

11. WEB SEARCH

The module conveys the necessary information for searching on Web information regarding a certain subject. Specific aspects of that type of information are shown, as well as the main types of the search tools that are at the user disposal (specific aspects, advantages, disadvantages and using opportunities being detailed). The module also contains recommendations for organising the information search activity, technical details regarding the way the search subject is formulated, indications concerning the evaluation criteria of the found information and recommendations regarding the quoting way of the bibliographical sources.

11.1. Introduction

The huge quantity of information existing on the Web and the specific manners in which this is being stored leads to certain peculiar search modalities of some sources referring to certain information. This chapter also includes some introductive issues referring to the very general principles after which the Web search of information can be unfolded.

11.1.1. *Can we search on the whole Web?*

The finding on the Web of the desired documents (pages, sites) can be very easy sometimes and other times it may seem almost impossible. This last problem can occur due to the huge WWW dimensions, that is estimated to contain approximately 25 thousand million documents (March 2009), or because of the fact that this information quantity is not catalogued in any way.

Unlike the search in the library's catalogue, in which the books are grouped by certain topics, the Web search is fulfilled by trying to guess what words could contain the pages we would like to find, or what topics, or categories somebody chose in order to group the information on many pages referring to a certain field. When we start what it is called the "Web search", it is impossible to search directly in this one. The Web represents the totality of an immense number of pages existing in computers all over the world. A computer cannot identify them and it cannot connect to all of them. It is impossible for anyone to apply a search on the whole Web. Any affirmation of this type, concerning the possibilities of a certain search tool, can only lead to the truth distorter.

11.1.2. *What can we hope for?*

What we can achieve by using a computer is to access one or a few of the available search tools. Usually we access the search tool of a data base or the one of a site collection, this representing in fact a trifling part of the whole Web.

The search tool provides us links to other pages. By accessing the latter ones, we are going to have documents, images and sounds from the whole wide world at our disposal.

There are many search tools categories, but only four of them are being generally recognized. In order to exploit the tools' potential of each category, there are different strategies that can be used.

11.2. Search engines

The objective of the chapter is to facilitate the understanding the way that search engines function. The search engine notion is presented and the way this gathers information from its own data base. It is described more precisely the meaning of the "search" notion by using such an engine. The advantages and disadvantages of using such tools are explained and some criteria that can differentiate them are presented.

11.2.1. *What are the search engines?*

The search engines are large data-bases, which contain Web pages files. The respective data bases are automatically constructed, with a minimum volume of human supervising.

The Web pages from the data base of a search engine are not organized on categories, but they are classified by a range of algorithms, the whole content of the original pages (word by word) from the Web being within them.

The search in the data base is made by specifying certain words – key words. The search engine offers as result the Web addresses of the pages from its data base in which there were found the respective words. The accessing of Web address given as a result by a search engine leads to the “navigation” toward the original page.

Although there are also specialised search engines, of lower dimensions, Web pages data bases have enormous dimensions, and very often they offer as a result of the search a huge quantity of information. It is estimated that some data bases contain information regarding over 90 % from the available Web pages. For this reason, some search engines allow performing the “under-searches”, only in the set of results obtained in the previous search.

11.2.2. *How do they construct the data base?*

The search engines construct their data bases with the help of the “spiders” or of the “robots”, specialised programmes that search through Web from link to link, identifying and searching the pages.

The search made by the “robots” consists in visiting the Web addresses from the pages that already exist in the data base of the search engine. A “robot” cannot effectuate judgments in order to imagine a new address (URL).

If the address of a Web page does not exist in any other page, it cannot be found. The owners of the new pages, that are not taken into consideration in other pages yet, can send the new addresses to the search engines, so that the eventual inclusion in the data base could be determined.

Once a “robot” started searching a certain Web site, it visits all its available pages. The moment it identifies a page, the “robot” transfers the page content toward another program. The latter “indexes” the page, by storing information referring to the words within this one in the data base, referring to the eventual present links or to other characteristics of the page content that will allow its identifying in case of a search.

11.2.3. *What does the search mean in fact?*

Every time a user accesses a search engine, specifying the key words or even phrases after which the search is to be made, it requires him to verify the existence of the respective texts in all of the Web pages from its data base.

It is important to memorize the fact that when a search engine is used, a Web search is not made as it exists in the moment of the search. The search is made on a part of the whole Web, part which has been copied in a data base at an anterior moment of the search.

It is difficult to verify how long it takes to add a web page to a search engine. Usually the “robots” visit again the original pages, in order to notify the eventual modifications, but the process of actualizing the information from a data base depends on the periodicity with which these revisiting are made, and on the necessary time for the proper actualization of the data base index.

The search engines have a “policy” after which they exclude certain Web pages or links.

11.2.4. *The advantages and disadvantages of the search engines*

The search engines allow the access toward a pretty extended part from the multitude of the available public Web pages, multitude that is in a continuous exponential development.

For the moment, the search engines represent the most efficient means in order to effectuate a Web search. Without their presence, in the multitude of the existent information at the global scale, it would be impossible to discover the needed part at a given moment.

On the other hand, the immense number of the words indexed by a search engine expands the probability for this to offer hundreds of thousand of irrelevant answers, as a result of a simple request. As a matter of fact, a search engine can offer the address of a large dimension document, even if the required word appears in it only for a single time.

11.2.5. *Differences between the search engines*

The search engines have specialised programmes that search the key words or the phrases in the data base and then offer results, usually grouped after a criterion which is called "relevance".

Although these programs can use similar algorithms, two distinct search engines will not offer the same search options and neither have they had the same speed characteristics of the answer, nor the same algorithms of establishing the relevance. The results of a search will be different, depending on the used search engine, and the differences can be sometimes significant. Recent estimations show that approximately 60 % of the results offered by a search engine superpose the results of different engines.

