Communicator’s Tools (II): Documentation and web resources

ENGLISH FOR SCIENCE AND TECHNOLOGY

“La web es un mundo de aplicaciones textuales… hay un gran conjunto de imágenes e incontables archivos de audio, pero el texto predomina no sólo en cantidad, sino en utilización…”

Millán (2001: 35-36)

Internet

- Global computer network of interconnected educational, scientific, business and governmental networks for communication and data exchange.
- Purpose: find and locate useful and quality information.

Internet

- Web acquisition. Some problems
  - Enormous volume of information
  - Fast pace of change on web information
  - Chaos of contents
  - Complexity and diversification of information
  - Lack of security
  - Silence & noise
  - Source for advertising and money
  - No assessment criteria

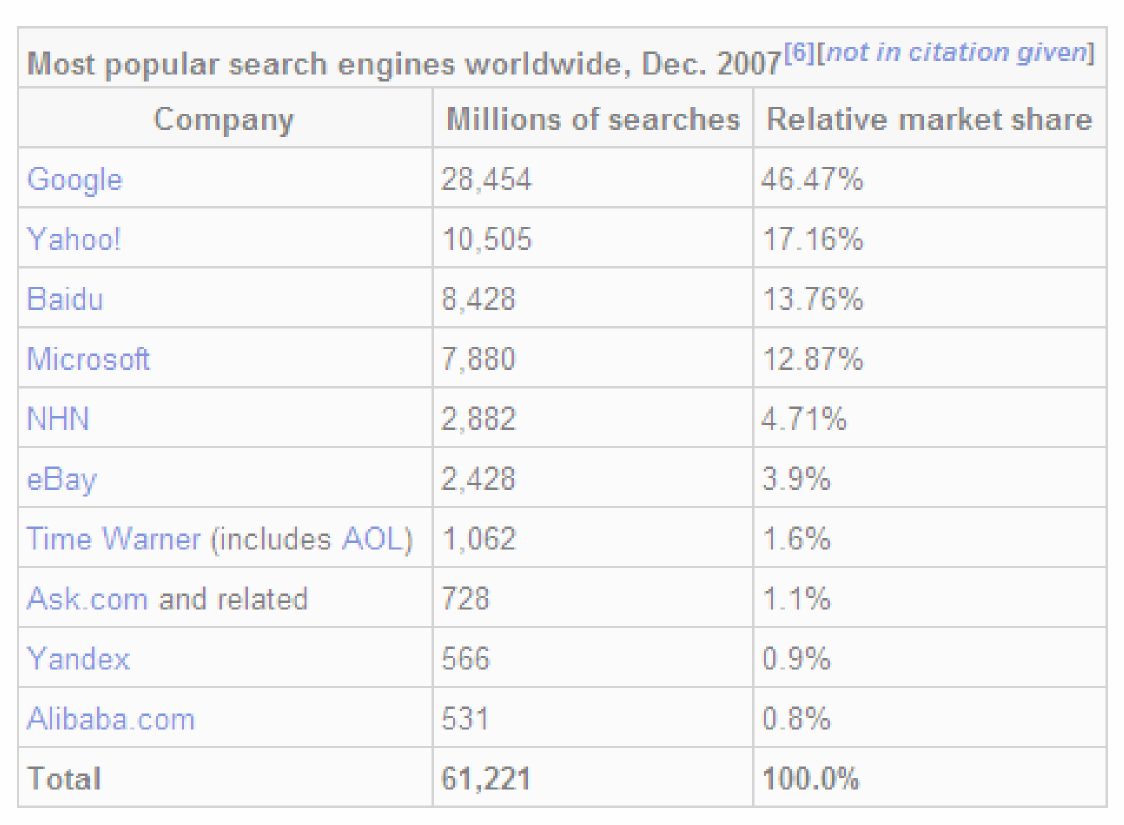
Search engines and web directories

- Differencies between search engines and directories
- Search syntax (*Google* y *Altavista*)
- Search strategies
- Evaluation criteria

