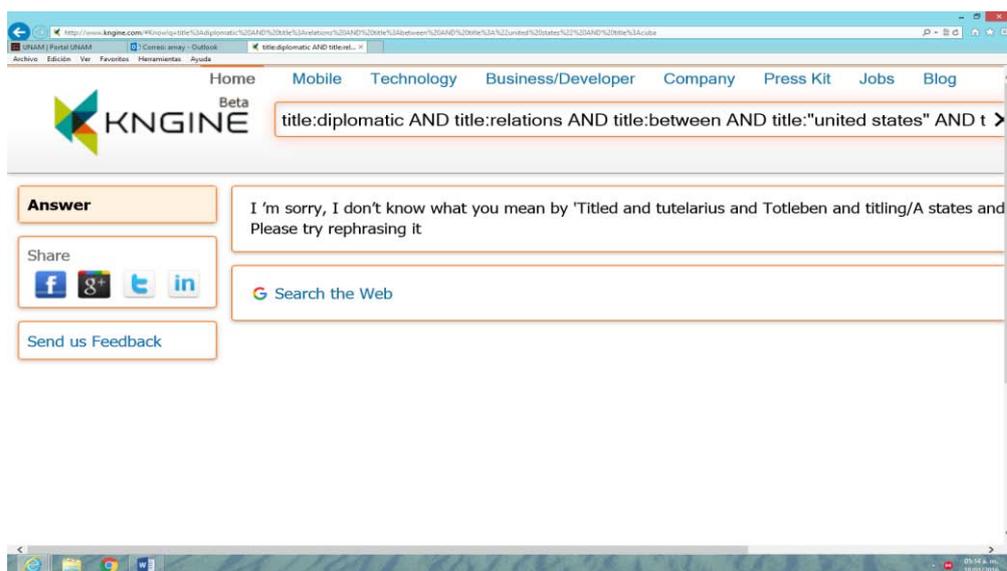
Imagen #3.100 Búsqueda con operadores lógicos en inglés en Kngine



Búsqueda por campos

Al realizar la búsqueda en el campo de título en inglés, no se obtuvieron resultados en el buscador semántico, como se puede apreciar en la imagen #3.101.

Imagen #3.101 Búsqueda por campos en español en Kngine



3.6.13 Resultados de la recuperación de información en Exalead

A continuación se presentan los resultados de la búsqueda de información en el metabuscador Exite con respecto a la búsqueda.

Búsquedas en español

Búsqueda libre

Se introdujo la frase completa en español en el metabuscador y nos entregó 40,790 resultados, divididos por año desde el 2002 hasta el 2015, como se puede apreciar en las imágenes #3.102 y #3.103.

Imagen #3.102 Búsqueda libre en español en Exalead



Imagen #3.103 Resultados visibles búsqueda libre en español en Exalead



Búsqueda con operadores lógicos boléanos

Al realizar la búsqueda en español, utilizando los operadores lógicos boléanos, los resultados fueron los siguientes, el buscador semántico recuperó 44,297 documentos, divididos por año desde el 2002 hasta el 2015, tal como se muestra en las imágenes #3.104 y #3.105

Imagen #3.104 Búsqueda con operadores boléanos en español en Exalead



Imagen #3.105 Resultados visibles búsqueda con operadores en español en Exalead



Búsqueda por campos

Al realizar la búsqueda en el campo de título en español, se obtuvieron 41 en el buscador semántico, divididos por año desde el 2014 hasta el 2015, como se puede apreciar en las imágenes #3.106 y #3.107.

Imagen #3.106 Búsqueda por campos en español en Exalead

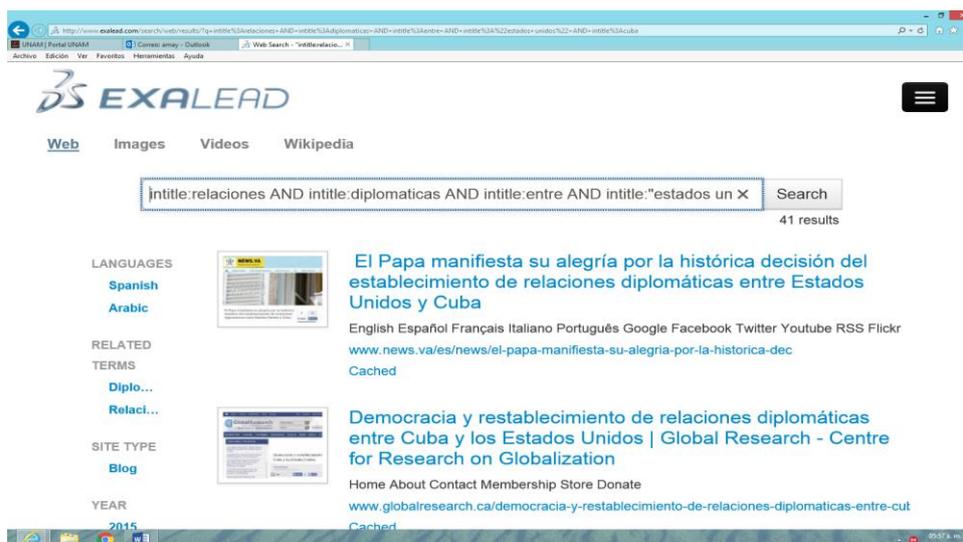
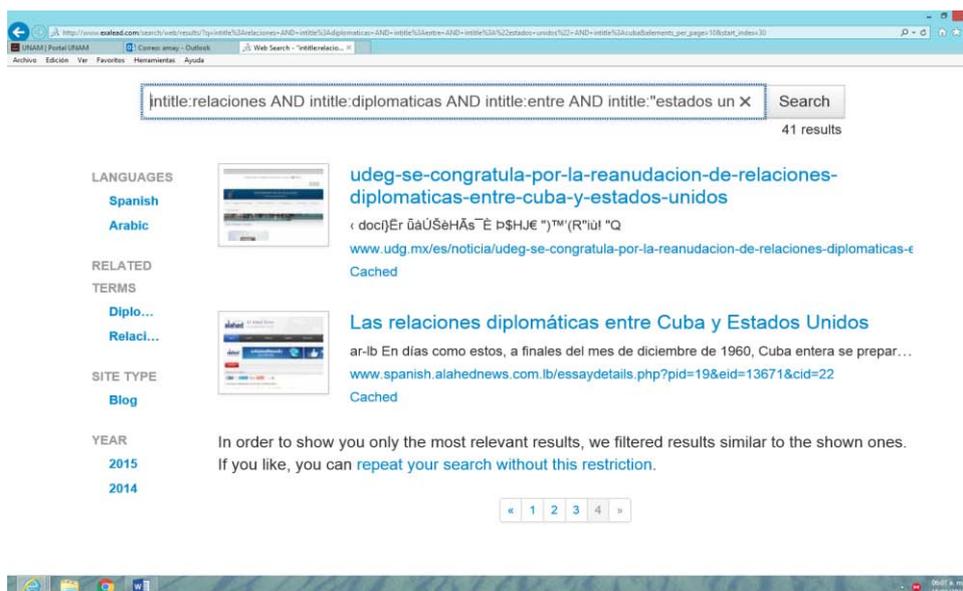


Imagen #3.107 Resultados visibles búsqueda por campos en español en Exalead



Búsquedas en inglés

Búsqueda libre

Con respecto a la búsqueda de información en inglés, el buscador semántico, recupero 153, 277 documentos, divididos por año desde el 2002 hasta el 2015, así lo demuestra las imágenes #3.108 y #3.109.