In order to appreciate the relevance of a page, a search engine has a set of rules, set which can vary from an engine to another.

Among the most used rules, there are those referring to the location on the page and to the frequency the key words or phrases appear on the Web page and even on the HTML META part (tag) of this one. The title of the page, its "header" and the areas situated at the beginning of the content are also verified.

Some search engines estimate the "popularity" of a page by the number of the existent links toward it from other Web pages.

11.3. Decision search engines

The objective of the chapter is to facilitate the understanding the decision search engines. The decision search engine concept is presented. The advantages and disadvantages of using such tools are explained below.

11.3.1. *What are the decision search engines?*

Today more and more users search the web as part as they job tasks. Also, the scientists are very concerned about the time spent discovering quality information. For these users, the simple list of search results is not enough. Scrolling the list is a tedious and inefficient activity. They need something better.

So, a new kind of search engines was generated. They were called decision search engines. They offer a more organised experience and simplify tasks and provide insight to web pages conducting to faster and more confident decisions.

The decision search engines, built to go beyond today's search experience, include deep innovation on core search areas including entity extraction and expansion, query intent recognition and document summarization technology as well as a new user experience model that dynamically adapts to the type of query to provide relevant and intuitive decision-making tools.

Also, the decision search engines offer specialised searching in domains like: images, videos, maps, news, health, shopping, travel, etc.

11.3.2. *The advantages and disadvantages of the decision search engines*

Some of the advantages of the decision search engines are:

- structuring the results;
- finding instant answers quickly;

- reduced number of clicks;
- deep links (offers links direct to the relevant page within a site);
- quick view.

An important disadvantage of decision search engines is that they are a little bit slower than a usual search engine. (See Bing.com versus Google.com)

11.4. Metasearch engines

The objective of the chapter is to facilitate the understanding the way the metasearch engines function. The metasearch engine notion is presented and the way it realizes a list of information sources referring to a certain subject. The advantages and disadvantages of using such tools are explained below, and also the differences in relation with the search engines and some suggestions regarding the using opportunity are offered here.

11.4.1. *What are the metasearch engines?*

In order to construct their own data base, the metasearch engines do not visit the multitude of the Web pages, but they access the data bases of many individual search engines.

The accessing of the individual engines set is made at the moment a user performs an interrogation by using an interface given by the metasearch engine.

The metasearch engines have usually two ways of transmission the search results to the user:

- In a single list, the results offered by the different individual engines being included into it after an anterior removal of the doubles;
- in separate lists, depending on the individual engine from which the results came, key that leads to the appearances of the same result on many lists.

Some metasearch engines permit the refining of the search (the search in a result list), the specifying of the individual engines to be accessed, the maximum time allotted for accessing each individual engine a.s.o.

Sometimes the facilities given by some metasearch engines need software installing on user's own computer, but in most of the cases the rolling is made on the host server.

11.4.2. *The advantages and disadvantages of the metasearch engines*

A metasearch engine permits individual engine identification, in which a possible search would lead to best results, but it does not have the search configuration options, that the individual engines offer.

Although they appeal to an important number of individual engines, most of the existent metasearch engines neither access Google, one of the largest data base engine, and nor the individual Northern Light engine, useful to the academic search. Similar to a supermarket, the metasearch idea is very tempting, but its implementation leads to the ability limitation of controlling the places to search into, this fact being one of the biggest disappointments made by the metasearch engines.

Though being very fast, there is the opinion that metasearch engines tend to be more and more limited in covering areas, relying on the results obtained from subject organised lists or from pay per view engines, generally speaking.

The metasearch engines do not perform "meticulous" searches, the speed coming from the superficial way of working. It is estimated that a metasearch engine offers approximately 10 % of the results that it finds into an individual engine which it accesses. Moreover, they have an unpredictable behaviour in case of complex searches.

11.4.3. *When are the metasearch engines used for?*

The use of the metasearch engines is recommended in cases in which the getting of the results has to be done fast, and they become useful when a relative low volume referring to a single subject or to a well defined term is required.

Efficient results can be also obtained in the case in which, for a simple search, the previous individual engines searches did not lead to the getting of useful documents.

11.5. Subject directories

The chapter facilitates the understanding of the way the subject directories function. The subject directories notion is presented below, and the way in which the information sets are constructed, that are at the users' disposal. It is made a classification of the main categories, the advantages and disadvantages of using these tools are explained and some suggestions concerning the opportunity of use are also offered.

11.5.1. *What are the subject directories?*

The subject directories (the guides), unlike the search engines, are created and maintained by human editors, and not by electronic "robots". The editors survey and select the Web pages that are to be included in guides, by using selection criteria, which have been previously established. Usually, the address lists within such a guide contain the editor's annotations. It should be mentioned the fact that the accuracy of the evaluation and the promptness of the bringing up to date decrease while the dimensions of a guide increase. The guides are of lower dimensions than the data bases of the search engines, usually having only the address of the home page. Many guides have search engines in their own lists, these engines having also the search option on the Web.

11.5.2. *How do the subject directories function?*

When a search is made on a subject guide, the key word or the phrase are searched inside the description (annotation) that is annexed in every address from the list. The subject guides can be classified on different categories, from those that have ordinary subjects to those that have academic, commercial, portals, and even vortal themes. The portals represent subjects that have been created or transformed for commercial purposes, and after that having been reconfigured in order to function as "gates" towards Web. Among the lists referring to the most popular subjects, the portals offer additional services as e-mail, news, stock exchange, travelling information, maps, a.s.o. The vortals, the vertical portals represent more specified guides, avoiding the extended range of subject and links that can be usually found into a portal. The differences between the subject guides and the search engines are continuously dimmed. Many guides use search engines in order to interrogate their own data base or to search for the additional sources of information on the Web. Many search engines are not also limited to the specific mechanisms, but they include (in a commercial manner) subject guides or they create their own guides.