Search engines

- Index millions of web pages
- How they work:
  - They work by storing information about many web pages, which they retrieve from the WWW itself.
  - Generally use robot crawlers to locate searchable pages on web sites (***robots*** are also called *crawlers*, *spiders*, *gatherers* or *harvesters*) and mine data available in newsgroups, databases, or open directories.
  - The contents of each page are analyzed to determine how it should be indexed (for example, words are extracted from the titles, headings, or special fields called meta tags). Data about web pages are stored in an **index** database for use in later queries.
  - Some search engines, such as Google, store all or part of the source page (referred to as a cache) as well as information about the web pages.
  - Unlike directories, which are maintained by human editors, search engines operate algorithmically.
  - They allow free browsing.
  - They allow faster updating

Search engines



Most popular search engines worldwide, Dec. 2007[6][*not in citation given*]

Company	Millions of searches	Relative market share
Google	28,454	46.47%
Yahoo!	10,505	17.16%
Baidu	8,428	13.76%
Microsoft	7,880	12.87%
NHN	2,882	4.71%
eBay	2,428	3.9%
Time Warner (includes AOL)	1,062	1.6%
Ask.com and related	728	1.1%
Yandex	566	0.9%
Alibaba.com	531	0.8%
Total	**61,221**	**100.0%**

Search engines

- Information retrieval non-intuitive
- User must plan and design search and browse strategies
- User training to:
  - Information need -> representation of query
- Result: list of documents dynamically created as a response to the information need represented.

Search engines

- Which search engine should I use?:
  - exhaustivity
  - flexibility and quality of query language
  - relevance of results
  - Value added services (email, news, groups, chat)
  - periodicity of updating database
  - velocity of information retrieval
  - User’s habits

