

Tema 3

Recuperación de la información en Internet

1.- Herramientas de búsqueda

Ya se han explicado las estrategias que se pueden utilizar para afrontar una búsqueda. En este tema veremos con qué herramientas contamos para acceder al universo Internet.

Internet ofrece miles de millones de páginas web con enorme cantidad de información de muy variado tipo. Esta información es dinámica y volátil, debido a que, a diferencia de otras tecnologías, Internet posibilita que la información se cambie en cualquier momento. La cantidad y variedad de información disponible en Internet determina la necesidad de contar con ciertas herramientas para obtener información que resulte significativa, es decir, útil, relevante y confiable. Para ello es necesario que al iniciar un proceso de búsqueda se consideren los siguientes aspectos:

- Conocimiento de los recursos involucrados: es decir, características de la red, programas de navegación, de administración de archivos y otras utilidades.
- Conocimiento de los sitios de búsqueda y sus respectivas estrategias de búsqueda.

Tramullas y Olvera¹ aluden a la complejidad de la recuperación de la información en Internet debido al dinamismo que tiene, a la hipertextualidad o a factores externos relacionados con las técnicas de recuperación. El proceso de indización automático que rige en Internet supone el que no todo lo recuperado sea preciso, ni fiable, ni siquiera riguroso.

En el universo de Internet no toda la información es homogénea, no todo lo que resulta de una búsqueda convencional es pertinente, sino que además puede recuperarse mucho ruido.

Conviene saber que, según el acceso, se puede hablar de varios tipos de Internet:

Por un lado está **Internet global**, libre y totalmente accesible mediante navegadores, chats, mensajes, intercambio de archivos. La red que todos conocemos.

Una **Internet invisible** en la que la información disponible solo es accesible a través de páginas generadas dinámicamente al realizar una consulta en una base de datos. Los procesos habituales de recuperación de información no sirven para acceder a la misma. La información recuperada en ella es de una calidad superior a la de Internet normal. Finalmente, **Internet oscura** totalmente inaccesible desde nuestro ordenador debido a motivos de seguridad nacional o militar.

¹ Tramullas Sanz, J.; Olvera Lobo, M.J. *Recuperación de la información en Internet*. Madrid: RaMa, 2000.

Tramullas y Olvera señalan una serie de limitaciones que atienden a:

- La disfunción que se da entre los procesos de indización automática, la representación del contenido del documento y el contenido mismo.
- La cobertura que ofrecen los motores de búsqueda no es exhaustiva.
- La actualización de las bases de datos no es automática. Los robots de indización visitan a intervalos cada vez mayores las páginas indizadas en sus bases de datos priorizando las visitas a aquellas más solicitadas.
- Los motores no reflejan la variabilidad espacial y temporal de las páginas web, dando errores debido al cambio de localización de algunas páginas.
- A veces se representan páginas de forma aislada sin atender a su pertenencia a estructuras mayores, consecuencia de la estructura hipertextual.
- Cada motor de búsqueda tiene sus particularidades en relación a los operadores con los que funciona.
- Las respuestas ofrecidas responden a las ecuaciones planteadas.

Los recursos que ofrece Internet para solucionar las necesidades de información de los usuarios pueden tener de manera general tres formas: motores de búsqueda, directorios y portales.

El usuario navega por unas colecciones de documentos que se sustentan merced a unas estructuras auxiliares de información

2.- Motores de búsqueda

Son las herramientas de consulta que permiten interrogar a Internet para localizar la información que se necesita. Están diseñados para facilitar el hallazgo de determinada información en internet. Tienen otras denominaciones como rastreadores, webcrawlers, agentes, etc.

Básicamente se componen de una base de datos, un programa de indización, el robot de búsqueda y la interfaz.

El robot lógico o “araña”, (spider) rastrea y compila grandes cantidades de páginas web que almacena en gigantescas bases de datos. Dichos recursos son almacenados y se someten a procedimientos informáticos que asignan, sin intervención humana alguna, una serie de términos representativos del contenido. Es a partir de esos términos lo que utiliza el motor como punto de acceso en posteriores búsquedas.