Imagen #3.108 Búsqueda libre en inglés en Exalead

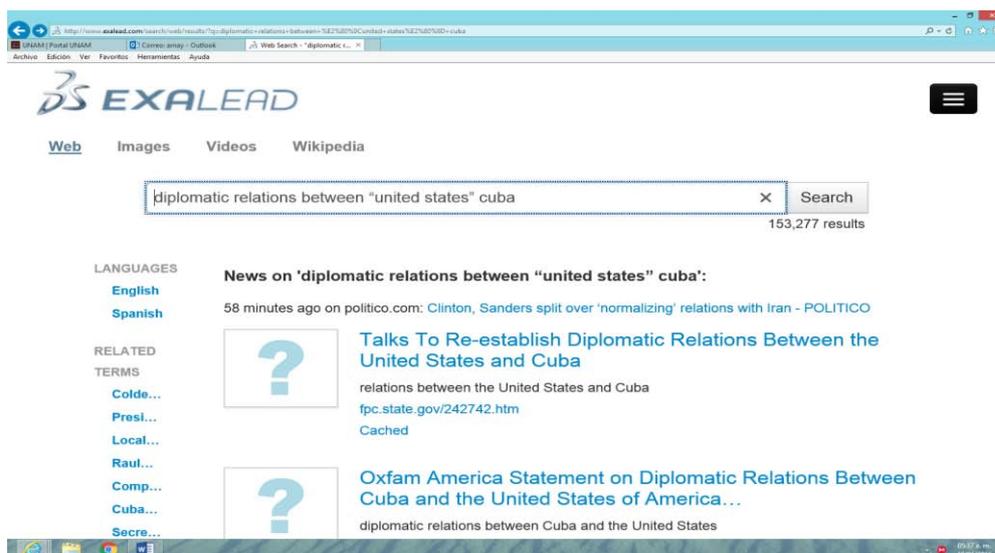
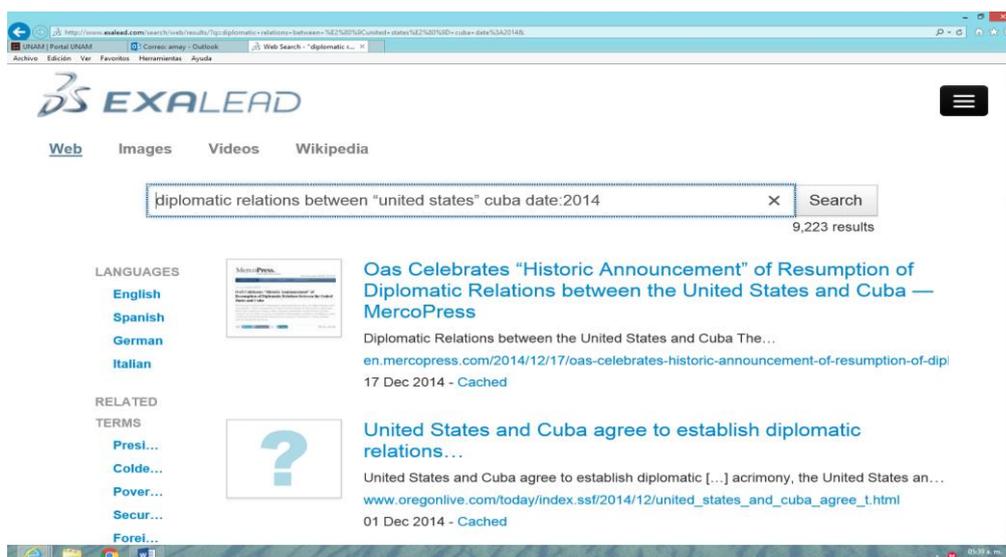


Imagen #3.109 Resultados visibles búsqueda libre en inglés en Exalead



Búsqueda con operadores lógicos boléanos

Al realizar la búsqueda en inglés, los resultados fueron los siguientes, el buscador semántico recuperó 94,753 documentos recuperados, divididos por año desde el 2002 hasta el 2015, tal como se muestra en las imágenes #3.110 y #3.111

Imagen #3.110 Búsqueda con operadores lógicos en inglés en Exalead

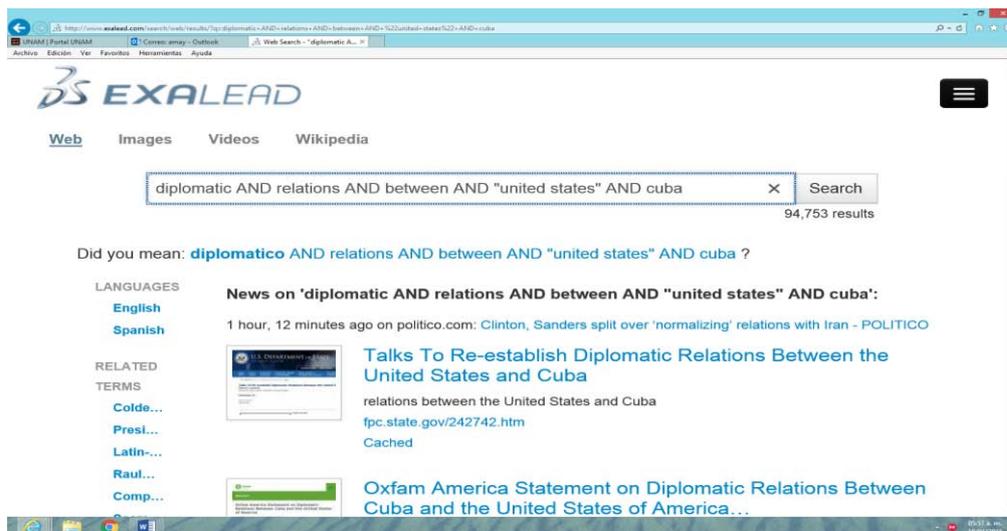
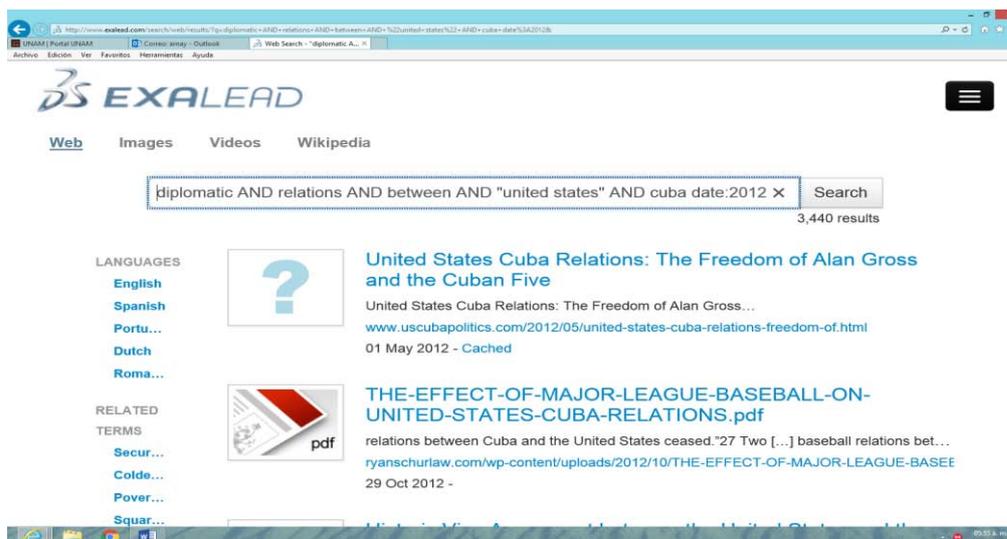


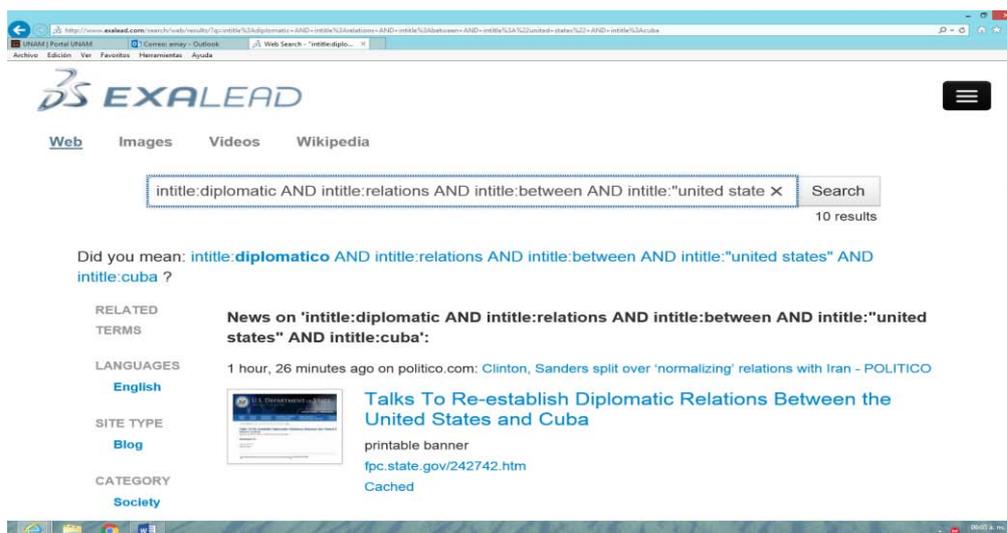
Imagen #3.111 Resultados visibles búsqueda con operadores lógicos en inglés Exalead



Búsqueda por campos

Al realizar la búsqueda en el campo de título en inglés, se obtuvieron solamente 10 resultados, como se puede apreciar en la imagen #3.12.

Imagen #3.112 Búsqueda por campos en inglés en Exalead



3.6.14 Resultados de la recuperación de información en DuckDuckgo

A continuación se presentan los resultados de la búsqueda de información en el metabuscador Exite con respecto a la búsqueda.