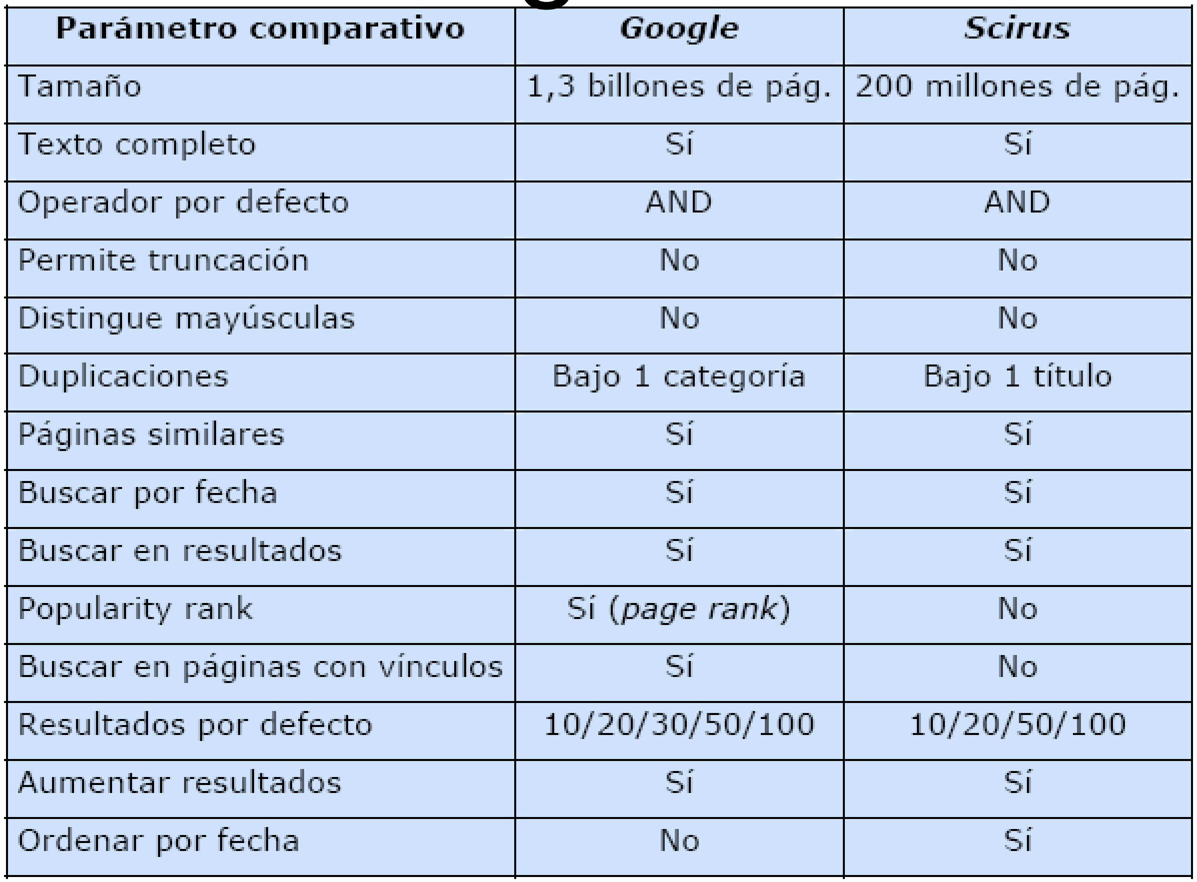
Search engines

- Google (www.google.es)
- Google Scholar: (http://scholar.google.com/) (academic bibliography, peer reviewed articles and papers, doctoral thesis, books, abstracts, technical reports, etc.)
- Altavista (www.altavista.com)
- Alltheweb (http://www.alltheweb.com/)
- Ask Jeeves (http://es.ask.com/?o=312)
- Netscape search (http://channels.netscape.com/search/default.jsp)
- Wanadoo (www.wanadoo.com)
- Lycos (www.lycos.es)
- Teoma (www.teoma.com)

Classic and alternative search engines

- Classic search engines (see previous slides)
- **Alternative** search engines
  - Well-defined documentary background
  - Scirus (engine *“for scientific information only*”) (www.scirus.com)
  - HighWire Press (http://highwire.stanford.edu/)
  - Google Académico (http://scholar.google.es/)

Classic and alternative search engines



Parámetro comparativo	***Google***	***Scirus***
Tamaño	1,3 billones de pág.	200 millones de pág.
Texto completo	Sí	Sí
Operador por defecto	AND	AND
Permite truncación	No	No
Distingue mayúsculas	No	No
Duplicaciones	Bajo 1 categoría	Bajo 1 título
Páginas similares	Sí	Sí
Buscar por fecha	Sí	Sí
Buscar en resultados	Sí	Sí
Popularity rank	Sí (*page rank*)	No
Buscar en páginas con vínculos	Sí	No
Resultados por defecto	10/20/30/50/100	10/20/50/100
Aumentar resultados	Sí	Sí
Ordenar por fecha	No	Sí

Web directories

- Link directory
- How they work:
  - They specialize in linking to other web sites and categorizing those links.
  - Retrieve only a small part of web resources
  - Human-edited databases created and maintained by editors who add links based on the policies particular to that directory.
  - They work by surfing through categories
  - Directed search (by trial & error)
  - Categories are hierarchically organized
  - Slower updating
- Result: a list of documents previously included within a category
  - Documents included within a category are subject-related to that category

Web directories

- Yahoo (www.yahoo.com)
  - since 1994
  - More than 1 mill. sites
- Excite (www.excite.es)
- Open Directory Project: Dmoz (www.dmoz.com). ODP is significant due to its extensive categorization and large number of listings and its free availability for use by other directories and search engines
- BULB (*Bulletin Board for Libraries*): directorio especializado (http://www.bubl.ac.uk/)

Metasearch engines

- Search engine that sends user requests to several other search engines and/or databases and aggregates the results into a single list or displays them according to their source.
- Enable users to enter search criteria once and access several search engines simultaneously.
- Operate on the premise that the web is too large for any one search engine to index it all and that more comprehensive search results can be obtained by combining the results from several search engines.
- Advantages: It may save the user from having to use multiple search engines separately.
- Disadvantages: fewer possibility for shaping searches
  - Metacrawler (http://www.metacrawler.com/)
  - Buscopio (http://www.buscopio.net)
  - Ixquick (http://www.ixquick.com)
  - Cyber 411 (www.cyber411.com)
  - Copernic (www.copernic.com)
  - Mamma (http://www.mamma.com)
  - Dogpile (http://www.dogpile.com)
  - Buscamúltiple (http://www.buscamultiple.com)

Multisearch engines

- Variant of metasearch engines.
- Windows from different search engines are showed within the same screen, so that the user may chose where to browse.
- MySearch: http://ks.mysearch.myway.com/search/default.jhtml
- Twingine: http://twingine.com/
- GuitarraNet: http://www.guitarra.net/buscax.htm
- TheInfo: http://www.theinfo.com
- Multibuscador de Antonio González: http://gva1.dec.usc.es/~antonio/otros/multibuscador.html

