Applied informatics

Advanced search, search robots

ZEMÁNEK, Z. – PLUSKAL,D. – ŠUBRT, Z.

european social fund in the czech republic
EUROPEAN UNION
MINISTRY OF EDUCATION, YOUTH AND SPORTS
OP Education for Competitiveness
UNIVERSITY OF DEFENCE
INVESTMENTS IN EDUCATION DEVELOPMENT

Advanced search, search robots

1. Fundamentals of advanced search
2. Use of search robots
3. Options for information retrieval from information databases
4. Assignments

european social fund in the czech republic
EUROPEAN UNION
MINISTRY OF EDUCATION, YOUTH AND SPORTS
OP Education for Competitiveness
UNIVERSITY OF DEFENCE
INVESTMENTS IN EDUCATION DEVELOPMENT

Aim of the lecture

1. To characterize the principle and use the advanced search
2. Clarify the importance of search robots
3. Explain the principles of information processing results of a survey of full-text resources and its practical use

european social fund in the czech republic
EUROPEAN UNION
MINISTRY OF EDUCATION, YOUTH AND SPORTS
OP Education for Competitiveness
UNIVERSITY OF DEFENCE
INVESTMENTS IN EDUCATION DEVELOPMENT

Information search

"The starting point of scientific studies must always be a careful study of the existing literature on the subject. In order not to reinvent the wheel. [1]

"Efficient information search and its application “at the right time with the right (knowledgeable and informed) people” is a just one, but a strategic aspect of succeeding in a modern environment. Ability to find relevant information belong among the competitive advantages, and no matter what field of work ". [2]

Catalogue search

- Data catalogue
- Links are sorted hierarchically
- Search engines: Google, Seznam,…

Disadvantages:

- Due to the substantial need for "manual" work is very limited in size.
- Each catalog has a differently structured group links.
- Duration (which is based on the structure - the user can spend a lot of time before finding the correct subcategory).
- Validity - invalid links - can be partially removed in an automated way.

Catalogue search

Yahoo! Directory

Arts & Humanities Photography, History, Literature...	**News & Media** Newspapers, Radio, Weather, Blogs...
Business & Economy B2B, Finance, Shopping, Jobs...	**Recreation & Sports** Sports, Travel, Autos, Outdoors...
Computer & Internet Hardware, Software, Web, Games...	**Reference** Phone Numbers, Dictionaries, Quotes...
Education Colleges, K-12, Distance Learning...	**Regional** Countries, Regions, U.S. States...
Entertainment Movies, TV Shows, Music, Humor...	**Science** Animals, Astronomy, Earth Science...
Government Elections, Military, Law, Taxes...	**Social Science** Languages, Archaeology, Psychology...
Health Disease, Drugs, Fitness, Nutrition...	**Society & Culture** Sexuality, Religion, Food & Drink...
New Additions 10/24, 10/23, 10/22, 10/21, 10/20...	**Subscribe via RSS** Arts, Music, Sports, TV, more...

Firmy.cz

Autobazary	Erotika	Letenky	Postele
Auto-moto	E-shopy	Mobily	Práce
Cestování	Finance	Nábytek	Reality
Deníky	Fitness	Nářadí	Řemeslníci
Dětské zboží	Hračky a hry	Obytné stavby	Stavebnictví
Doprava	Jazyk. školy	Okna a dveře	Stěhování
Dům a zahr.	Kuchyně	Pneumatiky	Školy
Elektro	Lázně	Počítače	Ubytování

Přidat firmu zdarma »