11.5.3. *Advantages and disadvantages*

Usually the guides' subject editors organize the information hierarchically, on categories and subcategories through which the navigation is possible, using general terms. The passing through several levels in order to reach the address of the page is sometimes uncomfortable, but this way of structure represents the important part of the guides. The human supervising of the guide guarantees high quality content, the search into a guide also offers, in comparison with the search engines, fewer results that are not going to harmonize in the context of the subject approached. Unlike the search engines, the subject guides do not contain their own data bases. They do not store pages, but only their addresses, their titles and short descriptions, and the search can be made only in a limited environment of this information.

The situation can bring about problems sometimes, because once the Web page has been accepted and included on the list of the guide, its content can be changed without the guide's editor realising this. Thus, the guide can keep on offering an address of a page that has been moved or that no longer exists. Such "dead" links represent a real problem of the subject guides, making them demonstrate a "guidance" trend of the user toward the e-commerce sites.

11.5.4. When are the subject directories used?

The subject guides are similar to the telephone books or to publications of "yellow pages" type, being useful sources of information referring to general interest aspects, organisations, and commercial issues or related to certain products. Usually, the subject guides are used when information regarding certain peculiar aspects from an interest area is searched.

11.6. Gateways and data bases

The chapter facilitates the understanding of the way the gateways and the specific subjects' data bases function. The gateway notions are presented below, and the way in which the sets of information, which they put at the users' disposal are constructed. It is made a classification of the main gateway categories, and some suggestions, concerning the opportunity of use, are also offered. The "invisible Web" notion is described below, and the main types of documents and information sources that are part of this category are also described, as well as the motifs that lead to these characteristics.

11.6.1. What are the gateways and the specific subject data bases?

There are two types of gateways: the libraries and the portals. They represent data bases collections and informational sites, organized on subjects that have been assembled (constructed), checked again and recommended by specialists, who are usually librarians. These collections support the searching activities or the bibliographical studies, identifying and recommending the academic orientation Web pages.

The specific subject data bases (the portals, the vertical portals) are devoted to a single subject, being created by teachers, researchers, experts, governmental agencies or interest groups of affairs. The specialists involved, have high level professional knowledge in the respective private domain and they possess significant quantities of information and data.

11.6.2. When are they used?

The library gateways are usually recommended to be used when very good quality information is required, and this is ensured by the taking the census of the sites' content and by their good evaluation, made by specialists.

The subject orientated data bases are recommended in case the search domain has a special feature, for example in the news' case, in multimedia links, archives, lists of addresses, personalities, job offer a.s.o.

More and more "main-stream" search engines, subject directories and portals comprise on their welcome pages links toward subject orientated data bases.

11.6.3. The semantic web

The **Semantic Web** is an evolving development of the World Wide Web in which the semantics of information and services on the web is defined, making it possible for the web to understand and satisfy the requests of people and machines to use the web content. It is meant to be an universal medium for data, information, and knowledge exchange.

At its core, the semantic web comprises a set of design principles, collaborative working groups, and a variety of enabling technologies. Some elements of the semantic web are expressed as prospective future possibilities that are yet to be implemented or realized. Other elements of the semantic web are expressed in

formal specifications. Some of these include Resource Description Framework (RDF), a variety of data interchange formats (e.g. RDF/XML, N3, Turtle, N-Triples), and notations such as RDF Schema (RDFS) and the Web Ontology Language (OWL), all of which are intended to provide a formal description of concepts, terms, and relationships within a given knowledge domain.

Humans are capable of using the Web to carry out tasks such as finding the Spanish word for "monkey", reserving a library book, and searching for a low price for a DVD. However, a computer cannot accomplish the same tasks without human direction because web pages are designed to be read by people, not machines. The semantic web is a vision of information that is understandable by computers, so that they can perform more of the tedious work involved in finding, sharing, and combining information on the web.

Semantic publishing will benefit greatly from the semantic web. In particular, the semantic web is expected to revolutionize scientific publishing, such as real-time publishing and sharing of experimental data on the Internet. This simple but radical idea is now being explored by W3C HCLS group's Scientific Publishing Task Force.

11.6.4. *What is the invisible Web?*

There is a wide part of the Web which is not accessible to the search engines, and cannot be catalogued by it. It was called the Invisible Web or the Deep Web and comprise, among the others, sites protected by passwords, documents hidden by "firewall" servers type, pdf files, archived materials, tools, interactive tools of computers and dictionaries type, some data bases content with own search facilities (e.g. UCB Library Catalogue Pathfinder or other libraries' catalogues) a.s.o.

Those that deal with the Web characterisation, appreciate that their invisible parts, made of thousands of such documents or data bases, represent approximately 60 – 80 % of the Web existent material.

These resources are not accessible to the search engines, being integrated in Web sites, in which the information is not static, but formed in a dynamic shape, as a result of the users' requests.

The invisible Web information categories are very diverse, starting from those of general or commercial interest and ending with the strong ones, dedicated to a certain subject.

11.6.5. *How can it be accessed?*

The accessing an invisible Web zone can be made only by specifying its address directly in the browser. The respective arias are often accessed by the human operators who administrate the library gateways or data bases orientated on subject, and rarely some subject directories, representing in this way valuable sources of addresses toward the invisible Web.