Other resources

- Web portals
  - Web portals often function as a point of access to different information on the www.
  - Present information from diverse sources in an unified way. Aside from the search engine standard, web portals offer other services such as e-mail, news, stock prices, infotainment and various other features.
  - Portals provide a way for enterprises to provide a consistent look and feel with access control and procedures for multiple applications, which otherwise would have been different entities altogether.
  - An example of a web portal is Yahoo!
  - Types of web portals:
    - General
    - Especialized
    - Corporate (UA)

Other resources

- Services of selective difussion of information (DSI)
  - The allow users to define a profile
  - They send information to user according to the profile defined
  - Most of them need subscription
  - Mynewsonline (http://www.mynews.es/)
- Subject-field databases
  - Allmovie (http://www.allmovie.com/)

Other resources

- The invisible web
  - Large portion of the web that is not picked up by robots that search engines use to find new sites and can be valuable in helping reseracher s locate new information (databases, etc.)
  - Causes:
    - Format of documents
    - Dynamic way of generating some web pages
    - Intranet sites

Search engine syntax

- Search strategies:
  - Basic
  - Logic or boolean expressions
  - Filters
- Search engines do NOT use the same search syntax
  http://www.lib.berkeley.edu/TeachingLib/Guides/Internet/SearchEngines.html#NewSearchEngines

Basic search operators

- +: industria glosario
- -: fabric –textile –clothes
- “...”: “science and technology”

Logic search operators

- AND: *bottom* AND *sole* AND *shoe*
- OR: *vamp* OR *front* OR *upper*
- AND NOT: *fabric* AND NOT *textile*
- (): (zapato AND suela) AND (pala OR copete)

Filter search operatos: Googles

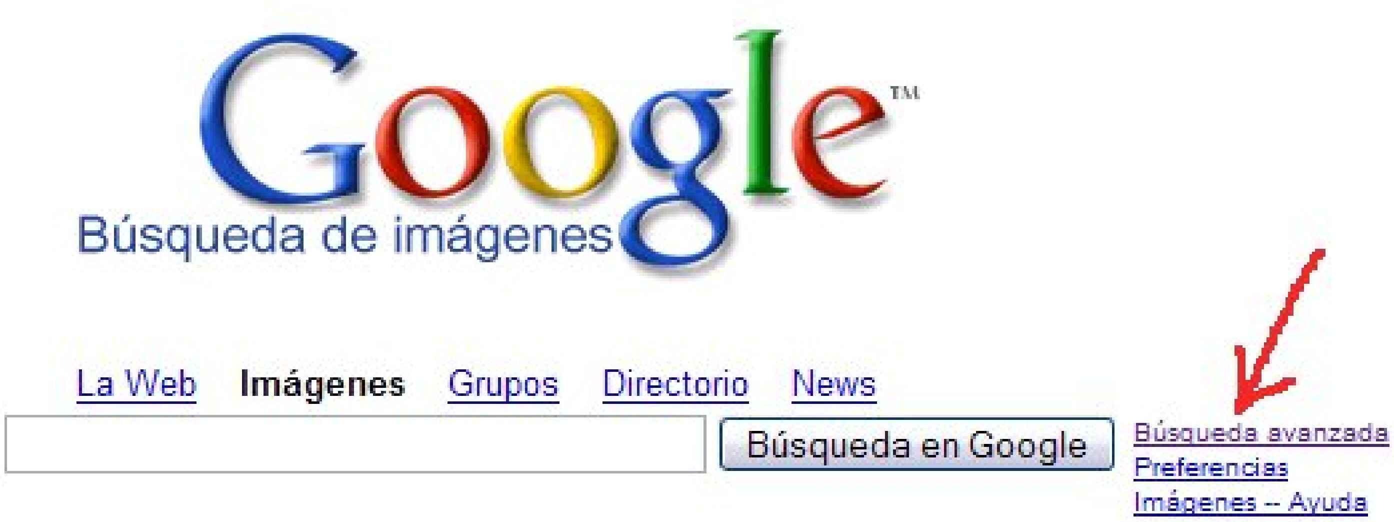
- **site:***sitename* (searches within the domain specified; p. ej. ***site:.***net; .org, .gov, .es .de, .uk.).
- **link:***URLtext* (p. ej. ***link:www.cueronet.com*** to find sites with links to that page).
- **allintitle:*text*** (pages containing the query word within the title of the document).

