Internet search

In this module we will introduce the issue of Internet search tools. We will briefly present the history of the Internet. We will explain what the WWW service is and present search services on the Internet, above all search engines, subject catalogues and metasearch engines.

Module objective:

- to summarize the history of the Internet
- to outline the characteristics of the Internet
- to introduce the WWW service
- to touch upon search services (search engines, subject catalogues, metasearch engines)

Basic terms:

Internet - a worldwide computer network based on TCP/IP protocols, which enable communication between public and private networks, on various types of communication media and various technical platforms.

Search engine - one of the basic types of Internet search tools; it is a system which, based on a keyword formulated by the user, searches in a database or an index, and communicates the search results to the user.

Subject catalogues - they enable access to a large amount of information sources arranged according to topics, based on a certain hierarchic pattern.

Metasearch engine - a kind of Internet search service which enables the user to conduct a parallel search through the databases of several search services based on one search query.

Hypertext - a text containing links to other documents which may be shown by means of selecting a given link.

Contents

1 INTRODUCTION

Each of you certainly uses the Internet. You can use it for your work, studies or entertainment. The Internet has become an inherent part of our lives. Whether you want to work with the Internet or you have to, it is necessary for you to be well acquainted with all the information it offers. And the amount of this information is great. To be able to effectively search for such information so that it is relevant requires an advanced knowledge of the Internet.

2 WORLD WIDE WEB

At the beginning we must remind you that WWW is not a synonym of the Internet itself. However, due to WWW the Internet has been popularized and opened to all kinds of users.

WWW is a service which originated as a means of sharing textual information based on the hypertext principle. It thus enables access to hypertext documents in the Internet environment, and is based on the **client/server** architecture.

The **client/server** model is a certain form of distributed processing when one program (client) **communicates** with another program (server) with the purpose of **exchanging information**.

This communication usually takes place in the following steps:

- The user launches the client in order to create a request.
- The client contacts the server.
- The client sends the request to the server.
- The server analyses the request.
- The server processes the request.
- The server sends the results to the client.
- The client presents the results to the user.
- The cycle is repeated as long as it is needed.
- The connection is closed.

2.1 WWW BUILDING BLOCKS

Let us now have a look at the building blocks of the World Wide Web. They are, above all, hypertext and HTML language.

Hypertext - it is an implementation of a hypertext access to information. Originally linearly written texts are divided into smaller units called **nodes** (e.g. documents, websites), and these nodes are then interconnected via **ties** (links).

HTML language (Hypertext Markup Language) - it is intended for the creation of documents which are to be presented on the Internet. This language offers the means of describing the document content. It is a formatting language.

2.2 WEB 1.0, 2.0, 3.0

Since 2004 there has been a new Internet phenomenon, Web 2.0. The main change lies in the creation and sharing of content. Have a look at the basic differences between Web 1.0 and Web 2.0 yourselves:

Area	Web 1.0	Web 2.0
Functionality	Information is organized in hypertext websites. Control via links causes fragmentation of the information displayed and a lower degree of transparency.	In addition to information, application functionality is provided. For example: webmail, discussion, blog, notice board, photo gallery, video shots. Control enables displaying information at one place without fragmentation.
Web content	Web content is created by the owner.	Visitors take part in the content creation. The owner, in addition to creating the content, plays the role of a moderator. Decentralization of authorities takes place.
Interaction	The owner has to interact; therefore, interaction is limited.	Interaction is welcome and usually has the form of assessment, discussion, chat or social network.
Content update	The content is updated and created by the owner.	There may be a great number of content creators. The content is created by the visitors, as well as by the owner.
Social ranking of the visitor	The visitor is a passive reader of information.	The visitors form a large community. Visitors with similar interests meet at one place and can actively communicate.

The table source: http://www.indextrade.cz/clanky/web-20-jako-dulezity-milnik-v-evoluci-webu

Even though there is no such thing as **Web 2.0** according to many experts, Jeffrey Zeldman used the term **Web 3.0** in 2006. In his concept Web 3.0 is a combination of Web 2.0 and semantic web.[1]

Among the technologies the use of which is considered within Web 3.0 are:

- artificial intelligence,
- automatic derivation,
- cognitive architecture,
- knowledge representation,
- ontology in terms of computer sciences,
- semantic web,
- semantic wiki,
- software agents.

3 INFORMATION SEARCH ON THE INTERNET

The Internet is a vast information space in which there are new online information sources available every day. The half-life of a website is around two years.

In order not to get lost in such a large amount of information and to make our way to it, we must be able to search for it.