Librarian's Index, Yahoo!, AcademicInfo and Infomine represent the best examples of directories and subjects in which the invisible Web data bases addresses can be found.

11.6.6. *Why are some pages "invisible"?*

There are two motifs for a Web page not to be included in a search engine: the technical barriers that can break the access to that page or an exclusion decision on that respective page.

If the single way to access a Web page is that of a visitor who has to write on the keyboard a set of characters or to select a certain option, the respective page cannot be found by a search engine.

The search engines' data bases are created by robot programs that "search" on Web, visiting only the pages whose addresses exist on the pages that are already in the data base. If the address of a page does not appear on search engine data base page, it will not be visited by the "robots", because they cannot "write" words in order to make possible the accessing of the respective page.

The invisible Web also includes a category of pages which the companies administrating the search engines exclude them for the very beginning and not because of lacking of the technical means. The exclusion criteria come from reasons connected to the huge dimensions of the search engine data bases, these ones requiring already high costs for their maintaining and bringing relatively low profits.

11.6.7. The pages generated dynamically are invisible

The data bases, for example, display the search results on the dynamically constructed Web pages, not stocked anywhere else, this way of work being easier and cheaper than storing on fixed pages that are to contain all the answers to every possible search, that the data base's users are about to effectuate.

The pages created as a result of the search in the data base are called "generated dynamic" pages. The results of a search are comprised in a page that is sent to the computer the search was commended from.

Several times, the generated dynamic pages are not stored, because it is almost impossible for the results which they contain to satisfy the search requests of other users, too. For the data base of the search engine, the dynamic regeneration of such a page is more efficient than its storing.

Unlike the generated dynamic pages, the "static" pages are stored on servers and they are identified through a unique URL. The robot programmes can find a static page if it's URL already exists in the search engine data base.

11.6.8. The content of the data bases is invisible

The content of specialized data bases, that allow searches inside them, can be entirely or only partially visible, depending on how many of its addresses exist on static pages.

Among the examples of the data base that allow searches, we can mention Google search engine, Northern Light or AltaVista. The content of the on-line libraries' catalogue, that does not need an accessing password (for example UCB's Pathfinder), is also included on the invisible Web.

The results of a search are generated dynamically. Sometimes, it is possible to hold the URL which the results list appears under, and to be used in order to regenerate the page created dynamically. But the respective page is not stored anywhere.

A site example with an accessible content, by direct search or through addresses searched by the robots is Yahoo! There are also many others directories, organised so that they could permit the search, but they also permit the access to their content by following some addresses. The "robots" will be able to use this second facility, while a human user will be able to use also the search facilities that are personal to the directories.

11.6.9. How is the exclusion of a page decided?

The exclusion decision of a page can be made because of the respective page format, when the respective format was not enough frequently searched by the search engine's users, or when the search was not finished successfully. Although there are no technical motifs for such exclusion criterion, this is used by many companies that maintain the search engines.

The data bases and the search engines robots are optimized for the HTML pages work the basic Web language. But there is a multitude of documents in different HTML formats. A HTML page may contain an address to such a document, but it cannot display their whole content in a text form, content which exists in a certain format. The pages that contain only images are often omitted by the search engines, because their content cannot obviously handle a key word search.

11.6.10. *Excluded pages because of their format*

PDF format files or those that contain HTML text format in a very low proportion (Flash, Shockwave, Word, WordPerfect, and PowerPoint).

Exceptions:

- Google gives the possibility to search in PDF files, converting them into text and introducing the respective text in a HTML format, for it to be stocked in the search engine data base. A search that leads to such a “translated” text will offer as a result the original address of the PDF file.
- The images data bases given by Google, AltaVista and other companies that maintain search engines, are purposely organized in order to handle file types which the text weight is lower within.

11.6.11. *The script pages are excluded*

A script represents a language type programme that can be used for accessing and displaying the Web pages. There are many types and ways to use the scripts. They can be used for the total or partial creation of a Web page or for the communication with a search facility data base. Many of the search criteria or the bases’ answers use the scripts.

When a page URL contain the “?” sign, this is a clue that a succession of “script” commands are used inside the respective page. The majority of the search engines are “instructed” not to access the sites or the pages that use the “script” technology, though this thing is technically possible.

When a “?” sign appears in a URL, the search engine “robot” avoids the respective page. The motif is given by the risk of finding inferior quality scripts from the programming point of view, or even “traps” created on purpose. Such a trap can drive the “robot” to an infinite loop, thus wasting the time allotted for the search. Because a “robot” does not have the ability to avoid such inconveniences, the exclusion of such pages can totally or partially hinder the search engine presence in the data base of a whole site.

11.7. **The search strategy**

The chapter conveys a series of suggestion regarding the way a search Web should be organised. Some general criteria are presented, and then many search steps are detailed and the most frequent errors that could lead to the diminishing of the efficiency are enumerated below. Search “declaration” forming techniques are also presented, by detailing the binary logical and the search elements of a certain information type.

11.7.1. *The first step*

It is always recommended a period of reflection before starting the search. A search strategy can be made by firstly searching the answer to this question “What would I like to do?”

To navigate only

- In this case in which is desired only the setting up of a general idea on the available information, it is possible to start with a subject directory of the Yahoo! type, and then it can go on by searching a few key words in a mega-motor of the lx quick type.