Búsquedas en español

Búsqueda libre

Se introdujo la frase completa en español en el buscador semántico y nos entregó 4 páginas de resultados, más de 100 documentos, como se puede apreciar en las imágenes #3.113. y #3.114

Imagen #3.113 Búsqueda libre en español en DuckDukgo

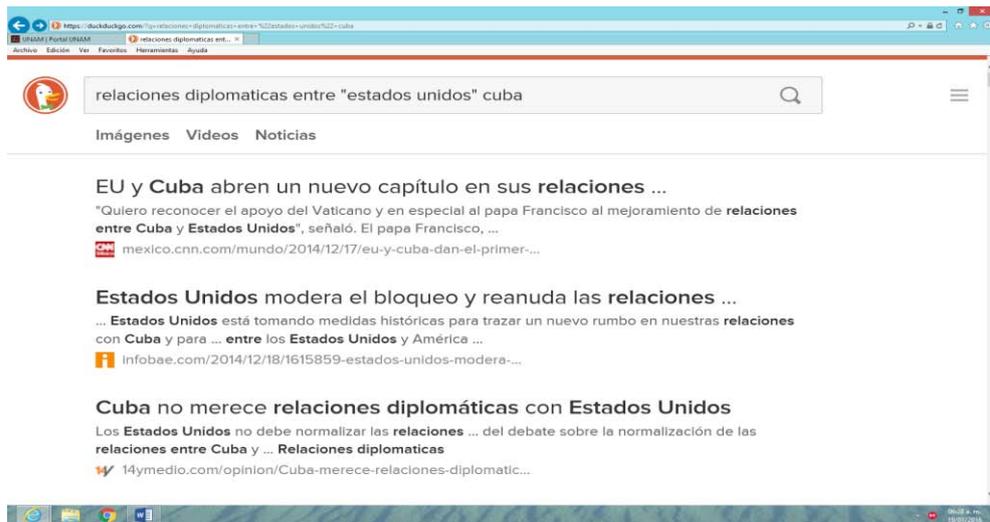
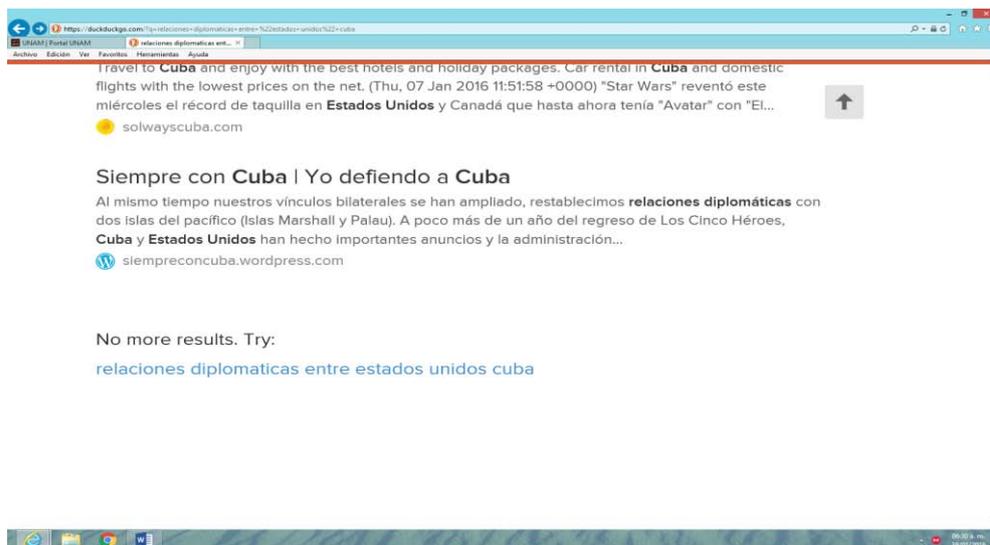


Imagen #3.114 Resultados visibles búsqueda libre en español en DuckDukgo



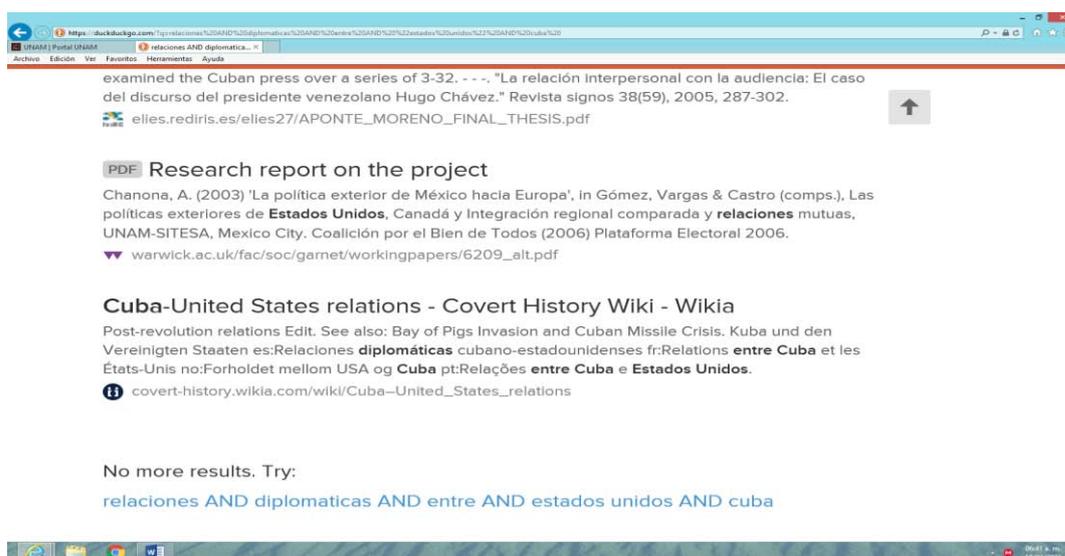
Búsqueda con operadores lógicos boléanos

Al realizar la búsqueda en español, utilizando los operadores lógicos boléanos, los resultados fueron los siguientes, el buscador semántico recuperó 4 páginas de resultados, más de 100 sitios recuperados, tal como se muestra en las imágenes #3.115. y #3.116

Imagen #3.115 Búsqueda con operadores boléanos en español en DuckDukgo



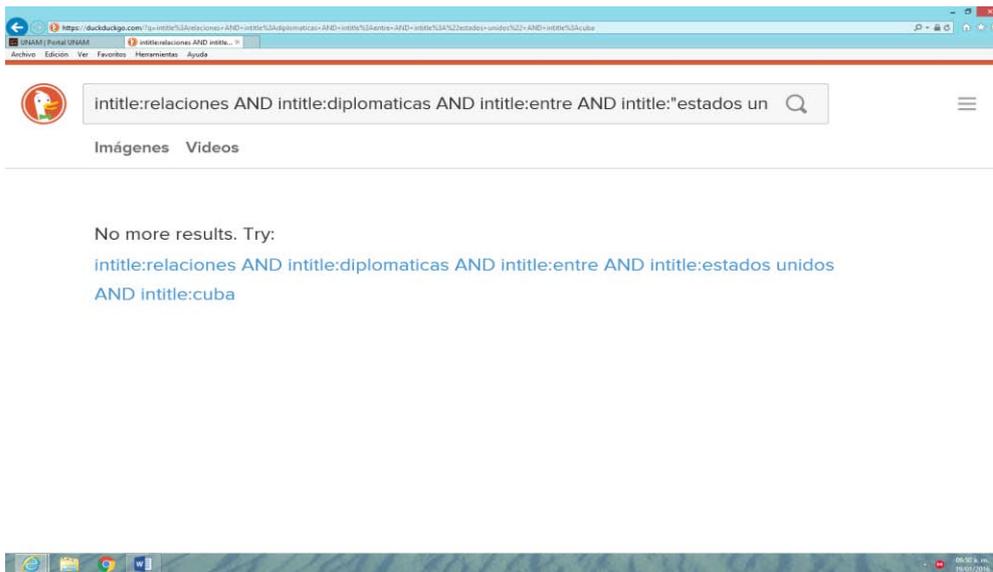
Imagen #3.116 Resultados visibles búsqueda con operadores boléanos en español en DuckDukgo



Búsqueda por campos

Al combinar operadores lógicos, comillas y realizar la búsqueda en el campo de título, no se obtuvieron resultados en el metabuscador, como se puede apreciar en la imagen #3.117.

Imagen #3.117 Búsqueda por campos en español en DuckDukgo



Búsquedas en inglés

Búsqueda libre

Con respecto a la búsqueda de información en inglés, el buscador semántico, recupero 4 páginas de resultados, más de 100 documentos, así lo demuestran las imágenes #3.118. y #3.119

Imagen #3.118 Búsqueda libre en inglés en DuckDukgo

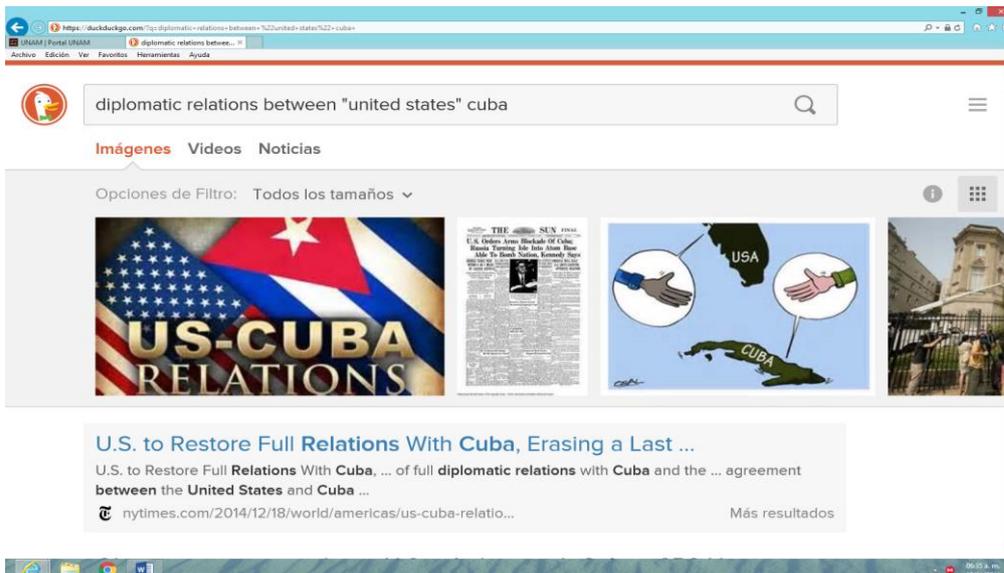
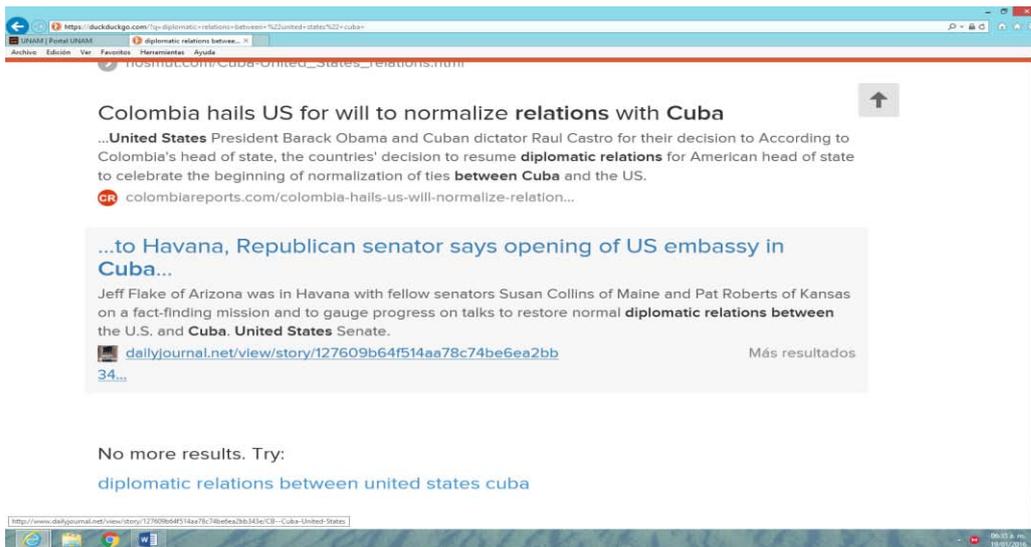


