Tema 3. Recuperación de la información.

VNIVERSITAT DE VALÈNCIA
OpenCourseWare

Índice.

1. Introducción
2. Buscadores de recursos
3. Técnicas de Búsqueda

Parte I: Introducción

1. Introducción.

- Ya sabemos cómo se generan documentos digitales y conocemos los distintos formatos.
- Veremos cómo se recuperan centrándonos en un medio específico: Internet.
- **Internet**: red de ordenadores conectados, con una enorme cantidad de sitios Web, y por tanto de información.
- En la Web tenemos una *gran base de datos* con información de todo tipo: texto, imágenes, audio y vídeo, y en múltiples formatos.

? Para vosotros

¿Qué características específicas tiene la Web que hace difícil recuperar información?.

1. Introducción.

- La Web tiene una serie de características específicas (*problemas intrínsecos de los datos*), como son:
  - La información está **distribuida** en muchos ordenadores distintos.
  - Hay una gran **volumen** de datos, que además son **volátiles**, ya que aparecen y desaparecen continuamente nuevas páginas.
  - No se conoce a priori la **estructura** de la información, y gran parte se genera dinámicamente mediante consultas a bases de datos.
  - Hay mucha **redundancia** de información (webs repetidas, webs con el mismo contenido).
  - Los datos son **heterogéneos,** con diferentes tipos de formatos de ficheros.
  - La **calidad** no es la misma en todas las fuentes de información.

Parte II: Buscadores de recursos

Buscadores de recursos

- **Los buscadores de recursos se pueden clasificar :**

1) Según su organización y funcionamiento en:

- Índices o *directorios,* que catalogan y organizan la información por categorías.
  - Son catálogos Web con recursos clasificados y organizados por categorías y subcategorías.
  - Existen directorios generales y directorios temáticos.

Ej: es.dir.search.yahoo.com/dir, www.dmoz.org

- *Motores de búsqueda*, que son programas que buscan a través de bases de datos de documentos html. Hay de dos tipos:
  - Buscadores sin robot.
  - Buscadores con robot.
- *Motores de decisión* , determinan cual es la respuesta o solución concreta a una pregunta o decisión
- *Buscadores de bitacoras*, buscan en el contenido de blogs o weblogs
- *Buscadores Temáticos* **,** ***buscan*** cualquier tipo de recurso o campo específico que podamos imaginar

Buscadores de recursos

2) Según el número de bases de datos a las que acceden:

- Acceso a una sola base de datos.
- Multibuscadores: a varias secuencialmente. Ej: www.compendio.com
- Metabuscadores: a varias simultáneamente.