To localize a certain information only

- A powerful search engine of the Fast (All the Web) Search or Google type can be appealed to, or a special data base like Voice of the Shuttle (for researches in human sciences) or Bureau of the Census (for statistics)

To store all that I find on a certain subject

- In this case the search of the same terms on different search engines is recommended

11.7.2. *Analyse the search subject to decide the start point*

- Are there distinctive words and phrases, such as “methemitha”, which have only a meaning, or “affirmative action”, group of words that has a specific meaning?
- Can't you find words or distinctive phrases? Do you have only common terms, of general use, that are going to offer error results? For example “order out of chaos”, it is used in many contexts, and “sundiata” can mean either a myth or a rock band.
- Are you looking for a vast subject, such as “the Victorian literature” or “alternative sources of energy”?
- Do you search a more detailed aspect of a large interest subject? For example, as part of the automobile recycling you are interested only in the actual research subjects, and in the future impact of design, and not in the oil recycling or other communitarian actions.
- Has the search subject got synonyms, equivalent terms, different orthography alternatives or terminations? for example, the “cold fusion energy” and “hydrogen energy” terms are used by different sources of information without being exactly equivalent. The “millennium” or “millenium”, “millennial” or “millenial” or “year 2000” terms can lead also to different results of the search.

11.7.3. *Choose a start point*

- If you have a distinctive word or an exact phrase that you want to search, firstly you can make a “warm up” search in Google (put your phrase between inverted commas). If the result does not satisfy you, search information referring to a wide concept in a subject directory, in which the notion could match (“what it is all about?”).
- If you DO NOT have a distinctive answer or an exact phrase, you can use a search engine, specifying many terms grouped between inverted commas, or you can navigate using a subject directory in order to identify a series of key words, which you are going to use afterwards.
- If you want a survey of a subject, the search engines ARE NOT recommended. Search for a subject directory with a larger rank of specialization on the respective subject.
- If you want to detail only a certain aspect of a larger domain, search for a subject directory, orientated on this domain or use the “Booleean” mode from the AltaVista search motor.

11.7.4. *Other advices to start the search*

- The subject directory ARE NOT recommended searching for synonyms, equivalent terms or variants. Apply the search in search engines that have facilities like the use of the logical function OR truncating or “Field” limitation type.
- When you have a clear idea concerning the information that you want to search, you avoid using the search engines. Access a subject guide, an encyclopaedia or a virtual library. In a situation of that type, do not avoid asking for help to the competent persons from a library.
- When you need to know data, concrete information, statistics, maps a.s.o., search to access a data base specialised on the invisible Web.
- If the search subject has a high degree of specialisation, firstly search for a subject directory that has the respective profile. Send an e-mail to a page's author that refers to the desired subject. Send a question to a discussion group or ask an expert.
- Never do neglect the chance. Try learning from any experience.

11.7.5. *Further on*

Learn in passing and apply what you have learned

- Do not start from the idea that you know perfectly what you search. Analyze the search results and try to discover clues with which you can adjust your initial opinion.
- Do not stagnate in a search strategy that doesn't offer any results.
- Oscillate between the search engines and directories. Especially search for specialised directories on the subject you search for. Imagine the types of data base that could exist in the domain and search for them.
- Resume the previous strategies better informed.

11.7.6. *Not recommended strategies*

Because of their inefficiency and often because of their venturesome and frustrating results, neither of the two search modalities is recommended:

- navigation through search facilities of subject directories
- Several times navigation is amusing, but it isn't efficient. The use of any search facility, which the respective directory, offers is favourite.

Through directory it is understood a Web resources collection organised on subject categories or with a certain hierarchical structure adequate to the content (subject directories, specialised data bases or gateway).

Through navigation, starting from the initial level, the widest one of the subject hierarchy, successive branches of the hierarchy are covered, and they are chosen on the base of impressions that are often subjective. At each level, the user tries to guess in which of the available categories the searched information could be found, similar to the manner in which periodically a way followed by a crossroad should be chosen more or less by chance.

11.7.7. *Other recommendations*

The use of a directory search tool will lead to the desired terms no matter the categories and the subcategories in which they were classified.

The following of the links toward the recommended site by their great number of accesses or commercial interests.

Several times, the results offered by the search engines are grouped after their great number of accesses toward the users or even after the taxes that were paid by them. Some addresses are even accompanied by "cool" type recommendations.

The fact that a site is visited often does not represent a guarantee that the users searched exactly for the desired information at that time. A site can be perfectly adequate to the search interest even if it hasn't been discovered by the public yet.

There is a series of motifs which can lead to undesired results. Sometimes, the system through which the search engine appreciates the pages' pointing out (system which is very hard to notice) can lead to some key words ignorance from the search of the subject. The fact that this thing takes place most often when the key words' displaying is the same with one of the predefined phrases from the data base of the search engine or even with a "stop word" from the internal list of the engine.

11.7.8. *Stop Words And Phrases ("Statements") For Searching*

The stop words are those words that most of the searching engines ignore when they are within texts or titles. For shortening the time, the searching engines ignore those short usual words (as adverbs, conjunctions, prepositions or expressions like "to be")

The list of such words is not the same for all searching engines, and such a list can be modified relatively frequently.

Such words are ignored even if they are inserted in quoted sentences excepting the Google engine, which is able to recognise them even and does not ignore them.

When defining the subject to be searched, it is recommended to take into account the following tips:

- Be as specific as possible
- If possible, use objects' names

- Enter the most important terms at the beginning of the list of the key-words
- Use at most three key-words
- Avoid using wide spread words unless they are part of a proposition
- Try and use the words that are likely to be in the web page content

11.7.9. *Other Advises*

Use lowercase. Using uppercase will lead the search to look only for these, while using lowercase will generate answers containing uppercase.

EXAMPLE: *president* retrieves both *president* and *President*

Use separation and “wildcard” signs as * (asterisk) for searching for different ways of typing what you are searching for.

EXAMPLE: *librar** returns *library*, *libraries*, *librarian*, etc.

Combine sentences and key-words by using quotas and signs as + (plus) and – (minus)

When searching for a key-word within a document, use the browser’s command “find”.