Los motores de búsqueda precisamente se diferencian de otras herramientas de búsqueda tales como los índices temáticos o directorios, en que las páginas almacenadas en aquellos no están clasificadas temáticamente, ya que el trabajo de almacenamiento lo realizan los robots, por lo que no se pueden realizar búsquedas jerarquizadas. La recuperación refleja las páginas con las primeras palabras o el párrafo donde aparecen.

Los buscadores contienen en sus bases de datos millones de páginas web, y por tanto el rango de resultados es importante. Los buscadores utilizan un algoritmo para ordenar los resultados. Aún cuando los programas lleguen a ser similares, no existen

dos programas de búsqueda exactamente similares en términos de tamaño, velocidad y contenido; no existen dos motores de búsqueda que utilicen de manera coincidente el mismo listado de relevancia y tampoco cada motor de búsqueda ofrece las mismas opciones de búsqueda. Por tanto la recuperación es diferente en cada motor utilizado. Aunque la diferencia podría no ser mucha, si resultaría significativa.

El objetivo del buscador es que aparezcan en primer lugar las páginas más pertinentes de acuerdo a los criterios de búsqueda. Hay que tener en cuenta que el usuario sólo va a detenerse en mirar las primeras páginas. Uno de los criterios es la aparición de los términos de la pregunta (búsqueda) en el documento. Si la palabra clave es muy común o tiene varios significados se pueden recuperar muchas referencias irrelevantes, o lo que es lo mismo, obtener ruido.

A este respecto, por ejemplo Google ha desarrollado el PageRank, criterio por el cual una página es más valorada cuanto mayor sea el número de otras páginas que enlazan con ella. Lo revolucionario es que PageRank determina que han de aparecer en primer lugar aquellas páginas que además de contener los criterios expresados en la búsqueda, tienen más enlaces, lo que las hace más relevantes.

Elementos de un buscador:

1. El programa robot 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 actúa de pasarela entre el servidor de documentos html y la base de datos.

El interfaz recibe la consulta del usuario (query), formada por la expresión de uno o más términos, realiza una consulta interna en la base de datos que contiene los recursos web indexados y ofrece una lista de aquellos recursos que cumplen una parte o el total de los requisitos establecidos en la consulta. Generalmente, los resultados aparecen ordenados según una puntuación (score) que el programa asocia automáticamente a cada recurso.

Para realizar una consulta es necesario tener en cuenta un conjunto de variables:

- a) Lenguaje de interrogación con la posibilidad de usar diferentes operadores.
- b) Posibilidad de refinar la búsqueda inicial.
- c) Campos limitadores que permitan reducir la búsqueda.
- d) Diversas alternativas de búsqueda: simple, avanzada...
- e) Funcionalidades: buscar diferentes recursos (texto, sonido, imagen), guardar y reutilizar búsquedas, diferentes formatos en los resultados de búsqueda.

2.1.- Principales motores de búsqueda:

- **Yahoo! Search** <http://www.alltheweb.com/>



- **MSN Search** que ha sido suplantado por **Bing** <http://www.bing.com/>



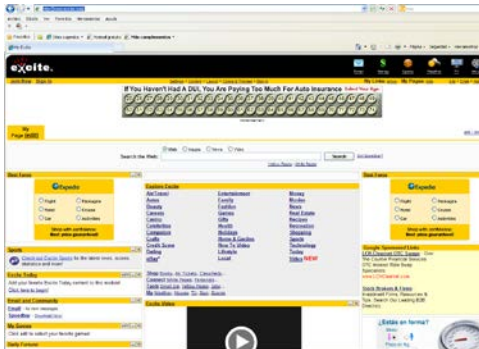
- **Ask Jeeves** <http://www.ask.com>



- **Blippex** <https://blippex.org/>
Interviene el factor humano ya que ordena los resultados basándose en la cantidad de tiempo que los usuarios emplean visitando cada sitio web o, lo que

es lo mismo, tras haber hecho click en cada enlace ofrecido. Sin embargo, esto no funciona si no se descarga la extensión que se encarga de cosechar este tipo de datos de manera anónima. Es uno de los buscadores más innovadores de los últimos tiempos.