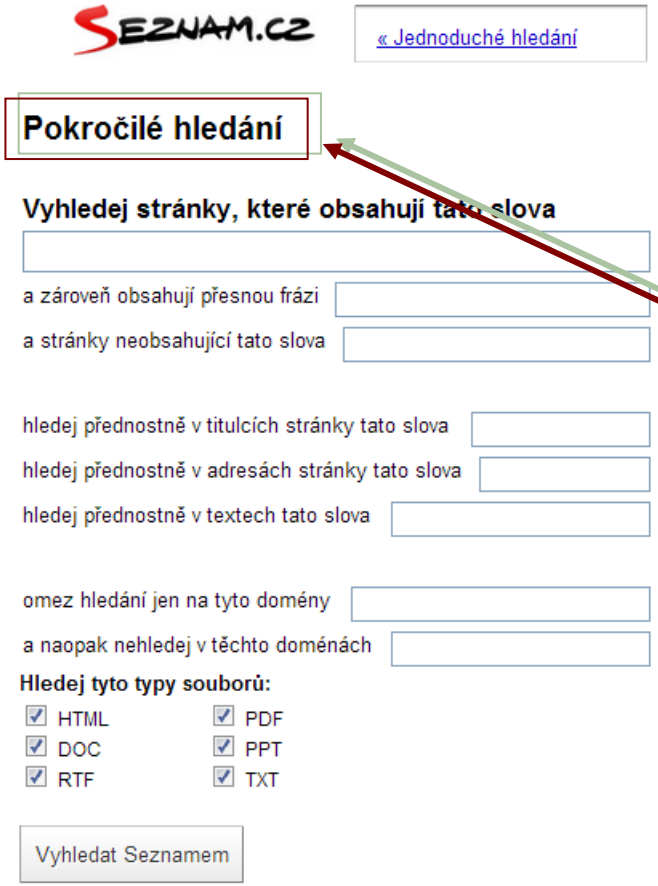
« Zpět na Seznam.cz	Encyklopedie	Kina	Obrázky	Slovník	Videa
Auto	Finance	Kurzy měn	Peněženka	SMS brána	Videoklipy
Bazar	Firmy	Lidé	Počasí	Software	Volná místa
Bulvár	Horoskopy	Lištička	Pro ženy	Sport	Výzkumník
Deníky	Hry	Mapy	Reality	Spolužáci	Zprávy
Dopravní informace	Internetové hledání	Mobilní služby	Reklamní systém Sklik	Sweb	
Dovolená	Inzerce, aukce	Moto	Seznamka	TV program	
Email	Jarmara	Nakupování	Seznam se bezpečně	Tip	

Advanced search

- The search operators are tools of the query language to exactly formulate the search query.
- Their importance can differ.
- An example (Google):
- http://www.google.com/advanced_search
- Or seznam.cz:
- http://napoveda.seznam.cz/cz/fulltext-hledani-v-internetu/pokrocile-hledani

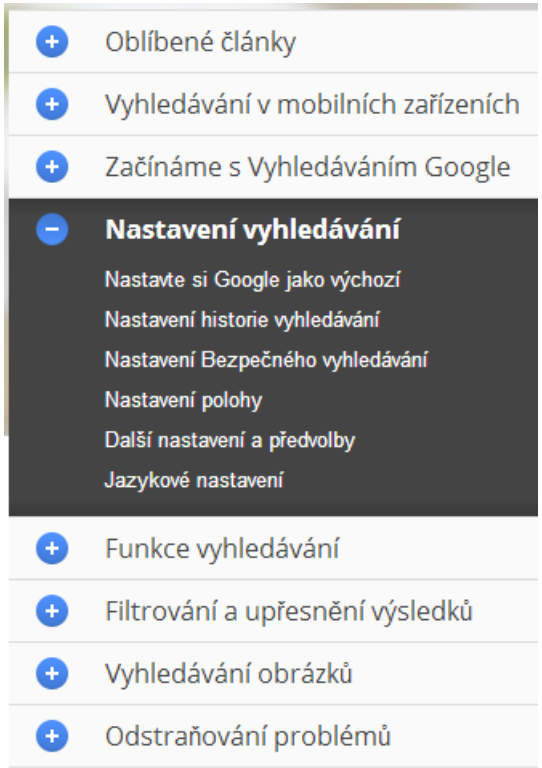
- In the advanced search are pre-set operators are:
- quotation marks ("""), comma (,), not (-), intitle, inurl, intext, site, and filetype.
- In the search list, you can also use the following operators:
- plus (+), guest-host and lang.