11.7.10. *What does “Boolean” means?*

Boolean logic was named after the British mathematician George Boole (1815 – 1864), who conceived a logical system designed for obtaining in case of searching the best results by asking more accurate questions. Boole named his system “calculus of thought”.

- Boolean operator “AND” allows searching only those documents that contain all the terms the user specified. Specifying more terms leads to a better search.
- Boolean operator “OR” allows searching the documents in which at least one of the specified terms is contained being usually used for similar key-words, like the synonyms. In this case, the more terms the user specifies, the more documents the searching engine will find.
- The operators “NOT” or “AND NOT” limit the search by selecting the documents containing the first key-word and not containing the second one. The documents containing the second key-word are not taken into account even if they contain the first one.

11.7.11. *Tips For Using The Search Engines*

Using the brakes is a very efficient way of combining a higher number of logical sentences for searching, and they are recommended for separating key-words when more than three or when using more than one logical operator. For more a more accurate search, it is recommended that the “statements” be between brackets. Using the Boolean logic is not a simple thing to do. Different search engines interpret the Boolean operator in different ways. For instance, only some of the searching engines accepts the operator NOT, some of them accepts the combination ANDNOT (as a single word), and other need separate words (AND NOT). Some searching engines request writing the operators with capitals. While others does not make any difference in this respect.

11.7.12. *Other Tips*

In case the subject being searched for includes more than one key-word not being separated by any symbol, a searching engine will automatically enter between them an AND or an OR. Many searching engine, including Google and AllTheWeb implicitly introduce between words the function AND if the sentence is not between brackets. Such change can significantly modify the result of the search, thus explaining the unexpected results. Knowing the implicit way a searching engine works under such circumstances is critical.

For some searching engines you can use menus allowing conceiving short sentences as a search with logical operators. Sentences like “All of the words” and “Must contain” have the same meaning as AND operator, sentences like “Any of the

words” or “Should contain” have the same meaning as OR operator, while sentence “Must not contain” has the meaning of the NOT operator.

11.7.13. *Implicit Boolean Operators*

Working with implicit logical operators is based on using signs as plus (+) and minus (-) in front of the words (without any space between the sign and the word) instead of the logical operators AND and NOT.

Including a phrase composed of key-words between quotations will lead to search them as a sentence in the exact order they were specified. Do not put between quotations when we type a single word!

While the normal form of the logical operators is accepted only in the advanced options of the searching engines, the implicit operators are accepted in the basic versions of most searching engines.

11.7.14. *Proximity Operators*

Proximity (positional) operators as NEAR, ADJ, SAME, FBY and so on are not included in the Boolean logic, but they can carry out similar functions when used for searching.

Not all the searching engines accept proximity operators, but some of them accept the operator NEAR in the advanced searching options.

- NEAR operator allows searching some terms situated at a significant distance one from the other, in a specific order. The closer the terms are, the better their position is in the list of the results. Using it instead of the operator AND leads to finding some more relevant results.
- Proximity operator ADJ (adjacent) is accepted only by few searching engine. ADJ means that the two key-words should be one near the other within a document, regardless to their order.
- Proximity operators SAME (key/words found in the same field) and FBY (followed by) are used in case of advanced searching methods in libraries or data bases with bibliographical references, not yet known by the searching engine.

11.7.15. *Searching Within Titles*

Searching by fields means using records in electronic format organised by distinct fields. For instance, a typical Web page is made of title fields, domain, site, URL and links. Some searching methods allow searching key-words only within certain fields of the Web pages, such facility being significantly efficient due to the fact that it allows to accurately specifying the area on the page to be searched.

Searching within titles (the texts in the upper area of blue coloured window of the Web browser) is useful when the user knows the subject of the searched page, the terms describing the subject being likely to be included in the title of the Web page. Searching a key/word within the titles generated more relevant results than searching within the content of the pages.

Searching within the titles does not come up with satisfying results if it used only one key-word.

11.7.16. *Searching Within A Field*

Searching within a field is used for finding information that is contained in a specific category of Web sites. In this case, the searching can be limited to one of the high level domain:

- **edu** – educational site
- **com** – commercial, business site
- **gov** – governmental site in USA
- **mil** – military site or a site of an agency in USA
- **net** – networks, providers, Internet providers (internet service providers), organisations

- **org** –non-profit organisations
will limit the search to educational sites containing information about Charles Darwin and his evolution theory.

Searching within a domain is allowed by several searching engines in the advanced searching options by choosing from the menu a domain to be searched in. The SearchEdu engine limits the search, in its basic version, exclusively to edu. domain. In November 2000, Internet Corporation for Assigned Names and Numbers (ICANN) approved seven additional suffixes, of which some are already operable:

- **.aero** – for aeronautic industry exclusively
- **.biz** – business sites
- **.coop** – co-operatives only
- **.info** – general information
- **.museum** – museums only
- **.name** – personal sites
- **.pro** – authorised professionals and professional entities only

11.7.17. *Searching For Other Fields*

Searching within a certain international domain can be effected by specifying the two letter code related to each country

Will limit the search only to the search in Great Britain containing information about the football team the user is looking for.

Due to the fact that the Internet was created in USA, the pair of letters **US** was not assigned as a domain code for the United States of America for the American sites, but it is usually used for governmental or local organisations' sites or even for the most public schools or community colleges.

Searching within a site, proper for the case when the user wants to search in a computer or server, can be affected by an instruction of the following type:

This fact will limit the search to the web page assigned to the Leonardo Da Vinci T&T Project.

Searching by URL is used when the user wants to find a folder the name of which is part of the URL of the host Web page.