- **Excite** <http://www.excite.com/>



- **Gibiru** <http://gibiru.com/>
Tiene la particularidad de que añade contenido censurado que no es mostrado por los buscadores en general. Mantiene la absoluta privacidad en las búsquedas.
- **Lycos** <http://www.lycos.com/>



- **Google** <http://www.google.es>

Después de diversos desarrollos en los motores de búsqueda, puede decirse que hoy se concentran en Google la mayor parte de las búsquedas realizadas. Es actualmente el más importante. Su origen comienza en 1995 de la mano de Larry Page y Serge Brin, alumnos de la Universidad de Standford (EEUU), que trabajaban en la creación de un buscador. En 1998 fundan la empresa Google Inc. No han parado de perfeccionarlo.

Google² es uno de los buscadores con mayor número de URL almacenadas en sus bases de datos. Además de español, dispone de ayudas en otros 100 idiomas.

²En esta dirección web del Observatorio tecnológico dependiente del Mº de Cultura encontraréis los desarrollos de Google muy bien explicados. Revisadla y leedla

El formulario de búsqueda simple, de sobra conocido, presenta una única ventana. Simplemente hay que escribir en el cuadro de búsqueda algunos términos clave que representan el tema que interesa. Se aconseja delimitar la búsqueda con palabras añadidas que remitan a una época, lugar, etc. No es conveniente emplear términos vacíos de significado como el caso de artículos, conjunciones. Pero para obligar a incluir esos términos en la búsqueda hay que utilizar el signo más (+), o sea si queremos un resultado exacto. O también, si queremos que busque resultados exactos del término escrito, hay que ponerlo entre comillas. El operador por defecto es AND, y si es preciso excluir términos se utiliza el signo menos (-) colocado inmediatamente antes del término que queremos excluir. El operador OR funciona bien en la Búsqueda avanzada.

Google solo encuentra páginas que contienen las palabras utilizadas en la consulta, de modo que agregando más términos a la misma, se concreta la búsqueda.

No reconoce paréntesis. No distingue mayúsculas de minúsculas. Se puede colocar el signo más (+) delante del término si queremos un resultado exacto.

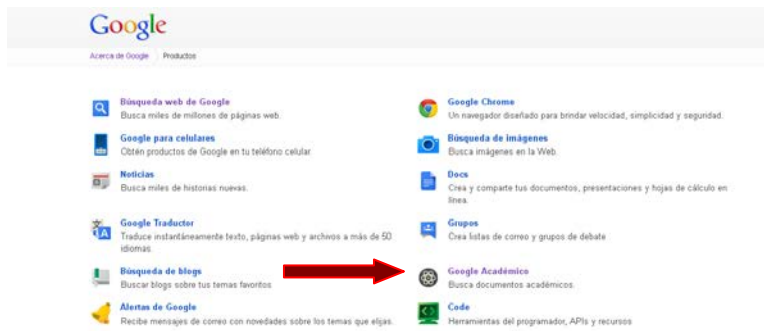
En la Búsqueda avanzada de Google http://www.google.es/advanced_search

Se puede acotar por idioma, dominio, periodo y tipo de fichero.

Los resultados de las búsquedas contienen el título de las páginas, dos líneas de texto donde aparecen los términos buscados resaltados y la dirección exacta de la búsqueda.

Google está continuamente introduciendo nuevos servicios. Queremos destacar el servicio *Google Académico* que permite el acceso a documentación de calidad: artículos de revistas, tesis, libros, repositorios académicos. Es la versión en español del Google Scholar.