Filter search operatos: Googles

- **allinurl:***text* (limits searches within a domain and a subdirectory).
- **filetype:***filetype* (limits the kind of file type; p. ej. ***filetype:pdf***).
- **allintext:***text* (pages containing the query word within the body of the document).
- **define:***text* (to find definitions of terms; p. ej. ***define:framework***).

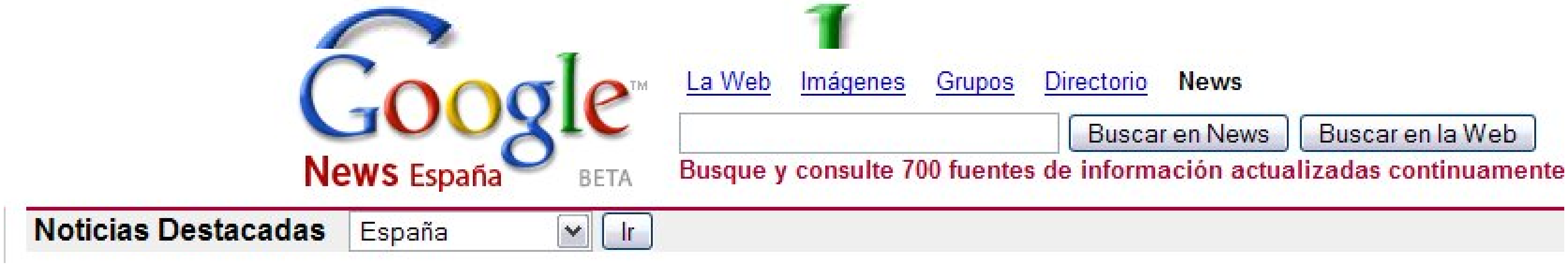
Google tools

- Images
  - Advance search: format of images, size, domain, etc.

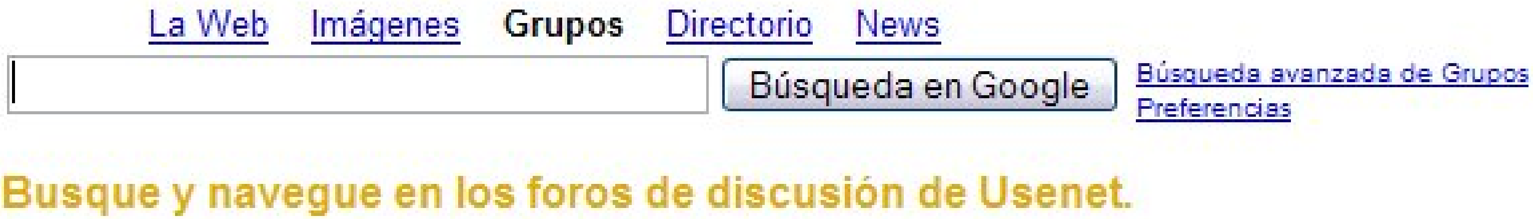


Google tools

- Groups and News
  - Search in specific fields and on subject-field matters.



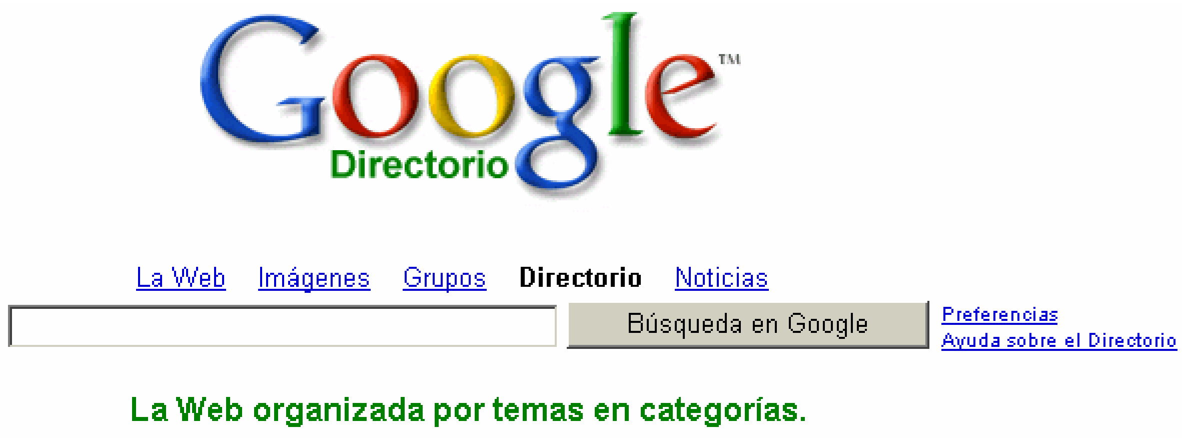
Busque y consulte 700 fuentes de información actualizadas continuamente