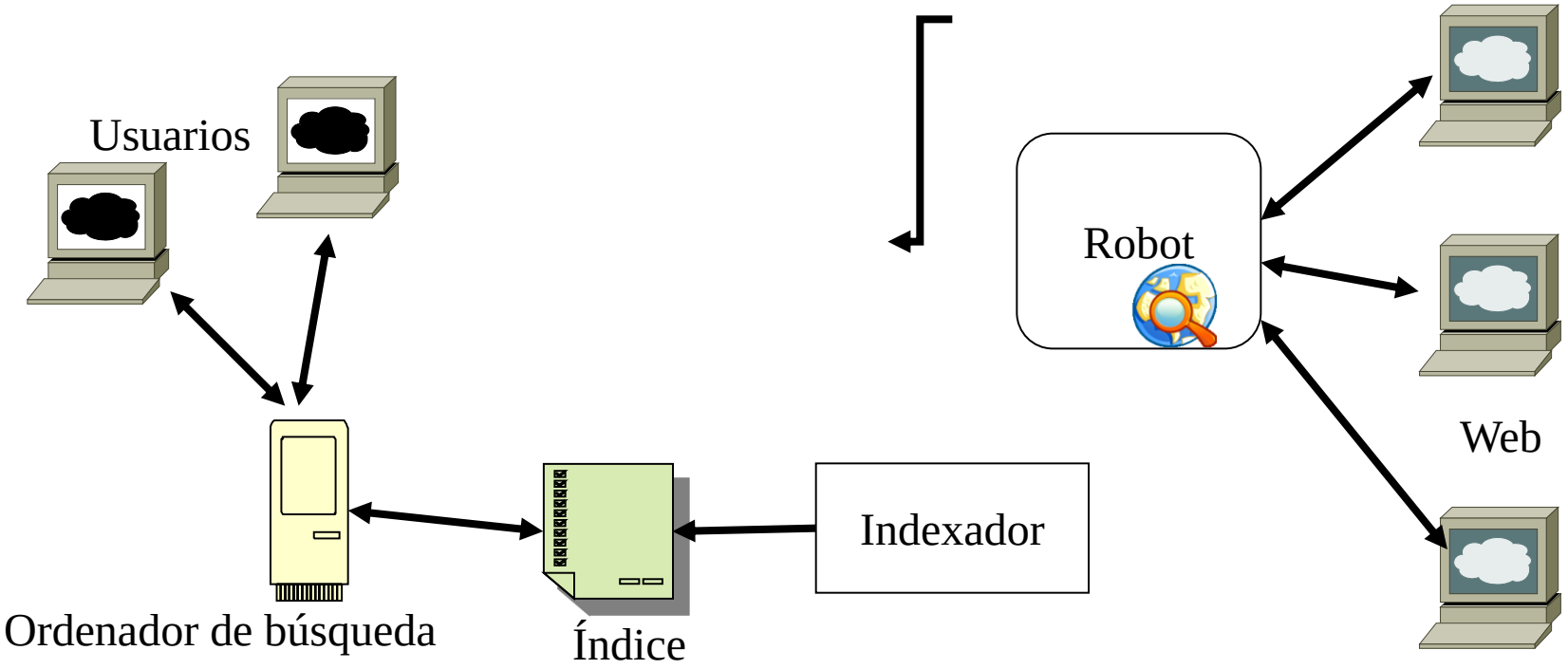
Ej: www.metacrawler.com , www.kartoo.com

Motores de búsqueda.

- Usan el paradigma de recuperación en texto completo.
- Todas las palabras de un documento se almacenan en un documento para su posterior recuperación
- Principal problema: recorrer la Web actualizando y agregando nuevas páginas.
- Motores de búsqueda **sin robot**:
  - Es necesario dar de alta las páginas para figurar en su base de datos.
  - Los contenidos en ocasiones son analizados por personas que visitan la dirección añadida y comprueban que cumple los requisitos para ser dado de alta.

Motores de búsqueda

- **Con robot**.
  - Son programas que buscan a través de la estructura del hipertexto recuperando enlaces.



Motores de búsqueda

- ¿Funciones de un robot?
  - Utilizan diferentes estrategias para elegir las Web a visitar.
  - Es habitual que almacene una lista histórica de URL's.
  - Cada página modificada o nueva que encuentra el robot es procesada.
  - Se analizan las páginas de la lista extrayéndose de ella otras páginas, que si son nuevas, se agregan a la lista de páginas a recorrer.
  - No es necesario dar de alta un sitio Web para aparecer en el buscador.

Motores de búsqueda con robot

- ¿Qué indexa un robot?
  - Normalmente se indexan los títulos HTML (etiqueta TITLE) y los primeros párrafos.
  - En ocasiones las palabras contenidas en el documento excluyendo las de uso común.
  - Los *metadatos* contenidos en las etiquetas META de la cabecera HTML→ es importante utilizar correctamente las palabras clave dentro del HTML.
  - También se indexan textos alternativos a las imágenes.

Motores de Decisión.

- No cuenta con una base de datos compuesta de páginas web indexadas, sino con una base de conocimiento y una serie de reglas que le permiten operar sobre ella

- Ejemplo motor de decisión: Wolfram|alpha

- Quiero saber la respuesta de :¿ población españa?

  - Quiero saber la respuesta de: valor x que minimiza

$$7 + 3x + x^2$$

Buscadores de Bitácoras.

- Se basan en nubes de tags→tienene un área de la página en la que aparecen las marcas más populares, normalmente con un tamaño proporcional al número de entradas publicadas que hacen referencia a ellas.

- No rastrean la Web como hacen los buscadores , sino que que las propias bitácoras cuando se actualizan envían una señal al buscador

– Ejemplo: http://es.wordpress.com/

Buscadores de Bitácoras.

