

Recuperación de información científica en la Web

Principios básicos de Cibermetría

José Luis Ortega

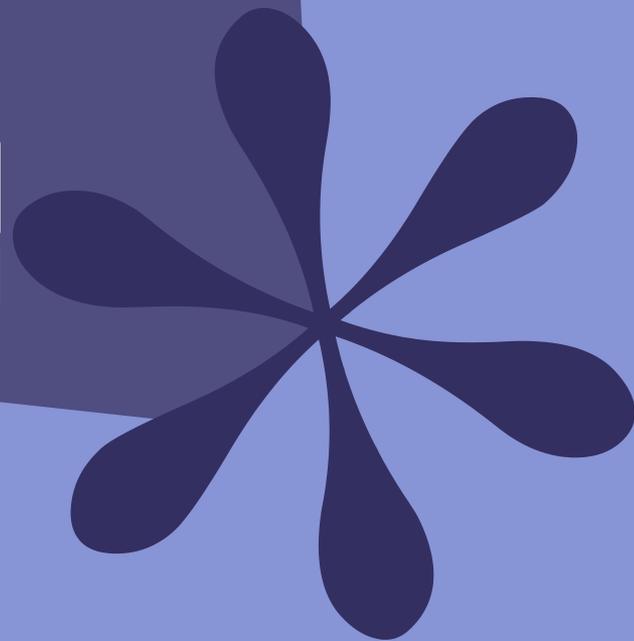
Vicepresidencia de Investigación Científica y Técnica

Consejo Superior de Investigaciones Científicas



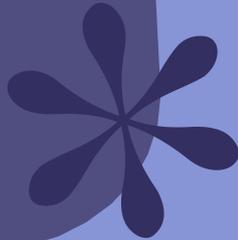
Colaborador del **CybermetricsLab**

jortega@orgc.csic.es



Agenda

- * Historia
 - Internet Archive
- * Motores de búsqueda
 - Crawler
 - Ordenación
 - Operadores
- * Buscadores académicos
- * Repositorios
- * Redes sociales científicas



Historia

* **YAHOO!** (1994): Primer directorio

*  (1994): Primeros buscadores

*  (1998)

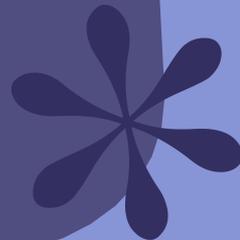
* Guerra de buscadores (2003)

– Yahoo! adquiere AltaVista y Alltheweb

*  (2005): se independiza

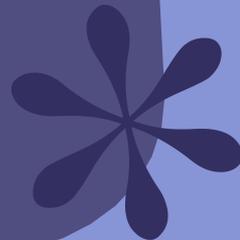
* Los tres grandes (2008)



Internet Archive

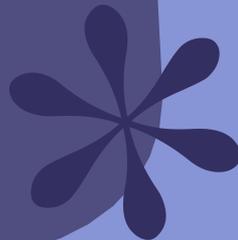
- * Archivo histórico de páginas web
- * 150 billones de páginas desde 1996
- * Audio, Video, software, etc.



Buscadores: funcionamiento

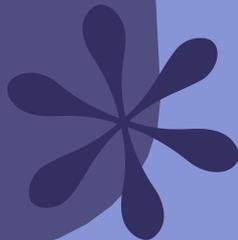
Se compone de tres elementos:

- * Base de datos y sistema de indexación
 - Robot o *Crawler*: Googlebot, Slurp, MSNbot, etc.
 - Texto completo
- * Motor de búsqueda
 - *Matching* y operadores de búsqueda
- * Interfaz de acceso
 - Ordenación de resultados



Buscadores: ordenación

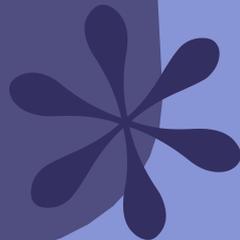
- * Miles de resultados por consulta
- * Peso del término de consulta (TF/IDF)
- * Sistemas probabilísticos, vectorial, difuso, etc.
- * PageRank (Bring & Page, 1998)
<http://www.miwebrank.com/>
 - Relevancia según enlaces entrantes (visibilidad)
 - Transmisión de la visibilidad (cadena de Markov)



Buscadores: operadores



Operador	Definición
cache:	Busca en la caché del buscador
link:	Busca paginas que enlazan a un sitio web
related:	Busca paginas relacionadas
define:	Busca definiciones de una palabra
site:	paginas de una sede web
allintitle:	Busca sólo en el título de las páginas
allinurl:	Busca sólo en la URL de las páginas
inurl:	Busca en paginas cuya URL contenga un cierto término
filetype:	Busca archivos de un determinado formato



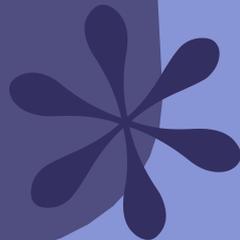
Buscadores: operadores

The image shows the classic red 'YAHOO!' logo in a white rectangular box with a slight drop shadow.