Imagen #3.119 Resultados visibles búsqueda libre en inglés en DuckDukgo



Búsqueda con operadores lógicos boléanos

Al realizar la búsqueda en inglés, los resultados fueron los siguientes, el buscador semántico recuperó 4 páginas de resultados, más de 100 sitios recuperados, tal como se muestra en las imágenes #3.120 y #3.121

Imagen #3.120 Búsqueda con operadores lógicos en inglés en DuckDukgo

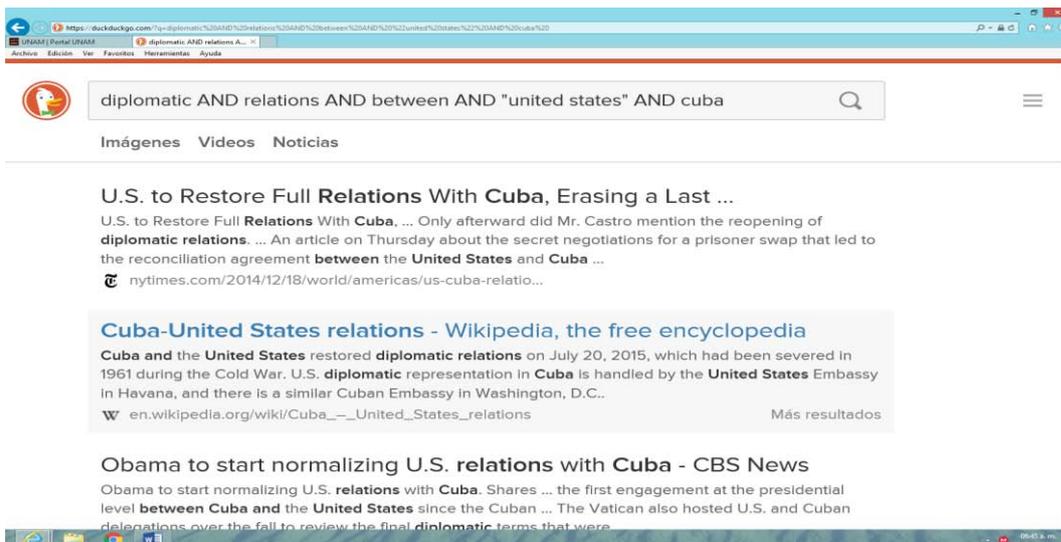
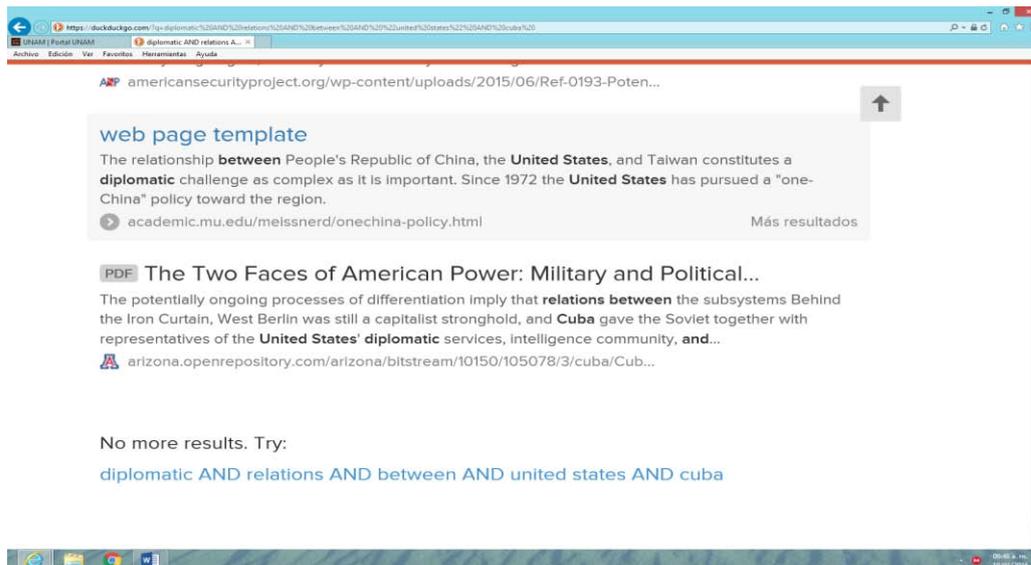


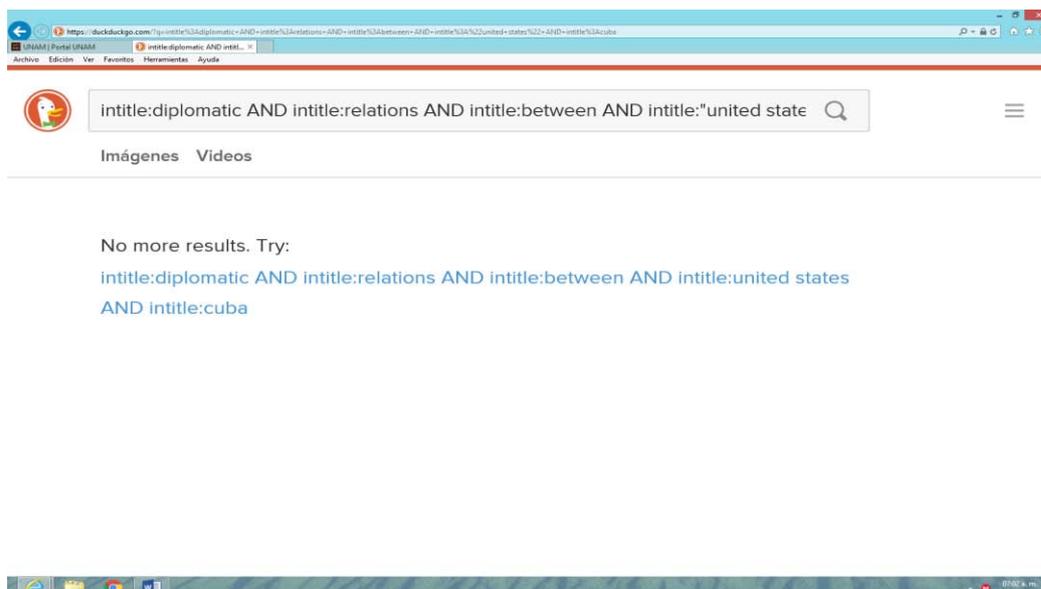
Imagen #3.121 Resultados visibles búsqueda con operadores lógicos en inglés en DuckDukgo



Búsqueda por campos

Al combinar operadores lógicos, comillas y realizar la búsqueda en el campo de título, no se obtuvieron resultados, como se puede apreciar en la imagen #3.122.