WordPress.com

Inicio Registrarse Características Noticias Sobre Nosotros Avanzado

Blogs sobre: Valencia Cf

Blog Destacado

David Albelda, un crack

David Albelda, ayer durante la entrevista.- TANIA CASTRO Un esguince en una clavícula deja a Villa tocado para la visita del Valencia, mañana, al Barcelona. El equipo de Unai Emery acude mermado a pes... more →

Jon Kepa

Atlético-Valencia, en cuartos de la Europa League

mikeberriv escribió 4 days ago: Sólo un equipo español avanzará en los cuartos de final de la Europa League MADRID, España, Mar. 19, ... more →

Etiquetas: Futbol, Liverpool FC, Atlético De Madrid, Hamburgo SV, europa league 2009-2010, SL Benfica, VfL Wolfsburgo, Standard Lieja, Fulham FC

Valencia y Benfica avanzan; la Juve se despide

mikeberriv escribió 5 days ago: En la Europa League, Valencia y Benfica obtienen su pase a cuartos de final,

Have *your* say.
Start a blog.
See our free features →

Sign Up Now!

Etiquetas relacionadas Todo →

Futbol SL Benfica
europa league
2009-2010
Standard Lieja Fulham FC
VfL Wolfsburgo

Directorios

- Directorios se crean de forma manual, recopilando las direcciones de los sitios y asociándoles a una o más categorías o descripción.

Ejemplos de Directorios:

-Librarian Index: www.lii.org

-Yahoo: dir.yahoo.com

-Google: directory.google.com

-About: www.about.com

-Webbrain: www.webbrain.com

Directorios

Search: the Web | the Directory

Search

Yahoo! Directory

Advanced Search Suc

Arts & Humanities
Photography, History, Literature...

Business & Economy
B2B, Finance, Shopping, Jobs...

Computers & Internet
Hardware, Software, Web, Games...

Education
Colleges, K-12, Distance Learning...

Entertainment
Movies, TV Shows, Music, Humor...

Government
Elections, Military, Law, Taxes...

Health
Diseases, Drugs, Fitness, Nutrition...

News & Media
Newspapers, Radio, Weather, Blogs...

Recreation & Sports
Sports, Travel, Autos, Outdoors...

Reference
Phone Numbers, Dictionaries, Quotes...

Regional
Countries, Regions, U.S. States...

The Spark: The Toons of Ub and Joe

By Dave Sikula
Wed, March 24, 2010, 12:01 am PDT


Ub Iwerks caught drawing Mickey

Fame's a funny thing. As the noted philosopher Heidi Klum has stated, "One day you're in, the next day you're out." Or vice versa. Consider the cases of animation directors Joe Barbera and Ub Iwerks, whose birthdays we celebrate today.

Joe Barbera toiled in anonymity as an animator and writer for such studios as Van Beuren and Terrytoons before settling in at MGM, where he was teamed with Bill Hanna, and the rest was movie history. Their first picture together was "Puss Gets the Boot," which introduced Tom and Jerry to the world -- and garnered the team the first of their 13 Oscar nominations. When MGM closed down its animation department in 1957, the team simply began producing shows for television, creating such megahits as "The Flintstones," "The Jetsons," "Jonny Quest," and "Scooby-Doo."

Ub Iwerks, on the other hand, started at the top. A childhood friend of Walt Disney, he was the animator and director for Disney's "Alice" and "Oswald the Lucky Rabbit" pictures, and is credited with the creation of the most famous cartoon star of all: Mickey Mouse. Walt and Ub eventually had a falling out, and Iwerks left to head his own studio. Despite the high quality of his cartoons, though, he was unsuccessful. He returned to Disney to mastermind the studio's technical innovations in relative anonymity until his death in 1971.

Iwerks has been rediscovered in recent years, but one wonders how entertainment history would have changed if the world had "flipped" for a frog and not for a cat and mouse.