Operador

Definición

link:	Busca paginas que enlazan a una página web
linkdomain:	Busca paginas que enlazan a un sitio web
define:	Busca definiciones de una palabra
site: o domain:	paginas de una sede web
intitle:	Busca sólo en el título de las páginas
allinurl:	Busca sólo en la URL de las páginas
inurl:	Busca en paginas cuya URL contenga un cierto término
originurlextension:	Busca archivos de un determinado formato



Buscadores: features y shortcuts

Operador

feature:audio

feature:homepage

region:mideast

Atajos (Shortcuts)

convert * [monedaA] [monedaB] Convierte el valor de una moneda a otra

convert * [medidaA] to [medidaB] Convierte el valor de una medida en otra

patent * Busca en la USPTO una patente

*** facts** Busca entradas en enciclopedias

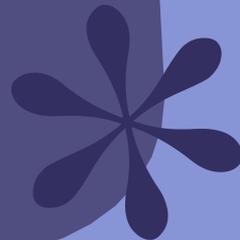
*** images** Busca imágenes

Definición

Busca páginas con archivos de audio

Busca paginas personales (~)

Busca páginas en la región de Oriente Medio

The image shows the classic Yahoo! logo in red, bold, sans-serif font with an exclamation point, set against a white background with a slight drop shadow.

Buscadores académicos

- * Scirus (2001)

- * 400 millones

- * Elsevier



- * Google Scholar (2004)

- * Sin cobertura conocida

- * Falla identificación de autores

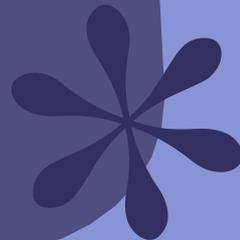


- * Microsoft Academic Search (2009)

- * Live Academic Search

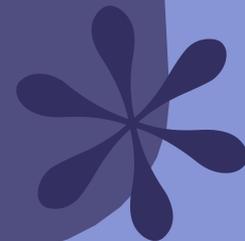
- * 8 millones

- * Indicadores bibliométricos



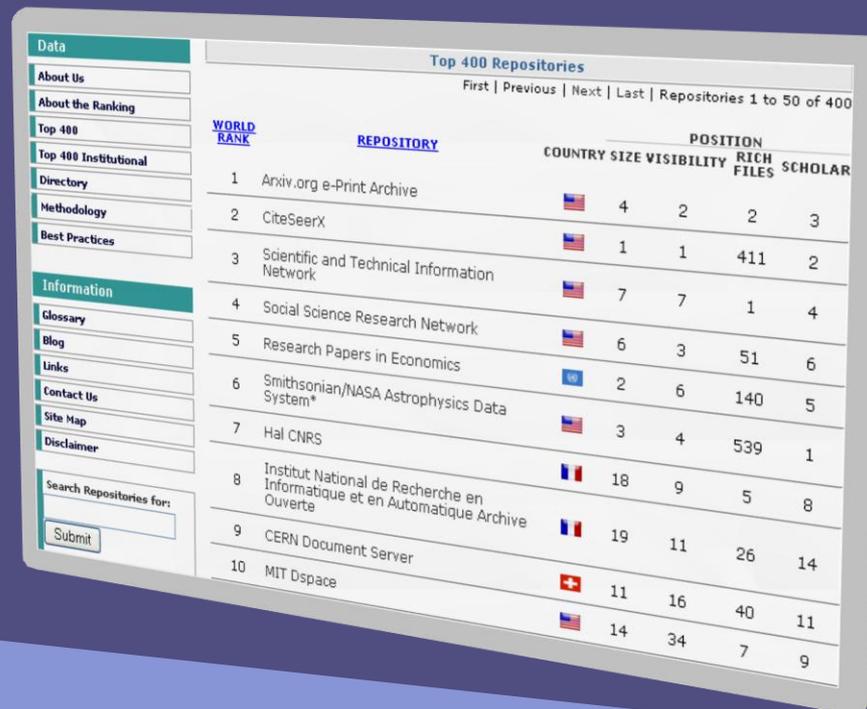
Buscadores de datos agregados

- * No recuperan documentos sino datos
- * Se nutren de variadas fuentes y bases de datos
- * Wolfram|Alpha
 - * Wolfram Research, 2009
 - * Multidisciplinar
- * ChemSpider
 - * Royal Society of Chemistry, 2007
 - * Especializado en Química