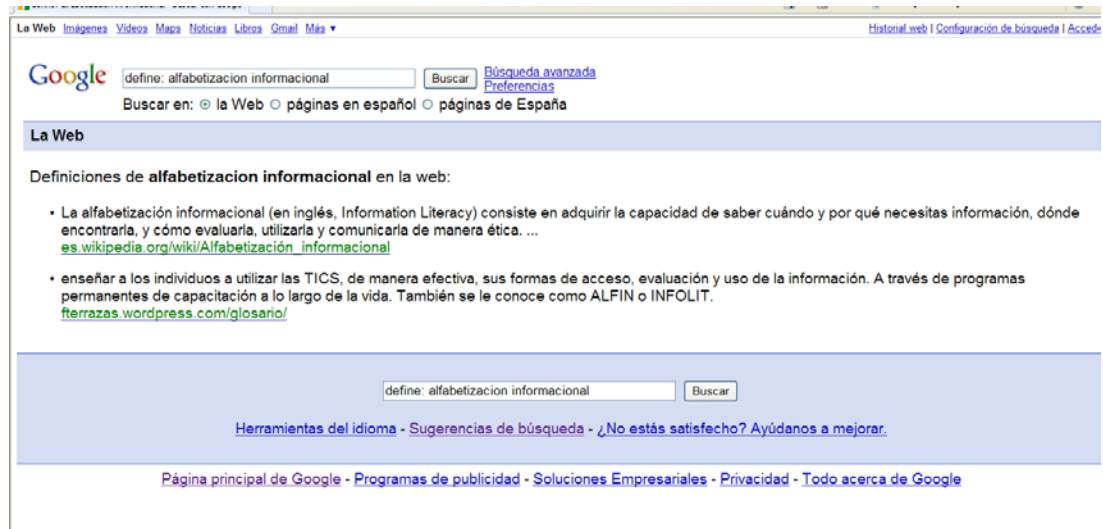
bien..<http://recursostic.educacion.es/observatorio/web/es/internet/recursos-online/1004-busquedas-avanzadas-en-google>



Para buscar un autor del que sabemos el nombre, lo mejor es introducir el operador *autor:* y el nombre completo, a continuación.

Una opción interesante que presenta Google es la de buscar definiciones. El procedimiento consiste en escribir en la caja de búsqueda el operador *define:* y el concepto que quiere buscarse.

define:alfabetización informacional



Con el término *intitle:* utilizado como operador, solo devuelve resultados en los que el término buscado, que debemos colocar a continuación, se encuentren en el título.
Intitle:bibliotecas

The screenshot shows a Google search interface. The search bar contains the text 'intitle: bibliotecas'. Below the search bar, the results are displayed. On the left side, there is a sidebar with various filters: 'Todo', 'Imágenes', 'Maps', 'Videos', 'Noticias', 'Shopping', 'Blogs', 'Libros', 'Más', 'León', 'Cambiar ubicación', 'La Web', 'Páginas en español', 'Páginas de España', 'Páginas extranjeras traducidas', 'Cualquier fecha', 'Última hora', 'Últimas 24 horas', 'Últimos 2 días', 'Última semana', 'Último mes', 'Último año', and 'Intervalo personalizado...'. The main results area shows several links related to libraries, including 'SUBDIRECCIÓN GENERAL DE BIBLIOTECAS. Co...', 'Bibliotecas españolas y de todo el mundo - Página...', 'BIBLIOTECAS PUBLICAS', and 'Bibliotecas Virtuales - Comunidad Literaria y Textos...'. The search results are approximately 13,200,000 results in 0.17 seconds.

Se pueden hacer búsquedas especificando que los resultados que nos muestre pertenezcan a un sitio web concreto o a un dominio determinado. Se utiliza el operador *site*:

Por ejemplo queremos buscar información en las páginas web del diario el País (diario digital) sobre “*crisis financiera*”. Sería por ejemplo: */crisis financiera site: elpais/*

También rastrea en dominios, por ejemplo: *.org, .es, .fr*.