1 semantics = the study of word meaning

With regard to the availability of information sources, we have to distinguish between:

- **Information sources available directly** (they are usually public and free of charge).
- **Information sources available indirectly** (e.g. professional and commercial database centres, while the Internet is only a method of access in such a case).

3.1 SEARCH SERVICES ON THE INTERNET

In functional terms, search services are composed of two main types:

- **Search engines** - they index words or terms found on WWW pages or in electronic documents available online.
- **Subject catalogues** - they classify documents or entire WWW pages according to a given subject classification or taxonomy.

3.2 SEARCH ENGINES

We may define a **search engine** as a system which, based on a **keyword** formulated by the user, **searches** in a database or an index, and communicates the search results to the user.

An important role while searching is played by the **keyword** with which the user tries to express their information need. In the field of search engines, the keyword is used in terms of a **search term** or **search expression**.

A **query** is then composed of one or more keywords/search expressions; this query represents a complete search request. **Operators** are often applied in queries; you became acquainted with them in the module How to search for information. A document corresponding with the query is called a **hit**. A successful query leads to one or more hits.

3.2.1 USE OF SEARCH ENGINES

We may use search engine services if:

- the subject of our interest is too narrow or contains unusual terms,
- we search for a special domain (website),
- we want to search through millions of websites,
- we want to find a great number of websites for specific research purposes,
- we want to find certain types of documents, files, languages etc.,
- we want to set the filter to the date of modification,
- we want to use benefits such as term clustering, ranking documents according to popularity etc.

3.2.2 EXAMPLES OF SEARCH ENGINES

Have a look at foreign search engines.

http://www.google.com,

http://www.av.com/,

http://www.bing.com.

Among Czech search engines are for example:

http://www.jyxo.cz,

http://www.google.cz,

http://www.morfeo.cz.

3.2.3 DISADVANTAGES OF SEARCH ENGINES

One of the main disadvantages is the fact that while using more search engines we might come across the issue of different syntax of commands, which is the result of development and a lack of search engine standardization. M**etasearch engines** eliminate the disadvantage of using different commands.

Search engine databases are created automatically by means of robots, as a result of which there are many websites **of dubious quality**.

3.3 SUBJECT CATALOGUES

Subject catalogues or directories present a second search option.

This search service covers a **smaller part** of the web space than search engines. Their advantage is a strict **hierarchic classification** and division into **categories** and **subcategories**. Subject search is used if we know the subject of our interest and want to find more sources concerned with this subject.

The user using a subject directory therefore knows what they are searching for and also knows to what subject category the given information belongs.

A subject directory may be characterized as a service which offers a connection to the Internet sources provided either by website creators or information workers. A subject directory is organized into subject categories, subcategories etc. based on its type and size.

3.3.1 USE OF SUBJECT CATALOGUES

We may use a subject catalogue

- if our topic is very broad,
- if we want to get a list of web domains,
- if we search for information on a web domain,
- if we search for products or news,
- if we search according to a web domain title,
- if we want to avoid documents with a low degree of content frequently returned by search engines etc.

3.3.2 EXAMPLES OF SUBJECT CATALOGUES

Among worldwide subject catalogues is:

http://www.yahoo.com (Yet Another Hierarchical Officious Oracle),
http://www.nbci.com,
http://www.dmoz.org (Open Directory Project).

From Czech catalogues we may mention for example:

http://www.seznam.cz,
http:/www.centrum.cz,
http://www.atlas.cz
http://www.zlatestranky.cz.

3.3.3 ADVANTAGES OF SUBJECT CATALOGUES

One of the main **advantages** of subject catalogues is the **quality guarantee** which is in fact given by the way catalogues are managed (they are edited only manually). Most directories furthermore assess and annotate information sources.

3.3.4 DISADVANTAGES OF SUBJECT CATALOGUES

Among the disadvantages of subject catalogues are above all: a **limited scope**, given by the way the catalogues are managed, **category structure**, since in every subject catalogue different classification patterns are used, therefore the orientation is more difficult, **time expenditure**, if the users do not perceive the arrangement as natural, they might spend a long time browsing through the catalogue before they find a relevant subcategory, **link validity**, many documents are in time moved or even deleted, the issue of less frequent **updates**, **subjectivity** while assessing and classifying information sources due to the human factor, **general descriptions**, descriptions are created based on a mere general examination of the information sources, they do not necessarily depict the exact contents of the sources.

While categorizing search services, we mentioned that current search services cannot be strictly divided into "search engines" and "subject catalogues". Now we are returning to the problem. Nowadays we may speak of **hybrid engines** which are based on the fact that their directories have search functions, and search engines also contain directories. The difference lies in what was the primary function of the given search service, whether a directory or a search engine.