Busque y navegue en los foros de discusión de Usenet.

Google tools

- Directory
  - – Limits queries within a specific category



La Web organizada por temas en categorías.

Search syntax- Altavista

- **domain:***domainname* (p. ej. ***domain:****.edu*).
- **link:***URLtext* (p. ej. ***link:www.cueronet.com***).
- **title:***text*.
- **host:***name* (pages within a specific host; p. ej. ***host:eu.int*** will retrieve pages in server **eu.int**).
- **image:***filename* (p. ej. ***image:elvis*** retrieves pages called “elvis”).

Search syntax- Altavista

- **url:***text* (p. ej. ***url:glosario*** to find all the pages containing the word “glosario” in any part of the host name or file name).
- **text:***text* (to retrieve pages with the specific text within any part of the page).
- **anchor:***texto* (to find pages with the key word or expression within the link text; p. ej. ***anchor:"Consulte AltaVista"*** you will find pages containing "Consulte Altavista" as a link.

Evaluation criteria

- Why evaluating Internet information ?
  - Anyone can post anything on the Web
  - Anonimous and anarchic nature of the WWW.
  - Lack of updating.
  - No quality filter to sift inaccurate or biased ifnormation from reliable information.
- Need to evaluate **veracity**, **relevance**, **credibility** and **accuracy** of data.

Parameters to evaluate quality

- Timeliness and maintenance
- Content and coverage
- Authorship and source

Timeliness and maintenance

- **When was it written? When was it last upated?**
- This parameter depends on our needs.
  - When discussing current technology, for example, we’ll need to locate the very latest information. However, if researching a subject with a broader scope, we may find that other criteria will be more helpful.
- Indicators:
  - Date of *copyright*
  - Date of *last updated*