Podemos indicar una búsqueda con los mismos términos pero realizada sólo en el dominio *.es*. Sería: */crisis financiera site:.es/*

The screenshot shows a Google search interface. The search bar contains the text 'crisis financiera site: elpais.es'. Below the search bar, the results are displayed. On the left side, there is a sidebar with various filters: 'Todo', 'Imágenes', 'Maps', 'Videos', 'Noticias', 'Shopping', 'Foros de debate', 'Libros', 'Blogs', 'Más', 'Buscar en la Web', 'Buscar sólo páginas en español', 'Páginas extranjeras traducidas', 'Todos los resultados', 'Búsquedas relacionadas', and 'Más herramientas'. The main results area shows several links related to the financial crisis, including 'Miquel Freijo - Entrevistas Digitales en EL PAIS', 'PDF "La palabra" toma Gijón de nuevo', 'Entrevistas Digitales en EL PAIS', 'EL PAIS: el periódico global en español', 'Cibern@is: Tecnología en EL PAIS', and 'Los afectados por la estafa filatélica'. The search results are approximately 202 results in 0.25 seconds.

Google permite localizar imágenes, videos seleccionándolos en el vínculo incluso se puede hacerlo con una búsqueda más precisa. Además cuenta con un amplio catálogo de libros de diversos temas y en distintos idiomas. Algunos muestran un número limitado de páginas impuesto por los derechos de autor, otros tienen acceso completo.

2.2 Motores de Búsqueda Académicos

Los buscadores especializados indexan de forma automática o semiautomática contenidos que están alojados en ciertos sitios o dominios de la web. No son tan rigurosos ni precisos como las bases de datos

Academia.edu <https://www.academia.edu/> es una plataforma para que los académicos puedan compartir sus trabajos de investigación. Los académicos utilizan a Academia.edu para compartir sus investigaciones, supervisar los profundos análisis de todo el impacto de sus investigaciones, y realizar un seguimiento de las investigaciones que los académicos siguen. Hay más de 33 millones de académicos inscritos, y que han añadido más de 10 millones de artículos y casi 2 millones de investigaciones.

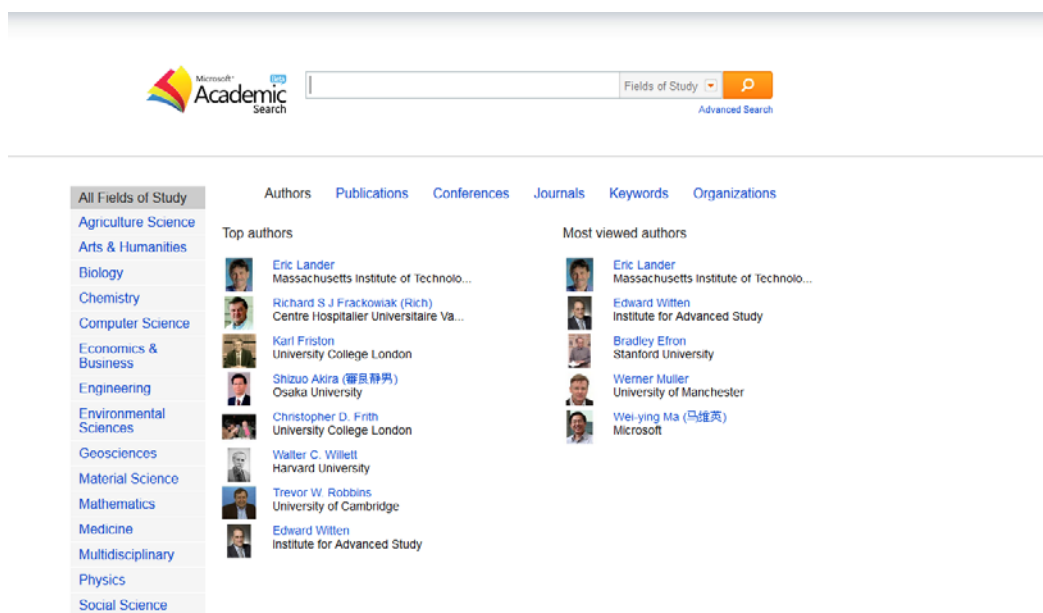
ERIC Institut of Education Science <https://eric.ed.gov/>

BASE (Bielefeld Academia Search Engine) <https://www.base-search.net/>
Recursos web académicos de acceso abierto, desarrollado por la Biblioteca de la Universidad de Bielefeld. BASE ofrece más de 80 millones de documentos de más de 4.000 fuentes. Se puede acceder a los textos completos de aproximadamente el 60-70% de los documentos indexados.