Points to think about!

Do you use subject catalogues? What do you search in them most often?

3.4 METASEARCH ENGINES

Metasearch engines enable us to **search** in various search engines or directories **at the same time.** While searching, search results are combined and duplicate entries are removed. They may also have the form of a list of search engines which may be entered from one domain (all-in-one).

3.4.1 EXAMPLES OF METASEARCH ENGINES

http://www.ask.com,
http://www.dogpile.com,
http://www.metacrawler.com,
http://www.profusion.com,
http://www.search.com,
http://www.kartoo.com .

3.4.1 ADVANTAGES OF METASEARCH ENGINES

Among the highlights of metasearch engines are:

- While searching, the user enters only **one website.**
- For an access to **more** search systems the user needs to learn to work with only **one interface** (which saves a lot of time, otherwise the user would have to get acquainted with all systems separately, since each of them has its own way of query entering, i.e. its own syntactic rules for the query formulation and its own form arrangement, they also have different output formats and structures of presented data on found documents).

- It is more effective and convenient to enter only **one** query rather than send queries to each search system separately; this saves time which would otherwise be lost e.g. while waiting due to a search engine overload.

3.4.2 DISADVANTAGES OF METASEARCH ENGINES

One of the disadvantages of this type of websites is that we do not come to know all the options the individual systems use, especially while using an extended search. The search request is then usually formulated by means of keywords, which is insufficient with an enormous number of websites.

3.5 PORTALS

Many search engines have lately been transformed into web **portals**. The purpose of portals is above all to **integrate services** in **only one user interface**. A portal may be described as a frequently visited server which offers complex services to its users.

Portals offer the following **services**:

- search for catalogue information,
- search for full-text information,
- news service,
- a possibility of personalizing the page,
- online entertainment,
- discussion forums,
- online sale (e-shops),
- free electronic mailbox,
- free web space,
- planning calendar etc.

Seznam.cz is such a portal providing most of the above-mentioned services.

4 RESULTS RELEVANCE

While searching for information, relevance means the significance or adequacy of found (selected) documents with regard to the entered query or search request. It may be measured as follows: the individual found documents are assigned a partial relevance which may be statistically evaluated, compared etc. Computer search systems often measure relevance as the rate of agreement between the entered key and the reference found.

The highest degree of relevance is usually found with documents containing the highest number of DIFFERENT keywords forming a query, and further documents in which such words occur with the highest density.

There are 3 types of relevance:

a) formal,

b) material,

c) pertinence.

Example Imagine you want to go swimming to Kraví hora, only you do not know what the opening hours and prices are. Try to enter "**kraví hora swimming pool**" in a search engine (without inverted commas, without Boolean operators). Google will provide you with a correct answer on the first position, Seznam will only give you the answer on the second page of results. In this simple example you can see how important relevance is and how it differs with each search engine. Therefore, if you are unable to find information immediately, try **another search engine** or **reformulate your query**. Most search engines contain their own search manuals. In these manuals you can often find interesting search tips.

5 HOW TO SEARCH ON THE INTERNET

Google has become a search engine which is used by more than a third of Czech users (it is 75% in Slovakia and more than 90% in Poland). That is why we will mention the possibilities of an advanced search (not only) in Google, which may really facilitate our search.

Boolean operator AND

If you enter several words in Google, Google automatically inserts the operator AND between these words, therefore it searches for all the words. You may use this operator and other operators in various search engines. Remember the module How to search for information.

Stop words

Google omits the so-called **stop words**. They are often function words (prepositions, conjunctions...) or words which do not affect the search (then, how, where). Google ignores such words while searching; otherwise there would be too many resulting entries of little relevance.

Hint! If we want Google to include these words in the query, we insert plus + in front of such a word. Do not use a space between plus and the word.

Another option is to search for exact phrases. We always insert the phrase in inverted commas, e.g. "word combination". Google then searches only for the exact phrase.

The following table contains a few other operators.