Searching within links is used when the user wants, for a certain Web address to get a list of the pages containing links to the specified address.

11.7.18. *Other Searches With Specified Field*

Searching for a certain image is possible by specifying a key-word contained in the name of the folder containing the image, as well as the extension describing the folder's type.

Among other types a search can be done we can mention anchor, applet, object, text, language, sound, pictures and date.

The field containing the data is difficult to access because, according to the searching engine being used, the user can know the date the Web page was last time created, the date the Web page was modified or the date the page was accessed by the searching engine "robots".

The names of the fields should be fully typed, while the others' name should not. In every case, the name of the name of the field must be followed by ":" and then by the key/word, with no space between them.

11.8. **Evaluating Web Pages**

This chapter provides a number of useful information for evaluating the accuracy and the usefulness of a Web page information.

This chapter explains the reason of evaluating the Web pages and describes the main categories of information to be identified and analysed in order to decide whether the Web page contains accurate actual and useful information

11.8.1. *Why Should The Web Pages Be Evaluated*

The Web represents an enormous source of information for research regardless to the domain. To introduce documents into the Web is simple and facile (or even for free), but it is not regulated or monitored.

The biggest advantage the Internet brought into our life is the possibility of expressing our opinions, exchanging thoughts, and establishing new relationships with people we wouldn't have met otherwise. Through Web pages' links everybody can have access to other's ideas and personality. There is a real treasure out there in the Web, but there are also lots of dangers.

This is the reason we have to thoroughly evaluate the information we find on the Web. The duty is all the user's. When accessing the information, one should establish its validity, actuality and integrity.

11.8.2. *Supplementary Arguments For Evaluation*

The documents can be easily copied and falsified or copied with errors and omissions in them - intentionally or accidentally. There are no editors to check the information on the Web, to reject a document not meeting the standards set by a publishing house (compared to the published information). Most of the pages discovered by the searching engines are created by individuals or by companies promoting their product or their opinion. Even the universities' sites or the libraries' ones can contain pages the institution does not supervise.

The Web must provide such freedom, but a user dedicated to serious research should be sceptical and critically evaluate the information he/she accessed.

For a successful evaluation of the Web pages the user should take the following two simultaneous measures:

- The user must be well trained for a quick identification of the information allowing him/her to evaluate the Web page;
- The user need to have a critical thinking, even a suspicious way of approach looking for answers to the questions leading to the decision whether the Web page is trustful.

11.8.3. *Verifying The Web Source And The Address*

We can expect to find sites containing commercial or personal trustful information as articles and scientific reports, complex documents, academic courses, but also stupid sites containing hoaxes or scandalous content. How can we sort all these?

One of the user's most important ability is to be able to understand a Web address, named also URL (Universal Resource Locator).

Is this a personal page?

Read very carefully its URL. Many personal pages contain in their URL a name preceded by a "~" symbol (tilde), "%" (percentage) or words like "users", "members", "people" and so on. The servers of the commercial companies (aol.com, geocities.com and so on) providing access to the Internet (ISP = Internet Service Provider) host lots of personal pages.

The personal pages should not be avoided necessarily, but the author of such pages should be first carefully "investigated". There are cases when the page has no editor or owner of the domain to guarantee the quality of the requested information.

11.8.4. *Identify The Domain and The Name of The Entity Who Published the Page*

The governmental pages can be identified by the domain name .gov, .mil or .us, the educational ones can be identified by the domain name .edu, and those non-profit, the name .org.

If the domain belongs to a country, search within the page the information regarding the domain.

Search and check if the nature of the information contained in a page correspond to the domain it belongs to.

Usually, the body that published a page is an entity or a person operating the server the page is stored on. The server's name is usually included in the URL between http:// sequence and the first "/" (slash).

Is the name of that entity known to you? Is this name in accordance with the site's name?

It is recommended that you search for information provided by the most authorised source. The news in a newspaper, for instance, try to search on the newspaper site directly, and the medical information from a medical site.

11.8.5. Who And When Was Written This Page?

Find out the name of the author, of the organisation, institution or agency responsible for the page content. In many cases, just an e-mail address is not enough.

If there are no indications about the page's author, try to cut the URL to find other page which contains information for which the author takes responsibility, or at least a page motivating the absence of the author.

All Web pages are created for a purpose, they do not appear instantaneously or without a reason. If the only information about the author is an e-mail address, send him a message politely asking for more detailed additional information.

It is important that in case of very dynamic domain or the pages with a very important content to check whether the information is not out of date. It is not recommended to use, for instance, statistical information if they are not dated. They are as useless as the anonymous ones.

In many cases, the date the page was created or updated can provide you information about the interest the author might still have in the pages. Many pages are abandoned by their authors.

11.8.6. Is The Author Competent For The Subject?

Is the author experienced or educated enough for publishing the subject he/she approached? Check out whether the chosen domain is nothing but a hobby for the author, or just claiming he/she is an expert or just an enthusiastic.

Try to see to what extent the page represent just the author's opinion and compare with other opinions expressed by other authors, too. Try to identify the pages containing extreme, distorted or exaggerated opinions.

If it is impossible to identify clear, relevant information about the author's competence, read carefully the bibliography that the author offered.

Anyone can enter information in a Web page in just few minutes. You have to distinguish the information that is trustful from the questionable information.

Many Web pages represent opinions. Try to judge a Web page in terms of competence in that particular domain and documentation volume according to the same criteria and at the same critical level you judge published information (book, magazine, newspaper).

11.8.7. Check The Bibliography Sources

As any other academic information, the information contained in a Web page should include bibliography sources list. Check if the links in a Web page connect you to a trustful source and if this links really work.