Google Scholar <https://scholar.google.es/> (Google Académico) Se puede buscar artículos, tesis, libros, resúmenes y opiniones... de las editoriales académicas, sociedades profesionales, depósitos en línea, las universidades y otros sitios web. Google Scholar ayuda a encontrar trabajos relevantes dentro del mundo de la investigación académica.

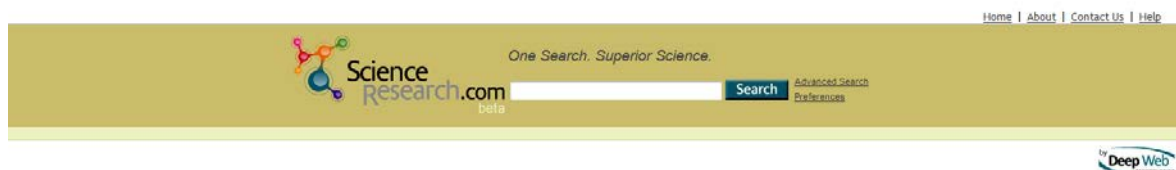
iSEEK Education <http://education.iseek.com/#/education> Motor de búsqueda específico que recopila cientos de miles de recursos de autoridad de la universidad, el gobierno y los proveedores no comerciales establecidos.

Microsoft ha tenido intentos de desarrollar un motor académico, de no muy buen resultado, como *Microsoft Academic Search* <https://academic.microsoft.com/>



Science Research <http://www.scienceresearch.com/scienceresearch/>

Es un sistema de búsqueda federada, el rastreo se hace en distintos sitios lanzando la búsqueda en repositorios, bases de datos del ámbito académico.



World Wide Science <https://worldwidescience.org/> es una puerta a la ciencia global de búsqueda de bases de datos de todo el mundo. Es multilingüe y proporciona en tiempo real la búsqueda y la traducción de la literatura científica globalmente dispersa.

Cabe nombrar aquí *Recolecta* <http://recolecta.fecyt.es/> donde se recoge la ciencia en abierto producida por las universidades y centros de investigación españoles



3. Multibuscadores

Finalmente decir que otras herramientas de búsqueda son los multibuscadores que permiten la consulta conjunta en varios buscadores. Son capaces de lanzar una ecuación de búsqueda de forma simultánea a distintos motores, sistemas de búsqueda que tienen en su aplicación. En el resultado evitan las duplicidades e indican donde han sido localizados. Son útiles para búsquedas sencillas, su capacidad de combinar términos es limitada ya que sus formularios son simples.

Ejemplos:

Ixquick <http://www.ixquick.com/esp/>

Es de los más potentes. Permite búsqueda avanzada. Los resultados tienen valoración con estrellas según la relevancia del resultado.



Metacrawler <http://www.metacrawler.com/>

Creado en 1994 es de los más antiguos. La búsqueda por defecto se realiza en todos los buscadores del sistema y se obtiene un resultado conjunto y otro opcional independiente por cada buscador.



DuckDuck go <https://duckduckgo.com/>

Tiene como norma el respeto a la privacidad en las búsquedas. Agrega los resultados de varios motores y los presenta de forma privada al usuario.

4.- Directorios

Puede decirse que son herramientas que recopilan y facilitan información sobre personas, actividades, instituciones, lugares.

Se trata de guías o listas agrupadas y ordenadas sistemáticamente por categorías y subcategorías realizadas por personas, que registran las direcciones y ofrecen una descripción de los sitios web.

Resultan fáciles de usar, y suponen una forma ordenada de navegar por la información, ya que clasifican los recursos incluidos en su base de datos para facilitar el acceso a los mismos. De hecho se puede realizar una búsqueda por categorías mediante los encabezamientos de materia que llevan a los recursos específicos de esos temas, o bien, la búsqueda por palabras clave introducida en la ventana recuperando mediante motor interno.