Cómo elegir la herramienta adecuada

- Cuando sabes dónde encontrar la información ir directamente al “site”
- No sabemos donde se encuentra, pero podemos determinar el campo que corresponde-→Directorios o bitácoras
- Si lo que buscamos es un sitio de recursos concreto→buscadores temáticos
- Si sabemos la palabras clave de lo que buscamos pero no el tema→motor de búsqueda o metabuscador
- Buscamos una pregunta concreta→motor de decisión o buscador de respuestas

Buscadores

- **Motores de Búsqueda genéricos:**
  - Altavista en español: http://es.altavista.com
  - Lycos en español: www.lycos.es
  - Excite: www.excite.com
  - AOL: www.aol.com
  - HOTBOT www.hotbot.com
  - ALLtheWeb: www.alltheweb.com/
  - GOOGLE: www.google.com
  - BING http://bing.com
  - ASK: ask.com
  - OZU. Ozu.es

Buscadores

- **Directorios**

-Librarian Index: www.lii.org

-Yahoo:http://espanol.dir.yahoo.com/Zonas_geograficas/Paises/Espana/

-Google: directory.google.com

-About: www.about.com

-Open Directory project: dmoz.org

-01WebDirectory: http://www.01webdirectory.com

-Internet Public Library (IPL): http://www.ipl.org

(Contiene referencias principalmente a recursos bibliográficos: libros, artículos en revistas especializadas y periódicos)

Buscadores

- **Buscadores de Bitácoras**

-http://www.technorati.com

-http://www.blogalaxia.com

-http://www.bloglines.com

-http://blogsearch.google.com

-http://blogpulse.com

-http://icerocket.com

Buscadores

- **Buscadores Temáticos:**

-http://ebay.es: eBay (subastas)

-http://paginasamarillas.es: Páginas Amarillas (datos de empresas y particulares)

-http://a9.com (sitios de comercio electrónico)

-http://expedia.es (Hoteles, viajes, vuelos)

Imágenes

-http://flickr.com (Fotografías)

-http://www.xcavator.com (Fotografías)

-http://www.picsearch.es

Contenido Audiovisual

-http://es.video.yahoo.com

-http://video.google.es

-http://www.open-video.org

Buscadores

- **Buscadores Temáticos:**

- Contenido Audiovisual

-http://www.findsounds.com (especializado en sonidos)

-http://www.dailymotion.com

-Noticias

-http://news.google.es

-www.abastodenoticias.com

-www.noticias.com

-Código Fuente

-www.koders.com

-www.google.com/codesearch

-Archivos en un determinado formato de escritura:

-www.docjax.com

Buscadores

- **Metabuscadores**

  - Dogpile: www.dogpile.

  -Kartoo: www.kartoo.com (ya no existe, pero era muy interesante)

  -Mamma: www.mamma.com

  - Metacrawler: www.metacrawler.com

  -Ixquick: http://ixquick.com

  -Clusty: http://clusty.com

  -Creative Commons: http://search.creativecommons.org: Creative Commons

- **Multibuscadores:**

  - Compedio: www.compendio.com

- **Motores de Decisión:** www.wolframalpha.com

Parte III: Técnicas de búsqueda.

? Para vosotros

¿Es lo mismo buscar:

"Asociación Española de Profesores Universitarios de Contabilidad"

“Documentos sobre el impacto del comercio electrónico en la Contabilidad“?

Técnicas de recuperación de la información con motores de Búsqueda

Procedimiento a seguir:

1. Definir bien el objetivo de la búsqueda

2. Utilizar estrategias de búsqueda de acuerdo al objetivo

3. Ordenarlas según su eficacia y eficiencia.

4. Replanteamiento de estrategia y/o buscadores de recursos (directorios, motores de búsqueda conceptuales) si no ha obtenido los resultados esperados .

Estrategias de Búsqueda en la Web

Características de tu Búsqueda	Estrategias
nombre o frase distintiva ? • Nombre de una organización o movimiento • Nombre de un individuo • Cadena de palabra asociadas con el tema de búsqueda. Puedes pensar en una organización, nombre o frase que estás buscando? Puede ayudar a cercar tu búsqueda	**PHRASE SEARCHING** es una característica que quieres en cada herramienta de búsqueda que elijas: -Requiere todos los términos aparecer exactamente en el orden que los introduces -La frase se introduce utilizando " " Ejemplos **"affirmative action"** **"world health organization"** **"a person's name"**
Algunos de los términos son **palabras comunes con muchos significados y contextos**? •*Children* con *television* y también con *violence* •*Censorship* como un aspecto ético en *journalism*	**BOOLEANO AND** ayudará : **children AND television AND violence** **journalism AND ethics AND censorship** **Google** and **AllTheWeb**a y las mayoría de buscadores ponen AND entre las palabras (por defecto): **children television violence** **journalism ethics censorship**
Anticipas muchos resultados con términos que no quieres? •Buscas *biomedical engineering* y *cancer* y te aparecen muchos programas académicos y lo que quieres son informes sobre este tema. Por tanto puedes excluir documentos que contengan " Departament of " o " School of".	**BOOLEA AND NOT** ayudará: **"biomedical engineering" AND cancer AND NOT "Department of" AND NOT "School of"** o **–excluye lo que es equivalente a**: **"biomedical engineering" cancer -"Department of" - "School of"**