Fundaments of advanced search



- Allows you to type in the Internet search engines more complex queries that can be combined with supported operators.
- You can search full-text phrase, which is located in:
  - the title page,
  - in the URL,
  - in the text of the page,
  - limit the search to a specific domain or vice versa the domain of the search excluded
- can set documents to be searched.

Use of settings and filters



- Search can be simplified by using the search engine settings
- In your profile
- using the search filters (by content)

Meta search engines

= Integration of search engines in one environment

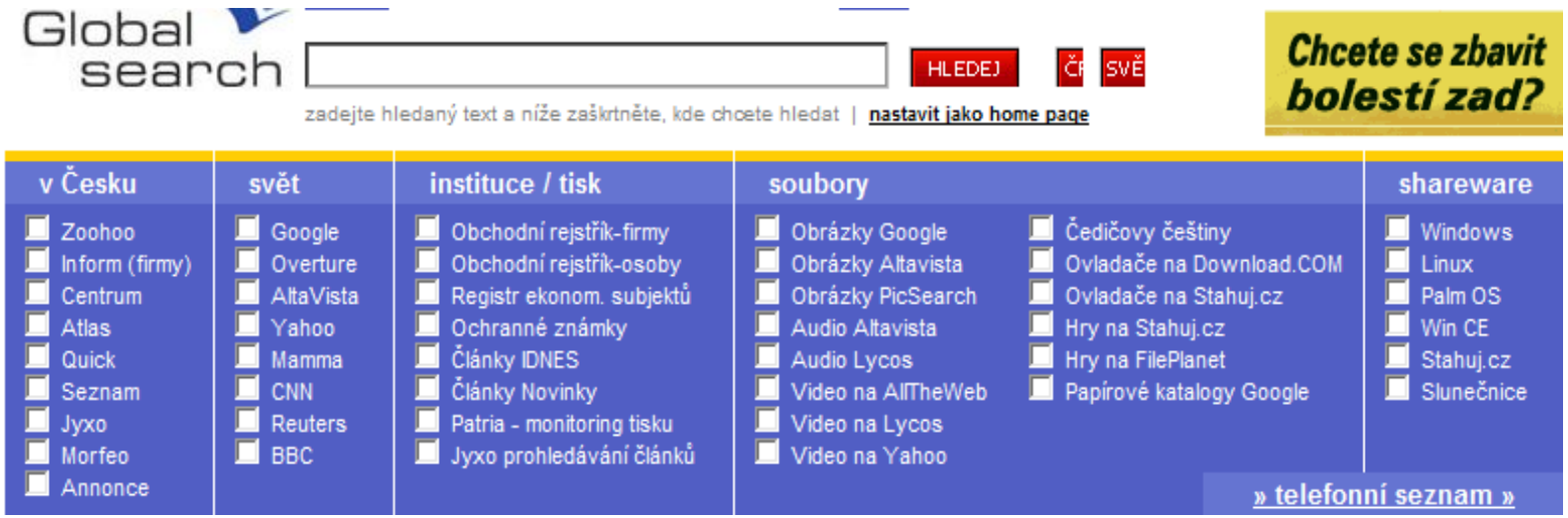
Alenka

Interesting Web portal with instant search in many Czech and foreign search engines.

Alenka is a classic meta search engine, but it allows you to pass a query to the selected search engine from a single place, without any further processing of the results.

Meta search engines

- www.alenka.cz
- www.globalsearch.cz
- www.odskok.cz/sluzby/robot.php



Meta search engines

Meta search engines do not index the Internet, but use existing search engines.

Advantages:

- ❑ removal of duplicates.

Disadvantages:

- ❑ slow – need to wait for several results of search engines

Meta search engines

1. ***The user query*** is sent to multiples independent search engines,which do their own search and then it presents the total result to the user.

2. ***The distribution mechanism*** is the basis of the meta search *(*the algorithms deciding which search engines will be asked).

3. ***Agent of the interface*** - converts a question from a metasearch form to the language that will be understandable by the specific search engine and then turn the results of each search engine to uniform shape to display them in metasearch.

4. ***The imaging mechanism*** - its mission is to eliminate multiple links (duplicates) in one document and verify their existence.