Operator	Function
Link: expression	This will show links which lead to a given web **address, e.g. link:muni.cz.**
Intitle: expression	It finds a given expression in the webpage title. E.g. **intitle:bioinformatika.**
Llintitle: expression	A similar principle to intitle. It is possible to use more words in this case. E.g. **allintitle:bioinformatika muni.**

Inurl: expression	It finds an expression in the web address. E.g. **Inurl:bioniformatika.**
Allinanchor: expression	It finds websites which contain the given word in their links. E.g. **allinanchor:Masaryk.**
Site: address expression	The command finds the expression in a given domain or network. E.g. site:www.idnes.cz univerzita. Another option is to limit the search to a certain domain, e.g. de, at, co.uk. The expression is then e.g. as follows: **site:de praha.**
Filetype:	It only searches through document types selected in advance (pdf, xls, doc, txt, sof, ps, ppt, rtf, wks, wps, wbd, wri, mw, lwp, wk1, wk2, wk3, wk4, wk5, wki, wku). E.g. **filetype:pdf bionformatika.**
Info: address	It writes out information on a given address. E.g. **info:muni.cz.**
Define: expression	It writes out a term definition. E.g. **define:bioinformatika.**
„* expression“	The asterisk replaces any word.
~expression	It searches for the given term including synonyms. E.g. ~copyright also finds intellectual property.
expression AND expression	It finds websites containing both expressions.
expression1 –expression2	It finds websites containing only expression 1. Expression 2 must not be included; e.g. Škoda -car finds websites containing only the word Škoda, not the word car.

You may enter these operators directly in the search field. You may, however, use an advanced search offered by Google, and thus facilitate your work.

6 INVISIBLE WEB

All of the above-mentioned search services index the web space based on statistic websites which are interconnected via other websites; this is the so-called **Surface Web**, which, however, does not comprise the whole of the Internet. In addition to it, there is **Invisible Web** or **Deep Web**. This comprises documents which are very to find using common search engines (i.e. these documents are invisible to them).

These are above all:

- information stored in databases,
- directories,
- specialized search engines,
- the so-called solitary websites which do not refer to other websites and there are no links to them anywhere,
- dynamically generated websites (after a basic interaction with the user) - e.g. library catalogues, calculators,
- websites protected by passwords.

Specialized search services enable you to find Invisible Web:

http://aip.completeplanet.com,
http://www.turbo10.com,
http://www.scirus.com,
www.findarticles.com,
http://www.infoplease.com,
http://www.informine.ucr.edu,
http://www.scholar.google.com.

Example

Try to find as much information as possible about yourself.

Solution: The fastest way is to visit a web www.pipl.com which specializes in people search. Enter your name there and you will get a list of documents related to it. You can limit your search to a country or a city.

7 TRENDS

7.1 EXPERIMENTAL SEARCH ENGINES

WolframAlpha is a new and, according to some experts, revolutionary search engine. It is neither a full-text search engine, catalogue search engine nor metasearch engine. The system is based on a comprehensive database where WolframAlpha tries to find you an answer (for example, if you enter Emil Zátopek, you will learn what medals he got and what times he reached). WolframAlpha contains information from the field of economics, culture, astronomy, weather, mathematics and many other.

Example

Try to find information about 1st January 1993

Solution*:* enter 1st January 1993 in the search field. WolframAlpha then gives you all the information contained in its database. 1st January 1993 was a Friday, the sunrise was at 7:49 a.m. CET and the sunset at 4:06 p.m. CET etc. If we were interested in weather on that day, it would be enough to add the keyword weather and a concrete spot.

Hakia is a semantic search engine. It still works only in the beta version, though. The whole system of result displaying is based on the current use of the following criteria at the same time: plausible websites recommended by librarians, websites presenting the newest information available, websites exactly corresponding to the query.

Google Squared was launched as an experiment in 2009. It does not display results in the form of links but in the form of a table of relations. Google Squared is still being developed.

Try to enter the question "Who was Václav Havel" in each of the search engines. Compare the results.
WolframAlpha will tell us that Václav Havel was a former Czechoslovak President from the Czech Republic. Hakia will offer many relevant links, concerned not only with his presidential term but also with his writing career. Google Squared will offer no relevant answer.

7.2 KNOWLEDGE SHARING

Knowledge sharing among users is a current Internet trend. Today users are no longer mere passive receivers of information; they actively participate in knowledge creation. A very good example of creating and sharing knowledge is the Internet encyclopaedia Wikipedia. You will learn more about these services in the next module called Effective Internet services.

8 MODULE SUMMARY

In this module you became briefly acquainted with the Internet history and the differences between Web 1.0 and 2.0. The greatest part was devoted to the Internet search services: subject, full-text and metasearch engines. You also became acquainted with search possibilities. We had a look at the issue of Invisible Web and semantic web. We briefly mentioned the trend of knowledge sharing.

Points to remember:

- There are other search engines than Google.com.
- There are special search engines for the Invisible Web.
- Most search engines have integrated manuals which will help you with your search.
- You can search for other document types on the Internet than just texts.
- Today we no longer speak strictly of either search engines or subject catalogues.