Estrategias de búsqueda en la Web

Características de tu Búsqueda	Estrategias
Hay **sinónimos, variaciones de una palabra**, o **deletreado extranjero** para alguno de tus términos? • *women,* females con *networking* • *Sarajevo, Sarayevo* con *peace* •*literature, litterature* con *French, francaise*	**BOOLEANO OR** ayudará: **(women OR females) AND networking** **(Sarajevo OR Sarayevo) AND peace** **(literature OR litterature) AND (French or francaise)** En **Google**, capitalize OR (no need to type "and"): **peace sarajevo OR sarayevo** **literature OR litterature french OR francaise** En AllTheWeb, usa paréntesis u omite OR: **peace (sarajevo sarayevo)** **(literature litterature) (french francaise)**
Estas buscando por **home pages** y/o otros documentos**, principalmente utilizando término(s)**? •La página de: the *American Dietetic Association* •Páginas principalmente sobre: *Affirmative Action*	LIMIT **Limítalo a los campos del TíTULO DE LOS DOCUMENTOS** **intitle:"American Dietetic Association"** **intitle:"affirmative action"** en **Google**, usa **intitle:"affirmative action"**
Estás buscando por términos que tengan muchas terminaciones? •*Feminism, feminist, feminine* •*Children, child*	Algunos sistemas buscan las terminaciones de algunos nombres de manera automática . Para estar seguro usa OR en las búsquedas: **children OR child**

Estrategias de Búsqueda Avanzadas (Google)

- Búsquedas en la dirección de la página:

  inurl: w3.org html

  -inurl: microsoft.com linux

- Búsquedas en el título:

  intitle:lenguaje **intitle:**programación

  (**inurl:** manual.pdf OR **inurl:** guia.pdf) **intitle:**casio

  allintitle: lenguaje de programación

- Búsquedas en los hipervínculos:

  inanchor: descarga **intitle**:libro

Técnicas de Búsqueda Avanzadas (Google)

- Documentos en un cierto formato:

  Sistemas operativos **filetype:**ppt

- Páginas que apuntan a otras:

  link: wikipedia.com

  link: microsoft .com **-inurl**: microsoft.com

Técnicas de Búsqueda Avanzadas (Google)

- Búsqueda de palabras cercanas:

  explosion * super nova

- Búsqueda de Definiciones

  define:computer

- Búsqueda de Sinónimos

  lenguaje c **intitle**:~curso

Técnicas de Búsqueda Avanzadas (Google)

- Información sobre un sito

  info: www.fcharte.com

- Búsqueda de Sitios relacionados

  related: www.astroseti.org

- Búsqueda dentro de un dominio:

  site:www.astroseti.org supernova

Técnicas de Búsqueda Avanzadas (Bing)

- Enlaces a documentos de tipo específico

  tenis **contains:**pdf
  tenis **filetype:**pdf

- Encuentra páginas que están alojadas en un determinado host que tienen la dirección **ip** que tu buscas

  ip:207.241.148.80

- Buscar en un determinado idioma:

  tenis (**language:** fr)

Técnicas de Búsqueda Avanzadas (Bing)

- Encuentra páginas que contengan una determinada palabra en el "body" de una página.

  inbody:tennis

- Encuentra páginas que contengan una determinada palabra en el "title" de una página.

  intitle:tennis

Técnicas de Búsqueda Avanzadas (Bing)

- Limita tu búsqueda a un dominio específico:

  site: .org

  site: .gov

  site: .edu

- Encontrar páginas que en el url contengan unos determinados términos

  url: about.com

Técnicas de Búsqueda Avanzadas (Bing)