Imagen #3.122 Búsqueda por campos en inglés en DuckDukgo



3,6.15 Resultados de la recuperación de información en Swoogle

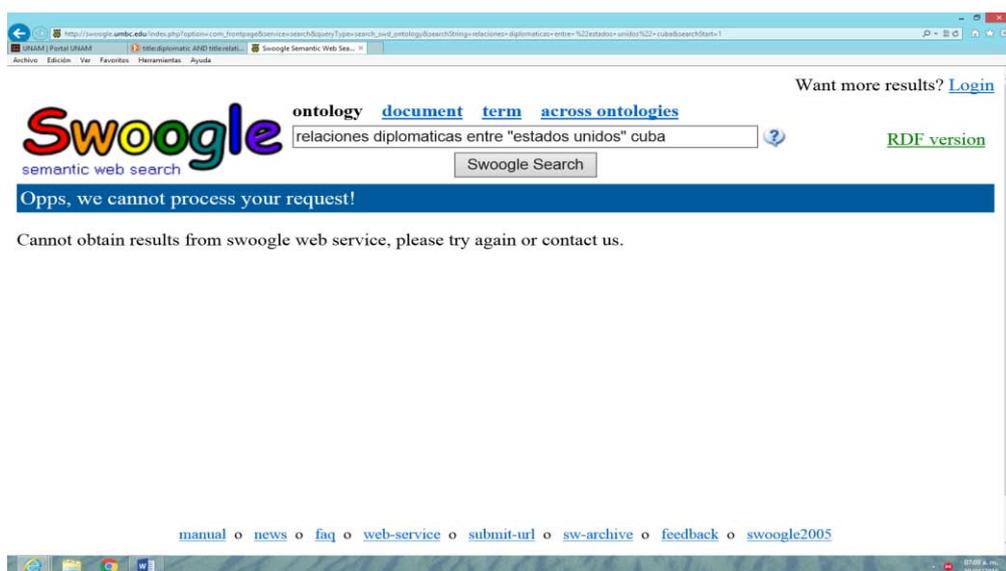
A continuación se presentan los resultados de la búsqueda de información en el buscador semántico Swoogle.

Búsquedas en español

Búsqueda libre

Se introdujo la frase completa en español en el buscador semántico y no recuperó resultados, como se puede apreciar en la imagen #3.123.

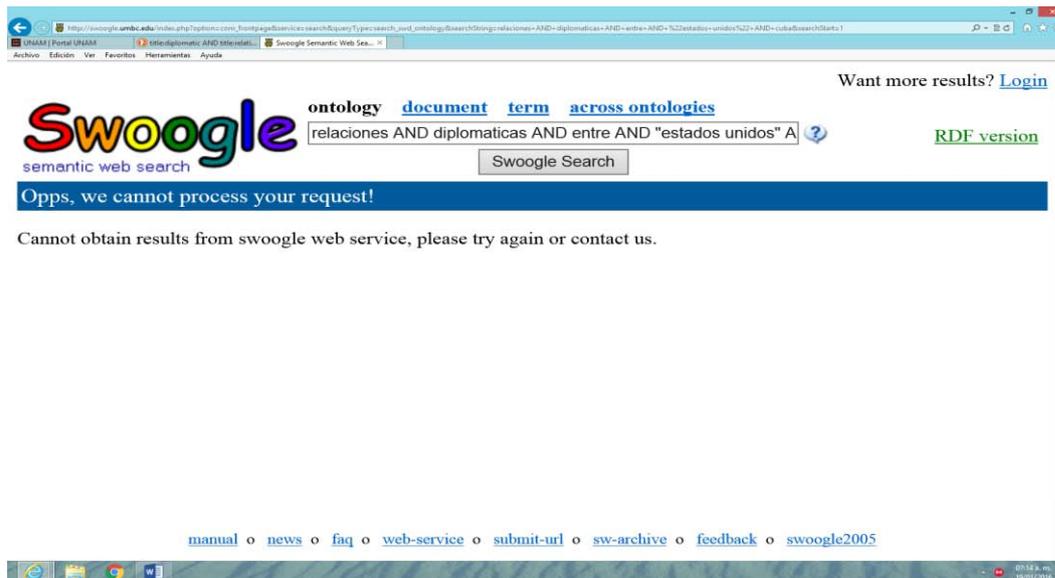
Imagen #3.123 Búsqueda libre en español en Swoogle



Búsqueda con operadores lógicos boléanos

Al realizar la búsqueda en español, utilizando los operadores lógicos boléanos, los resultados fueron los siguientes, el buscador semántico no recuperó documentos, tal como se muestra en la imagen #3.124.

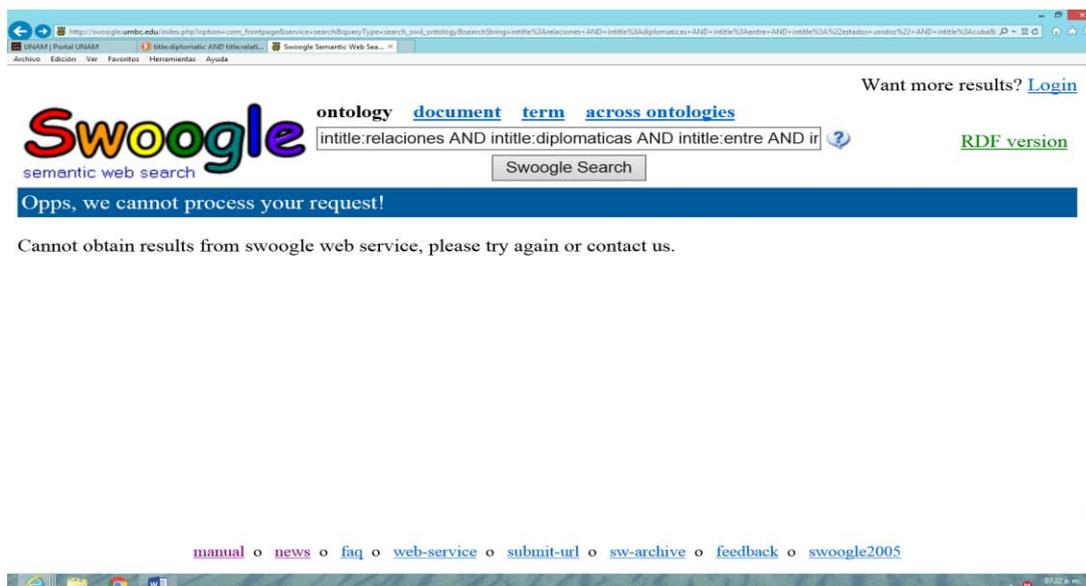
Imagen #3.124 Búsqueda con operadores lógicos en español en Swoogle



Búsqueda por campos

Al combinar operadores lógicos, comillas y realizar la búsqueda en el campo de título en español, no se obtuvieron resultados en el buscador semántico, como se puede apreciar en la imagen #3.125.

Imagen #3.125 Búsqueda por campos en español en Swoogle

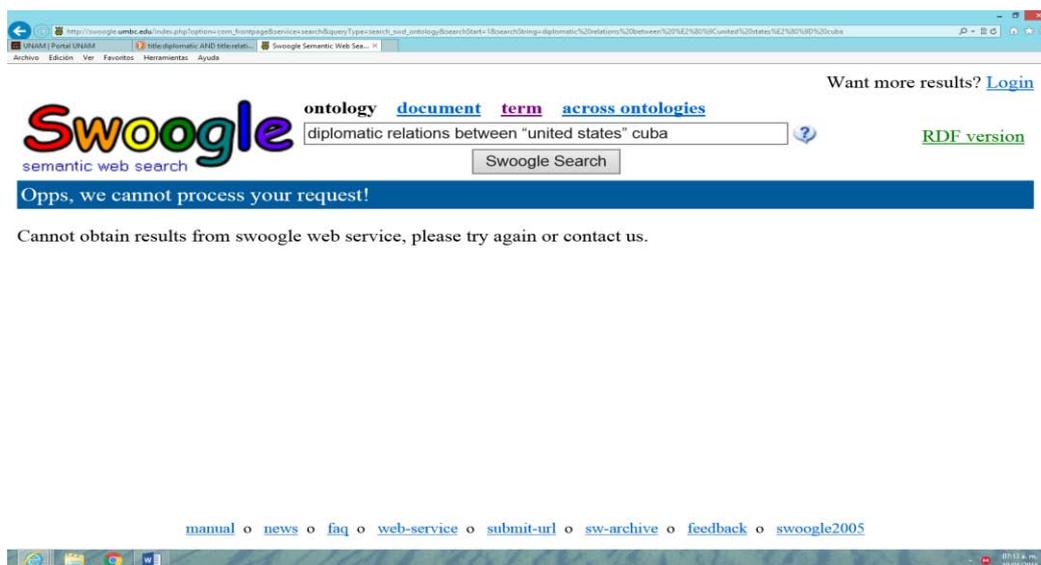


Búsquedas en inglés

Búsqueda libre

Con respecto a la búsqueda de información en inglés, el buscador semántico, no recupero documentos, así lo demuestra la Imagen #3.126.

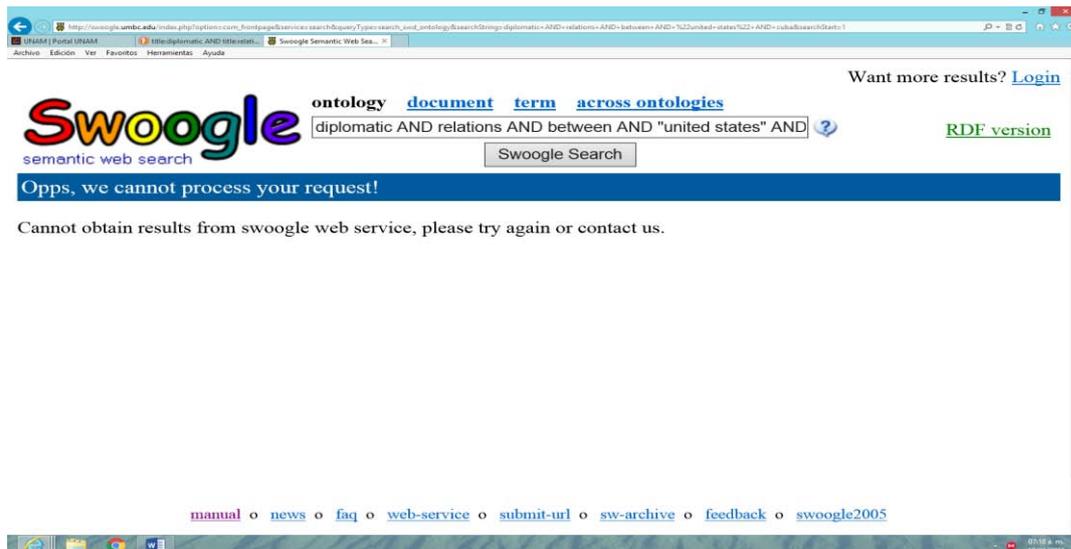
Imagen #3.126 Búsqueda libre en inglés en Swoogle



Búsqueda con operadores lógicos boléanos

Al realizar la búsqueda en inglés, los resultados fueron los siguientes, el buscador semántico no recuperó resultados, tal como se muestra en la imagen #3.127.