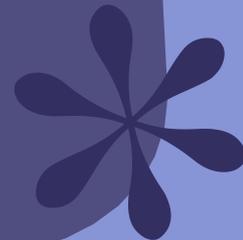
Repositorios

- * Alojamiento de documentos científicos
- * Libre Acceso (*Open Access*) 
- * Web Ranking of World Repositories



The screenshot shows a web page titled "Top 400 Repositories" with navigation links: "First | Previous | Next | Last | Repositories 1 to 50 of 400". The page features a table with columns for "WORLD RANK", "REPOSITORY", "COUNTRY", "SIZE", "VISIBILITY", "POSITION", "RICH FILES", and "SCHOLAR". The table lists the top 10 repositories, including Anxiv.org e-Print Archive, CiteSeerX, Scientific and Technical Information Network, Social Science Research Network, Research Papers in Economics, Smithsonian/NASA Astrophysics Data System*, Hal CNRS, Institut National de Recherche en Informatique et en Automatique Archive, CERN Document Server, and MIT Dspace. A sidebar on the left contains navigation links such as "Data", "About Us", "About the Ranking", "Top 400", "Top 400 Institutional", "Directory", "Methodology", "Best Practices", "Information", "Glossary", "Blog", "Links", "Contact Us", "Site Map", and "Disclaimer". At the bottom of the sidebar is a search box labeled "Search Repositories for:" with a "Submit" button.

WORLD RANK	REPOSITORY	COUNTRY	POSITION			
			SIZE	VISIBILITY	RICH FILES	SCHOLAR
1	Anxiv.org e-Print Archive		4	2	2	3
2	CiteSeerX		1	1	411	2
3	Scientific and Technical Information Network		7	7	1	4
4	Social Science Research Network		6	3	51	6
5	Research Papers in Economics		2	6	140	5
6	Smithsonian/NASA Astrophysics Data System*		3	4	539	1
7	Hal CNRS		18	9	5	8
8	Institut National de Recherche en Informatique et en Automatique Archive		19	11	26	14
9	CERN Document Server		11	16	40	11
10	MIT Dspace		14	34	7	9



Repositorios

* Temáticos

– ArXiv.org



- 600.000 artículos
- Física, Matemáticas e Informática

– PubMed Central



- CC. de la Vida
- 1 millón de documentos

– eCrystals



- 700 artículos
- Estructura de Cristales

* Institucionales

– CERN Document Server

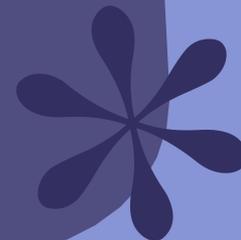


- 360.000 documentos

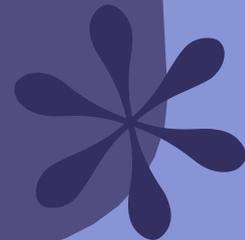
– Digital.CSIC



- CSIC
- 30.000 documentos

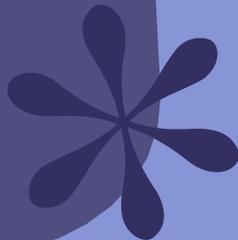


- * Universidad de Pennsylvania, 2010
- * No es estrictamente un repositorio, sino un recolector
- * 2500 artículos aprox.
- * Se estructura en tres partes:
 - * Chemical Entity Search: permite buscar formulas y nombres químicos
 - * formula: H₂O
 - * name: zeolites
 - * TableSeer: busca dentro de tablas
 - * Databases: recoge datos experimentales (Excel, XML, Gaussian y Charmm)



Redes sociales científicas

- * Plataformas en la que los científicos intercambian información y se relacionan socialmente
- * Servicios:
 - * Crear perfiles personales
 - * Listas de discusión
 - * Blogs
 - * Resultados de investigación: textos, gráficos, videos, etc.
- * Colaboración e internacionalización



Redes sociales

- * Nature Network



- * 25.000 miembros

- * Generalista

- * Biomed Experts



- * 266.000 miembros

- * Biomedicina y CC. de la Vida

- * UniPHY



- * 15.000 miembros

- * Física

- * LinkedIn



- * 90 millones de miembros

- * Red profesional

- * Specialty Chemical Network