What is a search robot?

- Program (SW), which repeatedly performs routine activities on the Internet.
- It runs on the portal.
- Typically, it collects data, sends and processes requests for services.
- Examples of robots are search part of the search engines (crawlers, spiders).

Why search engines?

- Helps you navigate the vast amount of information that is on the Internet.
- Very substantial acceleration time to find answers to the question.
- This is a full-text search according to the user's query.

Robot continuously collects WWW documents and creates the database (indexing).

Robots can operate either continuously or at certain time intervals.

How the robot works?

1. Robot goes through different websites looking for links to new pages (e.g. restricted in domains).
2. Indexes the content of any pages and links.
3. Content is continuously stored in the database, or existing records are updated.
4. Allows subsequent search (request-response).

Examples of search engines

+ Robot check links (LinkChecker). It parses the page and tests the references to (non-)existent pages.

+ Robots for management and maintenance of portals.

- Comment spam. Such robots locate forms on the web, and inserts ad text or commercial communication.

- E-mail addresses harvisting. To send unsolicited mail (spam).

Visible and invisible web

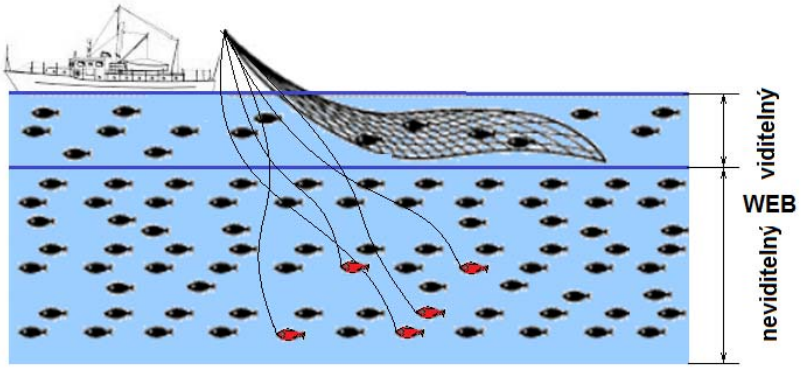
The concept of visible web (or "surface web") refers to a commonly available indexable pages.

The opposite is the invisible web (often referred to as "deep web") which contains documents that are readily searchable.

The reasons [3]:

- search engines can not index dynamically changing pages (information is generated from a database)
- Many search engines have restrictions on the number of pages indexed from a specific domain
- Most search engines prefer indexing of popular sites
- access to some sites is password protected

Invisible web



Characteristics of invisible web [3]***:*** [4]

- invisible web is up to 500 times larger than the surface web
- It is the fastest growing part of the site
- to 95% of the information in the Invisible Web is publicly available information that is available without charge

Intelligent Agents

- SW, which assists the user, navigates when working with PC applications, e.g. reading, filtering, sorting, searching, information management support:
  - artificial intelligence,
  - knowledge of user preferences,
  - principles of fuzzy logic,
  - neural Networks
  - other advanced algorithms
- incorporated into search engines (web spiders, web robots), application of competitive intelligence (incorporated in mobile technology).
  - under defined conditions (+ built-in knowledge of the user) filter and seek information on the level of the user (autonomous mode)
  - the ability to "learn" - to imitate our previous decisions in novel situations

Term: Review

From every serious professional-looking text the following three aspects should be visible:

- well-known knowledge;
- what the author came alone, ie what are his own opinions, attitudes, evaluations, measurements, etc.;
- what he had learned from others and how these ideas were taken, processed and cited. [5]

Searches are processed on requests. Their features are targeting and complexity (they include annotations, not just citations).

[5]

Term: annotation

Annotations may take the form:

- a brief summary of the document;
- notes to individual points of text, for example on the outskirts of the book (marginal);
- assessment or criticism of the document in terms of users or experts (brief review, the review notes, magazines, blogs and the like.).

Description:

- Briefly characterizes the content to facilitate user selection (e.g., annotations in the database, bibliography, in the publishing catalog);

Length: a range of 5 -10 lines ...