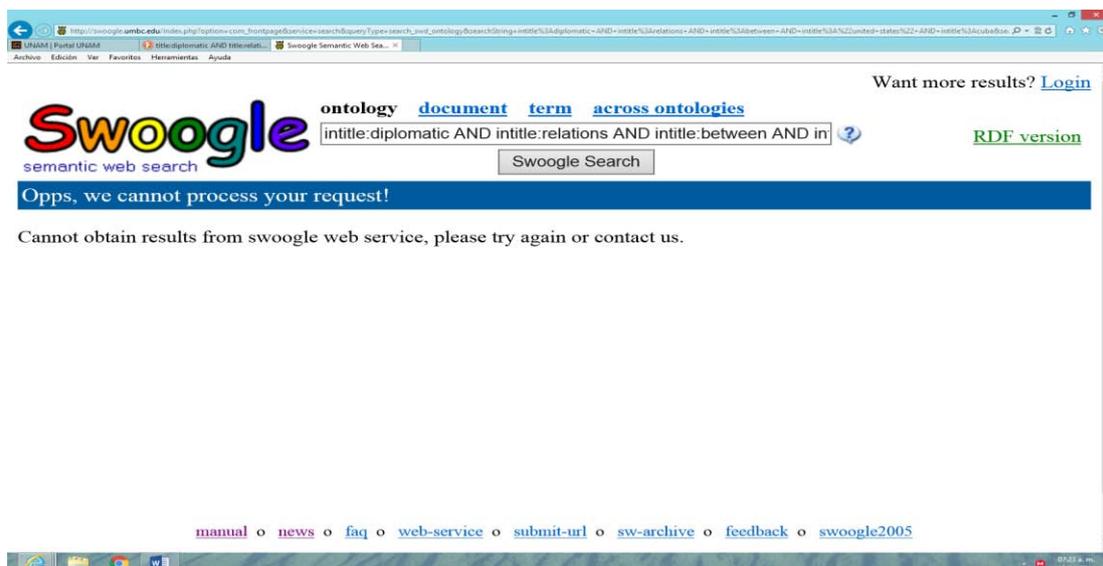
Imagen #3.127 Búsqueda con operadores lógicos en inglés en Swoogle



Búsqueda por campos

Al combinar operadores lógicos, comillas y realizar la búsqueda en el campo de título en inglés, no se obtuvieron resultados, como se puede apreciar en la imagen #3.128.

Imagen #3.128 Búsqueda por campos en inglés en Swoogle



3.7 Análisis de los resultados

Después de realizar el estudio, sobre la recuperación de información en diferentes herramientas de búsqueda, se encontraron una serie de situaciones que nos permiten entender la forma en la que buscan y recuperan información los motores, metabuscadores y buscadores semánticos más utilizados, que operan en la red.

A continuación se presentan dos tablas en las que se muestran los resultados de las búsquedas realizadas en español y en inglés de los motores de búsqueda seleccionados.

Tabla #3.11 Resultados en español motores de búsqueda

Motor	Libre	Operadores lógicos	Campo de titulo	Resultados visibles
Google	520, 000	457, 000	14	L= 331 O= 331 T= 13
Yahoo	145, 000	129, 000	5, 720	L= 183 O= 189 T= 121
Bing	2, 660,000	2, 290,000	1,860	L= 380 O= 423 T= 389
Ask	100	100	0	L= 100 O= 100 T= 0
AOL	608,000	462,000		L= 203 O= 186 T= 0

Fuente: elaboración propia

Tabla #3.12 Resultados en inglés motores de búsqueda

Motor	Libre	Operadores lógicos	Campo de título	Resultados visibles
Google	10,900,000	10,100,000	63	L= 352 O= 352 T= 27
Yahoo	7, 270, 000	606, 000	46, 800	L= 190 O= 200 T= 222
Bing	9, 960,000	4, 140,000	128	L= 445 O= 441 T= 126
Ask	100	100	0	L= 100 O= 100 T= 0
AOL	14, 300,000	12, 400,000		L= 192 O= 194 T= 0

Fuente: elaboración propia

En las dos tablas anteriores se puede ver claramente, que con respecto a las búsquedas libres, cuatro motores de búsqueda recuperaron una gran cantidad de información, solamente uno (Ask) recuperó en todas sus búsquedas solamente 100 resultados, con respecto a las búsquedas con operadores lógicos, aunque se redujeron los resultados, también se recuperaron una gran cantidad de documentos, finalmente en la búsqueda por el campo de título hubo una reducción significativa de los resultados. Por otro lado, hubo un aspecto importante que llamó la atención, al revisar los resultados de todas las búsquedas, es que cuatro motores solo permiten ver una cantidad muy pequeña de los resultados encontrados en la búsqueda, esto tal vez sea consecuencia de que no utilizan esquemas adecuados de metadatos como en los catálogos de una biblioteca, o también puede tratarse de una política del motor que no permite ver más de una cierta cantidad de documentos lo cual obliga al usuario a realizar varias búsquedas y estar más tiempo en el motor de búsqueda.

Metabuscadores

A continuación se presentan dos tablas en las que se muestran los resultados de las búsquedas realizadas en español y en inglés en todos los metabuscadores

Tabla #3.13 Resultados en español metabuscadores

Metabuscadores	Libre	Operadores lógicos	Campo de titulo	Resultados visibles
WebCrawler	730	670	0	L= 730 O= 670 T= 0
Dogpile	940	370	0	L= 940 O= 370 T= 0
Zapmeta	1, 650,000	33,800	0	L= 50 O= 50 T= 0
Info.com	300	320	0	L= 300 O= 320 T= 0
Exite	400	400	0	L= 400 O= 400 T= 0

Fuente: elaboración propia

Tabla #3.14 Resultados en inglés metabuscadores

Metabuscadores	Libre	Operadores lógicos	Campo de titulo	Resultados visibles
WebCrawler	430	300	0	L= 430 O= 300 T= 0
Dogpile	410	10	130	L= 410 O= 10 T= 130
Zapmeta	39, 600,000	1, 230,000	0	L= 50 O= 50

				T= 0
Info.com	400	380	10	L= 400 O= 380 T= 10
Exite	0	460	100	L= 0 O= 460 T= 100

Fuente: elaboración propia

Con respecto a las tablas de los metabuscadores, se observa que en cuatro de los metabuscadores, se recuperó una cantidad muy pequeña comparada con los motores de búsqueda, solamente en el caso del metabuscador Zapmeta que recuperó una gran cantidad de documentos en las búsquedas libre y con operadores lógicos, para el caso de la búsqueda en el campo de título no se obtuvieron resultados, por otro lado al revisar los documentos recuperados en las búsquedas cuatro metabuscadores, permitieron ver todos los documentos que habían recuperado en la búsqueda, el único metabuscador que solo permitió ver 50 resultados de los que había recuperado en la búsqueda fue Zapmeta.

Buscadores semánticos

A continuación se presentan dos tablas en las que se muestran los resultados de las búsquedas realizadas en español y en inglés en todos los buscadores semánticos.

.

Tabla #3.15 Resultados en español buscadores semánticos

Buscador Semántico	Libre	Operadores lógicos	Campo de titulo	Resultados visibles
Kngine	15	15	0	L= 15 O= 15 T= 0
Exalead	40,790	44,297	41	L= 40,790

				O= 44,297 T= 41
DuckDuckgo	100	100	0	L= 100 O= 100 T= 0
Swoogle	0	0	0	L= 0 O= 0 T= 0

Fuente: elaboración propia

Tabla #3.16 Resultados en inglés buscadores semánticos

Buscador Semántico	Libre	Operadores lógicos	Campo de titulo	Resultados visibles
Kngine	15	15	0	L= 15 O= 15 T= 0
Exalead	153,277	94,753	10	L= 153,277 O= 94,753 T= 10
DuckDuckgo	100	100	0	L= 100 O= 100 T= 0
Swoogle	0	0	0	L= 0 O= 0 T= 0

Fuente: elaboración propia

Con respecto a los buscadores semánticos, en las dos tablas se puede observar, que tres de ellos solamente recuperaron información en las búsquedas realizadas, el buscador semántico Swoogle no recuperó ningún resultado en todas las búsquedas que realizó, los buscadores Kngine y DuckDuckgo recuperaron la misma cantidad de documentos y finalmente el buscador Exalead fue el que recuperó la mayor cantidad de documentos en las búsquedas realizadas y permitió ver todos los documentos, aun tratándose de miles, esto se pudo constatar al revisar los resultados, debido a que este

buscador divide o agrupa en clusters a los documentos por año, esto permite que se puedan revisar todos.