- Obtener sitios web que contenga que cuentan con un sistema de subscripcion (RSS o ATOM)

  astronomia **hasfeed:** tennis

Comparativa de consultas entre diferentes buscadores

Acción	Como	En que buscadores?
Debe Incluir un término	**+**	**All**
Debe Excluir un término	**-**	**All**
Debe Incluir Frase	**" "**	**All**
Coincidir todos los Términos	**Automática**	**All**
Coincida cualquier término	**Por Búsqueda Avanzada**	**AllTheWeb, AltaVista, Google, Lycos, MSN Search, Teoma, Yahoo**
	OR	**AltaVista, AOL Search, Ask Jeeves, Google, HotBot, MSN Search, Teoma, Yahoo** *(se debe hacer en mayúsculas)* **AllTheWeb, Lycos** *(solo para dos palabras)*

Comparativas de Consulta entre diferentes buscadores

Acción	Como	En que Buscadores
	title:	AltaVista, AllTheWeb, Inktomi
	intitle:	Google, Bing Teoma
	allintitle:	Google
Búsqueda por título:	host:	AltaVista
	site:	Excite, Google (Netscape, Yahoo) Bing
	url.host:	AllTheWeb, Lycos (for AllTheWeb results only)
	domain:	Inktomi (HotBot, iWon, LookSmart)
	url:	AOL, DirectHit, HotBot, LookSmart, Lycos, MSN, Netscape, Northern Light, Open Directo

Comparativas de Consulta entre diferentes buscadores

Acción	Como	En que Buscadores
Búsqueda basada en el "URL"	url:	AltaVista, Excite, Northern Light, Bing
	url.all:	AllTheWeb, Lycos (for AllTheWeb results only)
	allinurl: inurl:	Google
	originurl:	Inktomi (AOL, GoTo, HotBot)
	u:	Yahoo
	none	AOL, Direct Hit, HotBot, LookSmart, MSN. Bing **Not yet updated, but may be still correct:** Open Directory
	link:	AltaVista, Google, Northern Light
	linkdomain:	Inktomi (AOL, HotBot, iWon, MSN) (NOTE: measures links to entire domains)
	link.all:	AllTheWeb, Lycos (for AllTheWeb results only)
	none	AOL, Direct Hit, Excite, HotBot, LookSmart, Northern Light **Not yet updated, but may be still correct:** Netscape, Yahoo (n/a)

Comparativas de Consulta entre diferentes buscadores

Carácter Comodín	*	AltaVista, Inktomi (iWon), Northern Light **Not yet updated, but may be still correct:** Yahoo
	?	AOL Search, Inktomi (iWon)
	%	Northern Light
	none	AllTheWeb, Direct Hit, Excite, Google, HotBot, LookSmart, Lycos, MS (MSN's help says it offers wildcard, but it failed to during testing)

3. Otros Operadores

- Otros operadores menos habituales:
  - ADJ (adyacente): cuando se desean encontrar documento con los términos cerca, en cualquier orden.
  - NEAR (cerca): cuando los términos deban aparecer en las 25 palabras próximas.
  - FAR(lejos): los términos aparecen con 25 palabras o más de distancia.
  - BEFORE (antes): similar a AND pero con los términos en un orden preciso.

4. Síntesis

- La Web presenta una sería de problemas intrínsecos (datos y usuarios) que hacen difícil recuperar información.

- Los principios básico de los Buscadores de Recursos (directorios, motores de búsqueda) han sido introducidos

- Procedimientos y Estrategias recomendables de búsqueda, así como los operadores más potentes han explicados.

Conclusión: La recuperación de Información en Internet presenta un gran reto tanto para usuarios como investigadores.

2.2.Buscadores

- Algunos tienen opciones como:
  - Buscar páginas en un determinado idioma.
  - Buscar documentos en un determinado formato (pdf, word).
  - Buscar páginas actualizadas recientemente.
  - Buscar por tipos de documentos: texto, imágenes, música.

Bibliografía

- Gestión Digital de la Información. Capítulo 14
- http:http://www.unav.es/fcom/mmlab/brasil2008/
- http://www.abcdatos.com/buscadores/