Bien es verdad que a veces puede resultar confuso decidir bajo qué categoría de materias puede estar incluido lo que nos interesa, ya que a veces no están muy claras las fronteras temáticas. Presentan el inconveniente de que proporcionalmente recogen una porción reducida de lo que hay en la red, por lo que resultan de utilidad en una búsqueda general de recursos. La actualización es limitada.

Tramullas y Olvera distinguen entre un directorio generalista y uno especializado. Los directorios generales incluyen páginas web sobre los más diversos temas. Pueden tener cobertura nacional y otros la tienen internacional. Un directorio especializado centra su actividad en un área temática concreta y los recursos son compilados por personas expertas que los organizan en bases de datos temáticas subdivididas en categorías específicas. Se les suele denominar también índices o incluso, bibliotecas digitales.

De entre los generales, el más conocido es *YAHOO* <https://es.yahoo.com/?p=us> que parte de una iniciativa de dos estudiantes de la Universidad de Standford cuando en 1994 comienzan a organizar temáticamente las direcciones de Internet que iban recopilando, ofreciéndolas a todos los usuarios. El crecimiento experimentado lo sitúa como uno de los directorios generales más importantes. Ha pasado de ser totalmente gratuito a utilizar un sistema mixto. Los sitios web que quieran ser incluidos pueden usar una fórmula preferencial que garantiza, mediante un pago, tener un lugar privilegiado en la visualización de los resultados. El servidor está en EEUU tiene una cobertura internacional, pero cuenta con centros locales en distintos países, p. ej. Yahoo España <http://es.yahoo.com/>.

Es un directorio generalista y en su clasificación se pueden ver apartados comerciales y de entretenimiento mezclados con categorías de carácter más académico y profesional. La búsqueda puede realizarse navegando por su clasificación temática. También se pueden realizar búsquedas definiendo los términos en la ventana para una búsqueda simple, o combinándolos con los operadores e incluso cruzándolos con las categorías temáticas.

Destacar *ROAD Directory of Open Acces Scholarly Resources* <http://road.issn.org/>
Servicio ofrecido por el Centro Internacional del ISSN y el Sector de Comunicación e Información de la UNESCO. Proporciona información sobre recursos en acceso abierto disponibles en la base de datos del ISSN. Publicaciones periódicas científicas, actas de congresos y documentos en repositorios. La información bibliográfica de cada revista o congreso se complementa con la cobertura de los recursos y con las bases de datos que indizan dichas publicaciones.

5.- Portales

Los portales son otra herramienta que ha surgido por la necesidad de hacer “la vida más fácil” a los usuarios de Internet. Se trata de servicios de información, de comercio electrónico y comunicación, que tiene como objetivo facilitar al usuario el acceso a todo lo que precise, de manera que no necesite visitar ningún otro sitio web para el ocio, compras, información y comunicación. Está formado por un buscador, un índice temático y una serie de servicios añadidos como noticias, compra electrónica, correo, foros. Se trata de un sistema comercial, cuyos sitios web se sustentan por la publicidad, por eso sus páginas suelen llenarse de anuncios. Tecnológicamente, un portal de Internet reúne un conjunto de tecnologías (HTML, XML, correo, fotos, directorios, bases de datos) que funcionan conjuntamente para facilitar el acceso ordenado a Internet. Aparecen en torno a 1997 cuando Internet ya se había difundido ampliamente y se quería proponer nuevos modelos de presentación de la información al público.

Existen varios tipos de portales: genéricos, especializados, corporativos, asociados a destinatarios concretos.

-Los portales generales u horizontales ofrecen información a un público amplio que demanda cuestiones relacionadas con el panorama de actualidad y con el ocio (música, juegos, cine, mensajería, email, noticias, buscador).

- MSN <http://es.msn.com/>

Propiedad de Microsoft ofrece información de entretenimiento, ocio y servicios de correo, noticias, horóscopo, compras, y un largo etc., pensado para no salirse de este sitio web.



Otros:

- La Guía del Ocio <http://www.guiadelocio.com/>
- AOL American On Line <http://www.aol.com/>

-Los portales especializados, también denominados verticales, intentan dirigir sus contenidos a usuarios que demandan información especializada. Por ejemplo en el caso de portales científicos, nos estaremos refiriendo a bases de datos especializadas, noticias especializadas, ofertas de empleo, directorios de profesionales, etc.