Finalmente, se muestra en la siguiente tabla algunos aspectos de los resultados con respecto a los criterios de Lavene, de la forma en la que los motores presentan la información recuperada en las búsquedas.

Criterios de Lavene	Buscadores Google, Yahoo , Bing, Ask, AOL
Localización:	Permiten realizar búsquedas simples y también búsquedas utilizando diferentes campos como el de título.
Frecuencia de aparición	Recuperan y entienden como relevante que los términos de la búsqueda aparezcan en repetidas ocasiones en el texto de un documento
Popularidad:	Recuperan en primera instancia las páginas que cuentan con muchos enlaces.
Precio:	Al presentar los resultados siempre aparecen páginas que ofrecen servicios por que pagaron por aparecer en primera instancia.
Interfaz:	Presentan una gran cantidad de resultados, los cuales parecen ser muy relevantes, pero al revisar la búsqueda se encuentran documentos que no tienen nada que ver con la búsqueda.

Como se puede apreciar en la tabla anterior, los motores de búsqueda, en una búsqueda simple, es decir cuando un usuario teclea en la casilla de búsqueda del motor una o varias palabras, en los resultados efectivamente se aprecia que aparece una gran cantidad de ellos, de los cuales los primeros si cuentan con las palabras que el usuario solicitó, aunque los resultados no sean relevantes para la búsqueda y los demás resultados podrían no servir para nada en la búsqueda. Asimismo, se habla de una mayor relevancia de los documentos, cuando las palabras que se buscan aparecen principalmente en el título, en este sentido los motores de búsqueda permiten hacer combinaciones con diferentes operadores y diferentes campos, lo que ayuda mucho a que los términos que se utilizan en una búsqueda aparezcan en este campo, como podría ser el de título, al buscar por este campo, además de utilizar los

operadores lógicos y el uso de comillas, permiten recuperar o hacen que los motores presenten resultados más relevantes.

CONCLUSIONES

A lo largo de la investigación, se encontraron aspectos importantes, que permiten mencionarlos a manera de conclusiones.

La evaluación sobre el funcionamiento de diferentes herramientas de búsqueda y recuperación de la información que se encuentran en la Web, así como la identificación de si existe relación entre la recuperación de la información y el uso de metadatos permitió confirmar los supuestos planteados al inicio del trabajo, en donde se menciona que, "Si se utiliza una estructura de metadatos adecuada en la organización de la información, se obtiene un mayor índice de recuperación de la información". Sin embargo muchas de las herramientas de búsqueda de información de la Web no cuentan con estructuras de metadatos adecuadas, lo que dificulta la recuperación de la información, en particular la recuperación temática.

En tal sentido, como aspecto de alta importancia se menciona el incremento de información en Internet, es una muestra del desarrollo que han alcanzado las tecnologías de información, al crear herramientas que le permiten a los usuarios de la red poder buscar y recuperar información para sus investigaciones, además de permitirles un espacio en sus servidores para que puedan subir sus documentos de interés a la Web, ocasionando que se incremente el número de información día con día, sin ningún elemento para su organización.

Asimismo, existe un gran porcentaje de usuarios de la red, que representan un segmento de alta importancia por la función que desempeñan e identificados como los profesionales de la información, quienes contribuyen de forma asertiva en los procesos de búsqueda y recuperación de información. En líneas generales, poder superar las debilidades que surgen al momento de recuperar la información por causas de las herramientas de búsqueda, llámese motores de búsqueda, metabuscadores y buscadores semánticos, requiere de la

identificación de una serie de características, las cuales al ser resueltas facilitan la recuperar la información, siempre y cuando el usuario sea capaz de identificarlas y las solventa adecuadamente.

En tal sentido, las herramientas que han sido analizadas, cuentan con diferentes formas de realizar búsquedas en sus índices para recuperar información, por lo que se requiere claridad sobre la manera de presentar la información y de no ser correcta, falló la forma de utilizar las estructuras de metadatos adecuadas.

Es por ello, necesario que al momento de realizar búsquedas de información se recuperen una gran cantidad de documentos, haciendo muy difícil para el usuario seleccionar los de mayor pertinencia entre tantos.

De tal manera que los motores de búsqueda convencionales, son y seguirán siendo una puerta para acceder a la información que se encuentra en la red, pero desafortunadamente tienen limitantes que no les permiten recuperar información relevante que se encuentran en algunos sitios de Internet.

Con respecto, a los metabuscadores, al utilizar las plataformas de los motores de búsqueda, tienden a recuperar menos documentos, pero también sufren diferentes problemas en la recuperación, debido a la falta de bases de datos.

En el caso de los buscadores semánticos, solamente el buscador Exalead, fue el único que recupero información de calidad y en la presentación de resultados también fue el único que permitió visualizar todos los documentos que recupero.

Los resultados de la investigación presentada, conllevan a establecer las bases para realizar otros que complementen el supuesto de la existencia de

una relación entre el uso de estructuras de metadatos y la recuperación de información eficiente.

Finalmente, la labor del bibliotecólogo es muy importante, debido a que es el responsable de organizar información para su posterior recuperación, por esta razón es de vital importancia que conozca de manera adecuada todos los aspectos relacionados con los metadatos, así como de sus estructuras, para poder organizar y describir adecuadamente la información.

REFERENCIAS

- Aguillo Isidro. (2000). Internet invisible o infranet: definición, clasificación y evaluación. En: Jornadas Españolas de Documentación. Bilbao: Universidad del País Vasco.
- Asociación Mexicana de Internet (AMIPCI). (2013) Hábitos de los internautas en México. [en línea] Disponible en Internet: <http://www.amipci.org.mx/>
- Barrueco, J. M.; Subirats Coll, I. (2003). OAI-PMH: protocolo para la transmisión de contenidos en internet. El profesional de la información, marzo-abril, v. 12, n. 2, pp. 99-106.
- Bergman M. (2000) The deep web: Surfacing hidden value. [en línea] Disponible en Internet: <http://brightplanet.com/.../12550176481-deepwebwhitepa..>
- Berners-Lee, Tim. (1997). Metadata Architecture, [en línea] Disponible en Internet: <http://www.w3.org/DesignIssues/Metadata.html>
- Berners-Lee, Tim. (1998). Semantic web road map, [en línea] Disponible en Internet: <http://www.hipertexto.info/documentos/metadatos.htm>
- Berners-Lee, T.; Hendler, J.; Lassila, O., (2001). The semantic Web, [en línea] Disponible en Internet: http://www-sop.inria.fr/acacia/cours/essi2006/Scientific%20American_%20Feature%20Article_%20The%20Semantic%20Web_%20May%202001.pdf
- Bin He y otros (2004) Accessing the Deep Web: A Survey. [en línea] Disponible en Internet: http://www.researchgate.net/publication/238436479_Accessing_the_Deep_Web_A_Survey

Bright Planet (2000) The Deep Web: Surfacing Hidden Value. [en línea] Disponible en Internet: <http://128.121.227.57/download/deepwebwhitepaper.pdf>.

Budapest Open Access Initiative. (2014) [en línea] Disponible en Internet: <http://www.budapestopenaccessinitiative.org/>

Burnett K. (1997). Control or management: a comparison of the two approaches for establishing metadata schemes in the digital environment Rutgers University (New Jersey), Florida State University (Florida). [en línea] Disponible en Internet:<http://www.scils.rutgers.edu/~sympark/asis.html>

Burnett K, Kwong Bor Ng, Soyeon P. A. (1999). Comparison of the two traditions of metadata development. JASIS (Special Topic Issue: Integrating Multiple Overlapping Metadata Standarts) (13):1209-17. [en línea] Disponible en Internet: <http://www.sims.monash.edu.au/subjects/ims2603/resources/week10/10.1.pdf>

Cabero Almenara, Julio. (1996). Nuevas tecnologías, comunicación y educación. En Revista electrónica de tecnología. No. 1. . [en línea] Disponible en Internet: <http://edutec.rediris.es/Revelec2/Revelec1/revelec1.html>

Caplan, Priscilla (1995). You call it corn, we call it syntax-independent metadata for documentlike objects. The Public Access Computer Systems Review, v. 4, n. 6. [en línea] Disponible en Internet: <http://epress.lib.uh.edu/pr/v6/n4/capl6n4.html>.

Codina L. & Rovira, C. (2006). La web semántica. pp.1-34 [en línea] Disponible en Internet: http://eprints.rclis.org/8899/1/web_semantica__.pdf

Córdoba González, Saray (2010) Los repositorios institucionales y de acceso abierto en Costa Rica. [en línea] Disponible en Internet: www.kerwa.ucr.ac.cr/bitstream/handle/10669/449/PONENCIA.pdf?