Term: Abstract

- Abstract is a brief summary of a scientific article, dissertation, report, or any in-depth analysis of any subject or discipline.
- It serves primarily as an aid to the reader to quickly get your bearings in the published work.
- Located at the beginning of the work.

- The length of abstract depends on the discipline, practice of the magazine or other media as well as the requirements of author.
- The typical length is between 100 and 500 words seldom more than one page.

Term: Abstract

- Academic abstract usually outlines four components essential to complete the work:

- The focus of the research (i.e., outlining the problem)
- Used research methods (experimental research, case studies, survey, etc.)
- The research results
- Overall conclusion and recommendations
- It may also contain concise reference.

Copyright law

Copyright law is the branch of law that deals with the legal relationship of the user and the creator of "works of authorship".

Copyright is protected by the Copyright Act.

Copyright does not protect ideas themselves; protected is only a concrete work, concrete expression of such ideas, works in objectively perceivable form. Copyrighted works are results of a unique creative activity of the author. A work is not a message, information, method, theory, formula, graph, table of physical constants, the output of a computer program, etc. [6]

Processing the results

The steps of the survey:

- ❑ Defining the objectives of the survey
- ❑ Choosing a database query techniques
- ❑ Selection of search terms, their combination
- ❑ Viewers of matching records
- ❑ Creation of the search catalog
- ❑ Evaluation of outcomes - relevant records
- ❑ Any change in search strategy
- ❑ Completion of the survey

Processing the results

- Finding information processed into bibliographic records with or without annotations.
- If the survey is carried out in a number of identical information sources - then compare individual records to avoid duplicates (Always give priority to the original).
- If two records of the same original source -> then choose the one that has the most complete, accurate, and the most recent data.

Processing the results

- Incomplete entries that can be complemented from other sources or records with other formal defects hindering the identification of the original source - Prefer to exclude them from the list.
- Check the content of the information according to established criteria.
- Then implement the decision, which records will be definitely included.
- It is clear that the search should not contain any information that would unreasonably be beyond the content of the assignment.

Assignments

- Explain the nature and advanced full-text search of information.
- Provide practical importance of robots in search of information.
- Process the survey results in the individual steps.
- Explain and test the possibilities of information retrieval from databases and information retrieval to the formation of case study.

european social fund in the czech republic
EUROPEAN UNION
MINISTRY OF EDUCATION, YOUTH AND SPORTS
OP Education for Competitiveness
UNIVERSITY OF DEFENCE
INVESTMENTS IN EDUCATION DEVELOPMENT

Resources:

1. ŠESTÁK, Z. *Jak psát a přednášet o vědě.* Vyd.1. Praha: Academia, 2002. 204 s. ISBN 80-200-0755-5
2. PAPÍK, R. *Vyhledávání informací I. Umění či věda?* Národní knihovna. Knihovnická revue. roč. 12, č. 1. 2001. s.18-25
3. Infogram: *Neviditelný web.* [online]. Praha: MŠMT, 2013 - [cit. 2013-11-5]. Dostupné z: http://www.infogram.cz/article.do?articleId=1765.
4. BERGMAN, M. "White Paper". *The Deep Web: Surfacing Hidden Value* [online]. Sioux Falls (SD, USA): BrightPlanet Corporation, September 24, 2001 [cit. 2013-11-8]. Dostupné z: http://www.brightplanet.com/images/stories/pdf/deepwebwhitepaper.pdf
5. KUŽELÍKOVÁ, L. - NEKUDA, J.- POLÁČEK, J. *Sociálně-ekonomické informace a práce s nimi* [online]. Brno: Masarykova univerzita, Ekonomicko-správní fakulta [cit. 2013-11-8]. Dostupné z: http://is.muni.cz/do/1456/soubory/oddeleni/svi/skripta/es2008-01.pdf.
6. Autorské právo. In: *Wikipedie: otevřená encyklopedie* [online]. San Francisco (Kalifornie): Wikimedia Foundation, 2002-2013, naposledy edit. 2013-06-4 [cit. 2013-11-7]. Česká verze. Dostupné z: http://cs.wikipedia.org/wiki/Autorské_právo.