-El portal del ciudadano, Administración: <http://www.060.es/>

Es una herramienta de comunicación entre la administración y los ciudadanos, pudiendo realizarse trámites administrativos.



<http://masdestacados.060.es/>



-PARES Portal de Archivos españoles <http://pares.mcu.es/>

Como proyecto del Ministerio de Cultura español, nace para difundir el Patrimonio Histórico Documental de nuestro país.



-UNIVERSIA <http://www.universia.es>

Es el portal de la universidad española. Agrupa alrededor de doscientas universidades españolas e hispanoamericanas y pretende ser una referencia fundamental para la comunidad universitaria ofreciendo contenidos académicos y científicos.



Los hay dedicados al Arte: <http://www.1arte.com/>

Derecho: <http://vlex.com/>

Educación: <http://www.educaguia.com/>

Historia: <http://www.thehistorynet.com/>

Medicina: <http://www.diariomedico.com/>

Naturaleza: <http://www.faunaiberica.org/>

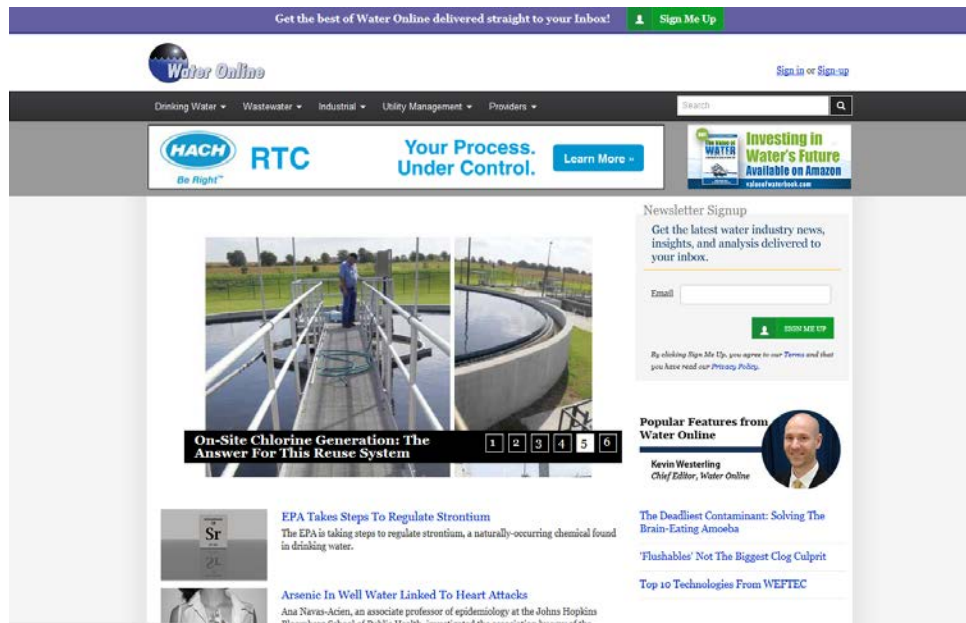
Discapitados: <http://www.discapnet.es>

- e-Libro <http://www.e-libro.com/> Se trata de una plataforma de distribución internacional de libros electrónicos de numerosas editoriales, en varias lenguas.



El paso del portal temático a la comunidad virtual (ahora denominada, red social) está dado ya. La capacidad para crear un entorno de servicios propicio para unos usuarios que comparten intereses, aficiones y que interactúan comunicando y comunicándose ha llevado a un desarrollo imparable. Un entorno en el cual los usuarios comparten sus intereses, aficiones o necesidades de comunicación, donde los servicios y productos animan a que el usuario se sienta miembro de esa comunidad y se anime a participar. En definitiva la creación de espacios o comunidades virtuales destinadas a compartir intereses comunes.

- Water online <http://www.wateronline.com/>



- Instituto Geográfico Nacional www.ign.es/