# 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

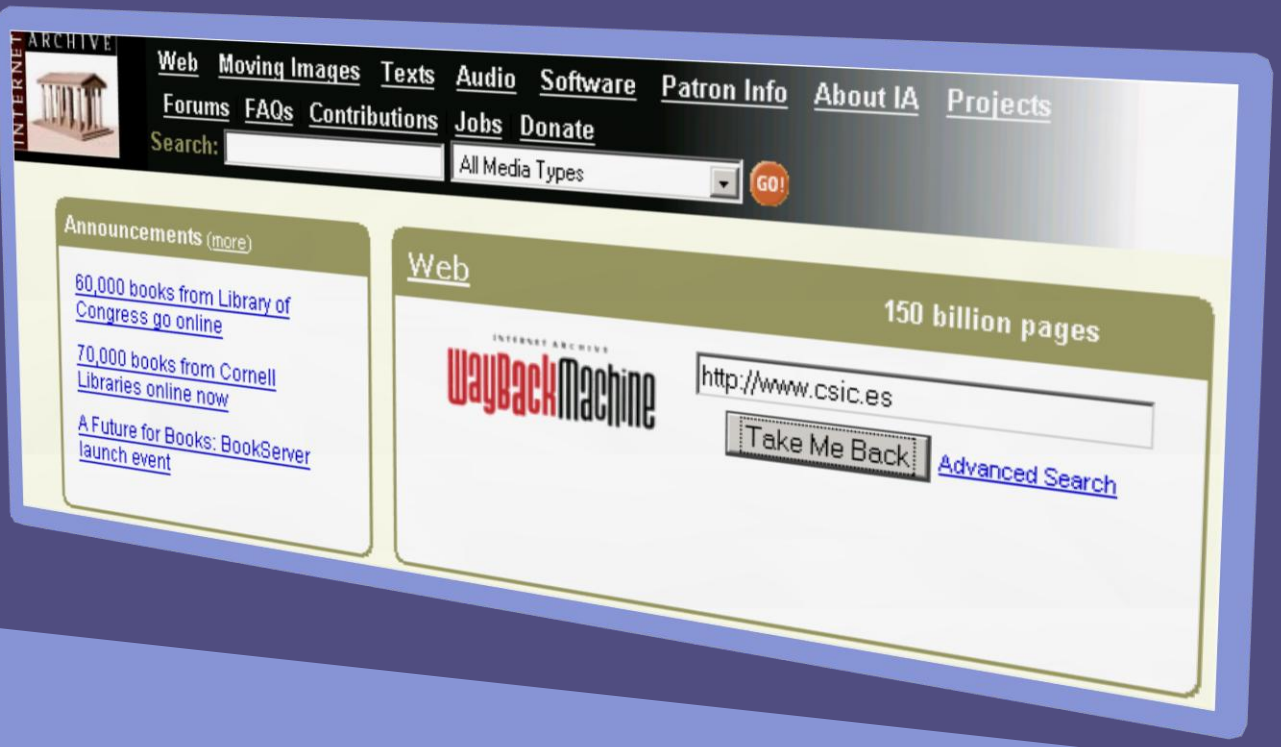
# Historia

* YAHOO! (1994): Primer directorio
* LYCOS Your Personal Internet Guide excite (1994): Primeros buscadores
* Google! BETA (1998)
* Guerra de buscadores (2003)
  - Yahoo! adquiere AltaVista y Alltheweb
* msn (2005): se independiza
* Los tres grandes (2008)

Google YAHOO! bing

# Internet Archive

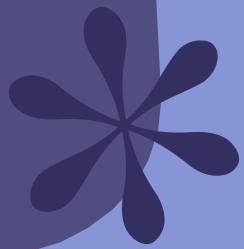
* Archivo histórico de páginas web
* 150 billones de páginas desde 1996
* Audio, Video, software, etc.



# Buscadores: funcionamiento

Se compone de tres elementos:

* Base de datos y sistema de indización
  - 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 probabilisticos, 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

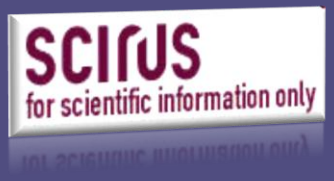

| Operador | Definición |
|---|---|
| **link:** | Busca paginas que enlazan a una página web |
| **linkdomain:** | Busca paginas que enlazan a un sitio web |
| **define:** | Busca definiciones de una palabra |
| **site: o domain:** | paginas de una sede web |
| **intitle:** | Busca sólo en el título 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 | Definición |
|---|---|
| **feature:audio** | Busca páginas con archivos de audio |
| **feature:homepage** | Busca paginas personales (~) |
| **region:mideast** | Busca páginas en la región de Oriente Medio |
| **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 |

# 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



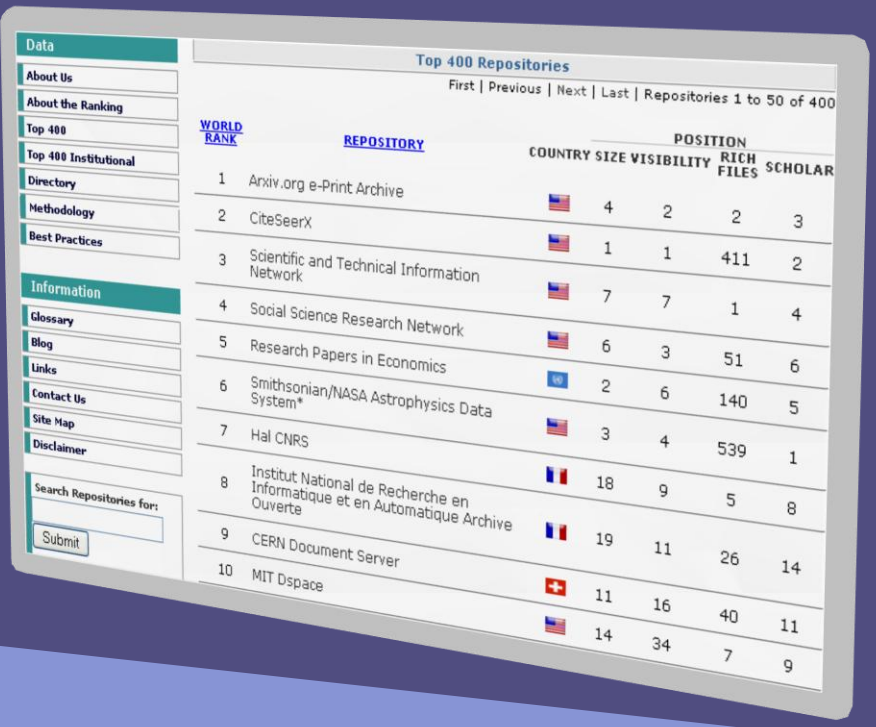


# Buscadores de datos agregados

* 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



# 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

# ChemXSeer


* 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