Data Never Sleeps How much Data is Generated Every Minute. (2012) [en línea] Disponible en Internet: <http://www.domo.com/blog/2012/06/how-much-data-is-created-every-minute/>

Daudinot Fournier , Isabel (2006) Organización y recuperación de información en Internet: teoría de los metadatos. [en línea] Disponible en Internet: <http://eprints.rclis.org/9246/>

Desire Project. 2000. [en línea] Disponible en Internet: <http://www.desire-project.eu/>

Directory of Open Access Journals. (2014) [en línea] Disponible en Internet: <http://www.doaj.org/>

Doan, Bich-Lien; Beingbeder, Michel; Girardot, Jean-Jacques; Jaillon, Philippe, (1998). Using metadata to improve organization and information retrieval on the WWW. [en línea] Disponible en Internet: <http://files.eric.ed.gov/fulltext/ED427698.pdf>

Domínguez, Lorena Mariel. (2008). Los usuarios, Internet y el bibliotecario: estudio de usuarios sobre los recursos en línea. *Inf. cult. soc.*, n.18, [en línea] Disponible en Internet: http://www.scielo.org.ar/scielo.php?script=sci_abstract&pid=S1851-17402008000100004&lng=pt&nrm=iso

Dublin Core Metadata Initiative. (2015). [en línea] Disponible en Internet: <http://dublincore.org/>

E-revistas, Plataforma open access de revistas científicas electrónicas españolas y latinoamericanas. (2014) [en línea] Disponible en Internet: <http://www.erevistas.csic.es/>

eBizMBA (2016) Top 15 Most Popular Search Engines. [en línea] Disponible en Internet: <http://www.ebizmba.com/articles/search-engines>

Fernández Laura, Pineda Enric, Mayayo Teo. (2005) Internet para farmacéuticos: navegando en la red farmacéutica. En Formación continuada para farmacéuticos de hospital. España, Ferrer Grupo. p. 68-106

Genealogy of Science According to Scopus. (2011) [en línea] Disponible en Internet: <http://www.wired.com/2011/03/best-science-maps>

Gilliland Swetland, Anne J. (2000). Introduction to metadata: Setting the stage. . [en línea] Disponible en: <http://ptarpp2.uitm.edu.my/ptarpprack/silibus/is772/SetStage.pdf>

Gómez, Laureano Felipe. (2007). Interoperabilidad en los Sistemas de Información Documental (SID): la información debe fluir. En Revista Códice Vol. 3 N° 1: 23-39 / Enero - junio.

Guajardo Salinas, Aldo. (2010). Z39.50 y OAI-PMH: Protocolos de Transferencia y Recuperación de Información.) [en línea] Disponible en Internet: www.bibliotecarios.cl/descargas/2010/11/guajardo_z3950.pdf

Hopkins, Judith, (2000) USMARC como una estructura de metadatos. En: Internet, metadatos y acceso a la información en bibliotecas y redes en la era electrónica, México, UNAM-CUIB, 112p.

Howe , D. Free on-line Dictionary of Computing (FOLDOC). 1993. [en línea] Disponible en: <http://wombat.doc.ic.ac.uk/foldoc/>

Kaiser Family Foundation. Generación Rx.com: Cómo usan los Jóvenes el Internet para obtener Información sobre Salud. [en línea] Disponible en Internet: <http://kff.org/>

Kashyap V, Sheth A. (2000). Information Brokering Across Heterogeneous Digital data: A Metadata –based approach. Boston : Kluwer Academic

Publishers. pp. 241. [en línea] Disponible en:
<http://www.nongnu.org/babe/papers/kluwerbook.pdf>

Lamarca Lapuente, María Jesús. (2013) Hipertexto: El nuevo concepto de documento en la cultura de la imagen. [en línea] Disponible en Internet:
<http://www.hipertexto.info/index.htm>

Lancaster, FW; Pinto, M. (2001). “Hacia la modelización del documento científico” en Procesamiento de la información científica. Madrid: Arco/libros, p. 41-67

Laney, Doug. (2011) Robust metadata representation standards emerge to tackle information integration troubles. [en línea] Disponible en Internet:
http://www.carnation.com.cn/knowledge_show.php?id=35&type_news=2&menuid=7

Lara Navarra, Pablo; Martínez Usero, José Ángel (2006) Agentes inteligentes en la búsqueda y recuperación de información. p. 61 [en línea] Disponible en Internet:
<https://core.ac.uk/download/files/418/11880151.pdf>

Levene, Mark (2010) An introduction to search engines and Web navigation. [en línea] Disponible en Internet:
<https://luwj.googlecode.com/svn/trunk/Ebook/An%20Introduction%20to%20Search%20Engines%20and%20Web%20Navigation%202nd.pdf>

Listofsearchengines (2016) Top 10 Search Engines List. [en línea] Disponible en Internet: <http://www.listofsearchengines.org/>

Lopez Guzman, Clara; García Peñalvo, Francisco; Pernias Peco, Pedro.(2005). Desarrollo de repositorios de objetos de aprendizaje a través de la reutilización de los metadatos de una colección digital: de Dublin Core a IMS. En RED. Revista de Educación a Distancia. [en línea] Disponible en: <http://www.um.es/ead/red/M2/lopez27.pdf>

Mejía Argueta, Miguel Angel G. (2012). La interoperabilidad y el intercambio de los metadatos en la red. En Revista Digital Universitaria. Vol. 13, No.9. [en línea] Disponible en Internet:
<http://www.revista.unam.mx/vol.13/num9/art95/art95.index.html>

Méndez Rodríguez, Eva María. (2002). Metadatos y recuperación de información: estándares, problemas y aplicabilidad en bibliotecas digitales. Gijón : Ediciones Trea, D.L.

Michán, Layla (2009), "Las revistas y la institucionalización de la sistemática en América Latina," Revista Mexicana de Biodiversidad 80 (1): 105-117.

Ministerio de telecomunicaciones y Sociedad de la Información de Ecuador. (2014) En el Ecuador 60 de cada 100 usuarios usan Internet una vez al día. [en línea] Disponible en Internet:
<http://www.telecomunicaciones.gob.ec/en-el-ecuador-60-de-cada-100-usuarios-usan-internet-una-vez-al-dia/>

Moyano Grimaldo, Wilmer. (2004). Sociedad de la Información: metadatos y futuro de la Internet en la recuperación de información de calidad. En Bibliotecas & Tecnologías de la Información, vol. 1, n. 1. [en línea] Disponible en Internet:<http://eprints.rclis.org/7010/>

Perfil de Metadatos IDEBA, (2011) [en línea] Disponible en Internet:
http://www.gobierno.gba.gov.ar/subsecretarias/ideba/docs/perfil_de_metadatos_IDEBA.pdf

PubMed-Medline. (2014) [en línea] Disponible en Internet:
<http://www.ncbi.nlm.nih.gov/pubmed>

Reliablesoft (2016) Top 10 search engines in the World. [en línea] Disponible en Internet: <https://www.reliablesoft.net/top-10-search-engines-in-the-world/>

Rovira, Cristòfol. Marcos Mari-Carmen. (2006). Metadatos en revistas—e de Documentación de libre acceso. El profesional de la información, v. 15. N.2 marzo-abril. pp. 136-144

ScienceDirect. [en línea] Disponible en Internet: <http://www.sciencedirect.com/>

Sherman, Chris, Price, Gary. The invisible Web. Searcher. 2001; 8(9):62-74.

Senso José, Rosa Piñeiro, Antonio de la. (2003) El concepto de metadato. Algo más que descripción de recursos electrónicos. Pp. 95-106 [en línea] Disponible en Internet: <http://www.scielo.br/pdf/ci/v32n2/17038.pdf>.

Studer, S.; Benjamins, R.; Fensel, D. (1998) “Knowledge engineering: principles and methods”.pp. 161-198. [en línea] Disponible en Internet: <http://citeseerx.ist.psu.edu/viewdoc/download;jsessionid=51BD89B29D495B74B006AB6B19B45FE8?doi=10.1.1.41.1007&rep=rep1&type=pdf>

Tendencias digitales. (2011) Usos de Internet en Latinoamérica. [en línea] Disponible en Internet: <http://tendenciasdigitales.com/>

Taylor, Arlene. G. The organization of information.(1999) LIBRARIES UNLIMITED, Englewood, Colorado, pp. 417.

El universo digital: ¿Cuánta información existe en Internet?. (2010) [en línea] Disponible en Internet: <http://w5.cl/2011/09/14/el-universo-digital-%C2%BFcuanta-informacion-existe-en-internet/>

Valle Gastaminza Del, Félix; Sánchez Jiménez, Rodrigo; Pérez Agüera, José Ramón, (2005) Descripción e identificación de recursos en internet: metadata. En Revista Digital Universitaria. 2005, Vol. 6, No.7.pp. 1-12 [en línea] Disponible en Internet: http://www.revista.unam.mx/vol.6/num7/art72/jul_art72.pdf

Valle Gastaminza Del, Felix. (2005) Los buscadores y la indización tradicional. [en línea] Disponible en Internet: <http://www.revista.unam.mx/vol.6/num7/art72/art72-2b.htm>

Velásquez Paulus, Cristian. (2005). Metadatos: introducción e historia. [en línea] Disponible en Internet: <http://users.dcc.uchile.cl/~cvasquez/introehistoria.pdf>

Won, Kim (2005) On metadata management technology: Status and issues. pp. 41-47 [en línea] Disponible en Internet: http://www.jot.fm/issues/issue_2005_03/column4/column4.pdf

RDF Working Group. (2014). Resource Description Framework (RDF). [en línea] Disponible en Internet: <http://www.w3.org/RDF/>