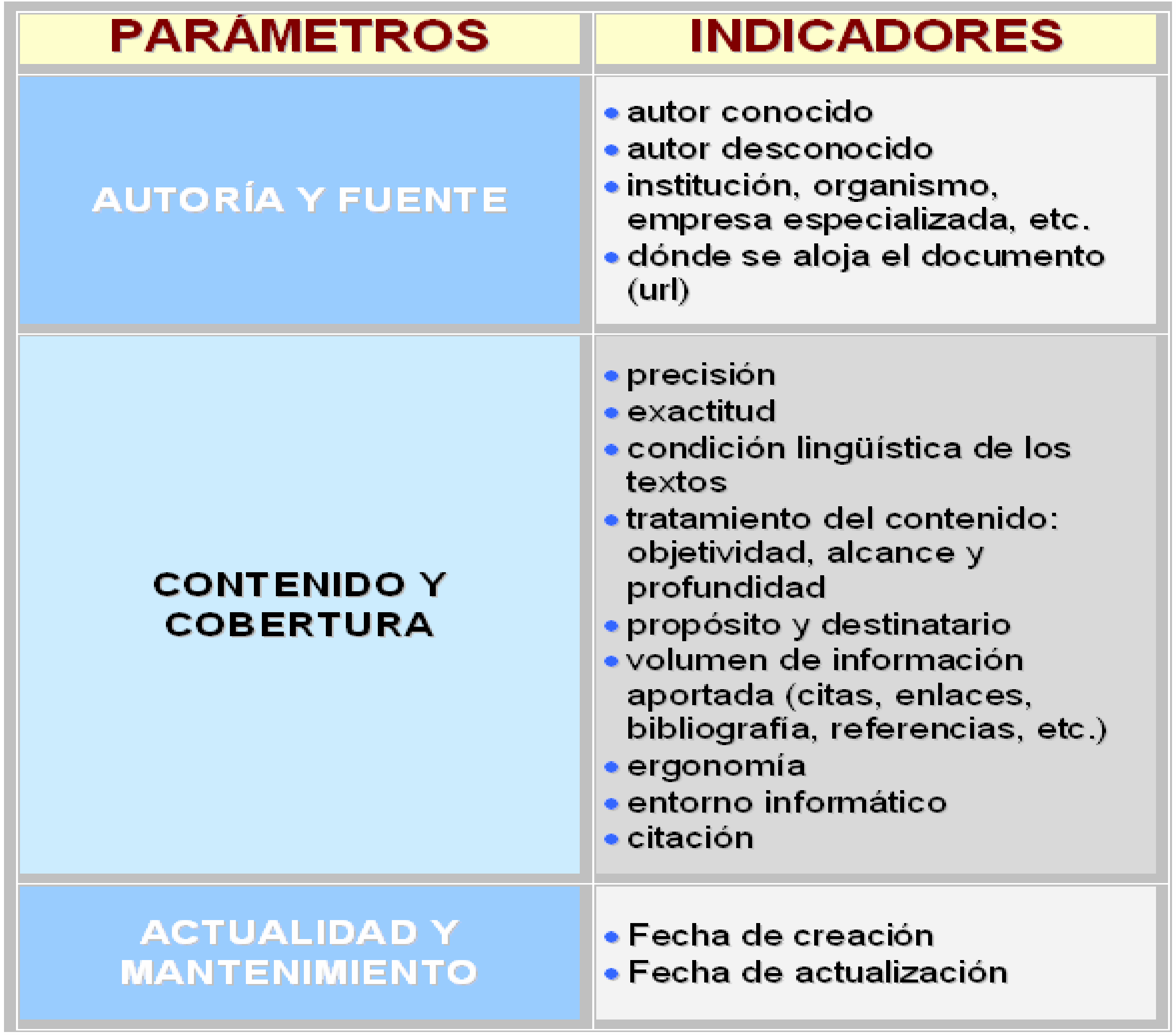
Content and coverage

- **Is information rigorous and accurate? (Quality)**
- Indicators:
  - Language and linguistic condition of texts
    - Popularization of scientific, technical, professional and academic knowledge in English
    - Fewer documents in other languages
    - Translations from other languages
  - Factuality of data
- **Is information complete? (Quantity)**
- Indicators:
  - Volume of information and additional information provided (dates, authors, figures, links, tables, pictures, graphics)