Publishing information in a domain without mentioning the bibliographical resources is equivalent to expressing an opinion or a point of view. Try to establish a "trustfulness level" necessary in you search.

Journalism, the consecrated one, can represent an exception, but this kind of information is not of a scientific nature.

The links that do not work to the documentation resources reveal the low quality of the resources in many cases, thus diminishing the trustfulness of the information contained in the page.

11.8.8. *Is the information complete, unaltered, and true?*

If the information was rewritten (by typing), it can be easily altered. The existence of the permission to reproduce information and the existence of mentions regarding royalties should be checked first. Which is the reason it was reproduced that information instead of making possible the direct access to it by a link on the page? The ideal method to check the correctness of information is that of finding the initial resource of it. If the links on the page do not lead to the original resource, it is very likely that the information is altered and illegally provided. A document in a recognised newspaper can arouse suspicions if it is not accompanied by a mention about the right to publish or the royalties.

11.8.9. *Are There Any Other Links To Other Information Resources On The Same Topic?*

Are the links well chosen, well organised and evaluated or commented? Are the links to other resources working? To what extent the information is just a point of view or an opinion? Is the author subjective toward the links provided in choosing the links to be provided?

Many Web pages well articulated provide links to other pages inviting the user to compare the information. Try to find out whether the links between pages lead to different opinions. It is recommended to search for information regarding the subjectivity especially when you are tempted to have the same point of view.

11.8.10. *What Do The Other Say?*

How many links are contained in a page? Can the user define a predominant category of sites containing these links? Are there links to this page in a subject director?

In many cases there are links to a page only from other areas of the same site, which does not represent a good reference for the page. If there are links providing point of views pro and against the subject, try and acknowledge yourself with both opinions.

Subjects directories refer only to a small fraction of the whole Web, thus if the page is mentioned is a remarkable thing. Read the comments within the directory to make sure they are positive.

Searching for opinions on the author using a searching engine (Googling someone) can be relevant. Take also into account the nature of the comments resource. If the point of view expressed by the author is radical or under dispute, we can expect to find lots of opponents. Try to judge correctly all the comments and all points of view.

11.8.11. *Why Was This Page Inserted In The Web?*

Among the reasons to insert a page in the Web we can mention: to inform, to provide with facts and information, explanations, persuasion, selling products, disclosure and so on.

The Web is a public place, opened for everyone. There are a lot of possibilities and human intentions that can be hidden behind a Web page.

Is the approach on a page ironic, representing a satire or a parody?

Analyse the "tone" used in a page. Is this the author overreacting? Do the arguments seem exaggerated? Are there any scandalous pictures or are the images improbable? Are the points of view proved based on examples that are eventually impossible?

Is the page, as an information source, as useful as searching in a library or an index?

11.9. Providing The Sources In The Internet In Electronic Format

*This chapter provides useful information about the method proper for providing the sources of the accessed information through Web.
The following pages contain the elements such reference consists of, a classification of the main methods the sources can be mentioned and a description of the basic formats for mentioning the bibliographical sources.*

11.9.1. *Elements Of Providing Sources*

Primary elements of a bibliographic reference are the same for most of the documentation types, although their presentation order can vary. Such elements include the name of the author, the title, place of the publishing, name of the editor, date of publishing and a reference to the place of the information within the document (usually as a page number).

In many cases, the reference to the sources includes a description of the publishing environment.

11.9.2. *Specific Aspects Of The Electronic Sources*

For the sources of an electronic nature, some elements can miss or need to be transformed into elements in terms of our new era. For instance, instead of the name, the on-line authors use identification login type names or nicknames. Instead of the title you can enter only the name of the folder. The place of publishing and the name of the editor are replaced online by the protocol being used and the Web address, and mentioning the publishing date is replaced by the date the page was accessed.

On the Web, an address is represented by a page, regardless to its length. Thus, the layout remains only a characteristic of the publications which is less important or even senseless for electronic documents or folders.

Due to the fact that most of the Web browsers and the text editors allows the user to search a word or a sentence within a document, specifying the location of a reference in an electronic document can be sometime redundant.

When in doubt, is always recommended to provide more information instead of providing insufficient information.

11.9.3. *Mentioning The Sources Within A Text*

The references included in a text or between brackets to the printed publications usually include the name of the author and the page number ("humanist" method) or the name of the author, publishing date, and page number ("scientific" method). Often, the references between brackets for electronic sources include only the name of its author or in case it is not available, the name of the folder. For the folders, which do not contain the name of the author, individual or organisation, the name of the folder is mentioned between brackets. (for example, l60.html).

For the "scientific" method, the page contains the publishing date or the date of the last updating or, if not available, the date the page was last accessed. The date is in the day -month-year format (for instance, 27 March 1993).

In case of printed sources, the consecutive references of the same document will not repeat the name of the author, but there will be mentioned each time the page number or the location of the information in the document. The name of the author can be repeated as a unique possible solution in case of electronic documents, such documents being lack of a set layout or other ways of delimitation.

11.9.4. *Basic Formats for References To The Bibliographical Sources*

- "humanist" style
surname of the author, name of the author.
"Document's Title." *Complete Work Title* [if available].

The number of the folder version [if available]. The date of the document or the date it was last revised [in case they are different from the date it was accessed]. The Protocol and the address, the path and the directories (accessing date).

- "scientific" style

name of the author, the initials.

(Document's date [if it is different from the accessing date]).

Document's title. *Complete Work Title* [if available].

The number of the folder version [if available]. (edition or version [if available]).

Protocol and the address, the path and the directories (accessing date).

Some text processors automatically format the internet addresses by changing the colour of the fonts and underlying them. In such cases, use the implicit settings of the editor.