Authorship and source

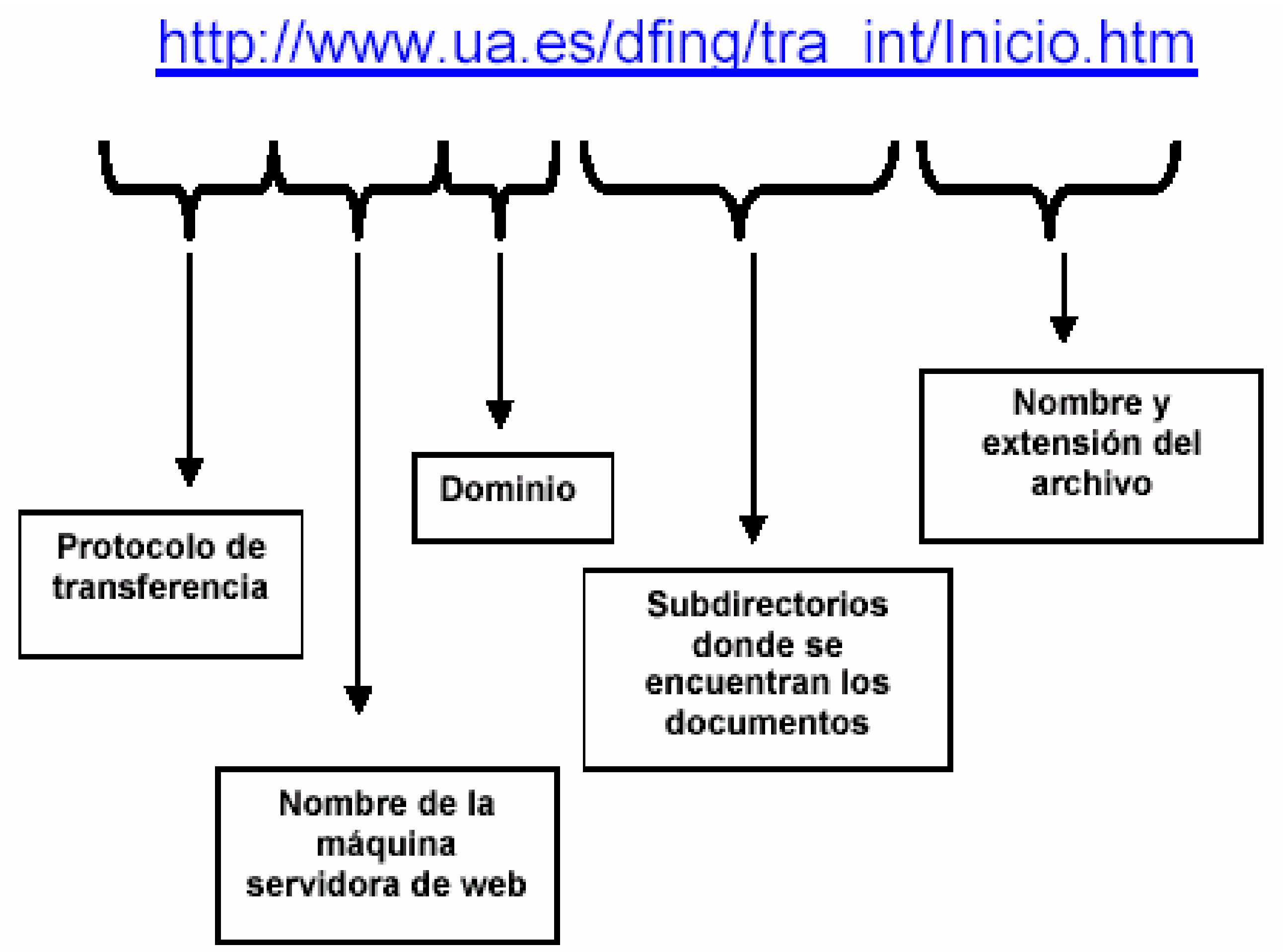
- Who wrote the source? What are the author credentials?
  - Indicator: **Autor**
    - Well-known author,
    - Unknown author
    - Institution, organization, company, etc.
- Where is the document located?
  - Indicator: **URL**
    - Personal page
      - *members, users, home, usuarios, people...*
    - Institution, organization, government agencies, etc.
    - Domains
      - *edu., com., org., net., gov. uk., es...*
- Competent server
- Appropriate domain name according to subject
- Experts cited, other sourced, listed bibliography (*Additional sites, Related links, About us, Biography Background, Philosophy, Who am I*).
- Truncate URL
  - www.ua.es/dfing/tra_int/Recursos.htm
- What do others say about the page?
  - “nombre * apellido”
  - `link`:ua.es/dfing/tra_int/Recursos.htm

Quality parameters and indicators



PARÁMETROS	INDICADORES
AUTORÍA Y FUENTE	• autor conocido • autor desconocido • institución, organismo, empresa especializada, etc. • dónde se aloja el documento (url)
CONTENIDO Y COBERTURA	• precisión • exactitud • condición lingüística de los textos • tratamiento del contenido: objetividad, alcance y profundidad • propósito y destinatario • volumen de información aportada (citas, enlaces, bibliografía, referencias, etc.) • ergonomía • entorno informático • citación
ACTUALIDAD Y MANTENIMIENTO	• Fecha de creación • Fecha de actualización

Anatomy of a *URL*



http://www.ua.es/dfing/tra_int/Inicio.htm

Protocolo de transferencia

Nombre de la máquina servidora de web

Dominio

Subdirectorios donde se encuentran los documentos

Nombre y extensión del archivo

Other parameters

- Why is the page/site on the WWW?
  - Author intentions, advertiser’s bias
- Is it the WWW the best place to find the information I am looking for?
- Always CHECK, CHECK, CHECK