Pronalaženje informacija na Internetu

Miroslav Milinović

miro@srce.hr

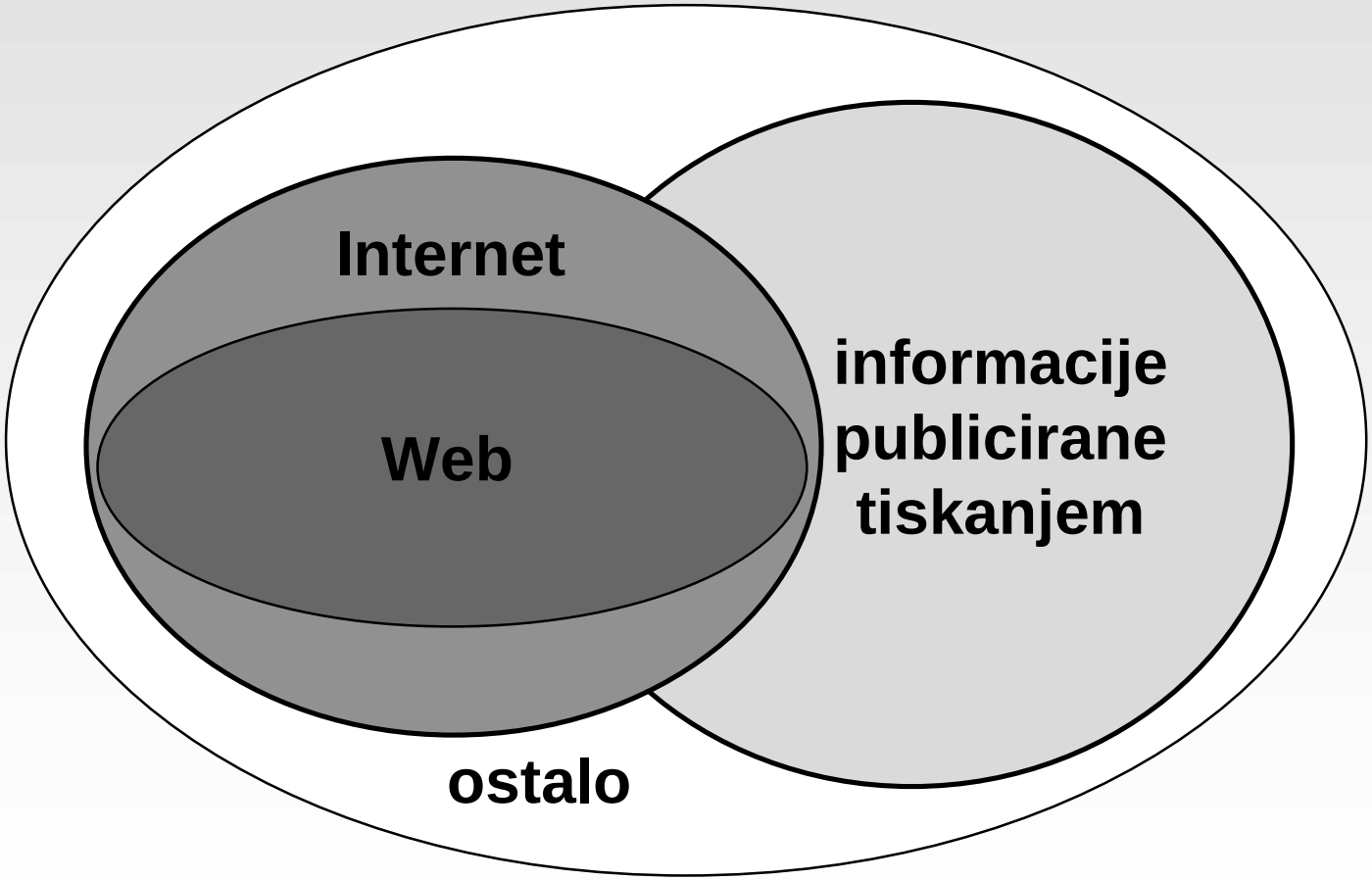
2. Stručni skup Knjižnice hrvatskih učilišta

Split, 29.-30., listopada 2001.

Sadržaj

- Internetski prostor informacija
- Mrežni izvori informacija (resursi)
- Identifikacija mrežnih resursa
- Meta podaci (metadata)
- Pretraživanje mrežnih resursa (posebno Weba)

Prostor informacija



Internetski prostor informacija

- NIJE UREĐEN - unificiran
- Postoje različiti izvori informacija (resursi)
- Mnoštvo tema
- Informacije su dostupne u različitim formatima
- Pristup je moguć pomoću različitih alata (programa)
- Postoje informacije koje (još) **nisu**:
  - publicirane u elektroničkom obliku
  - dostupne putem mreže

Internetski prostor informacija

Mnoštvo dostupnih tema i formata:

- dokumenti različitog formata
- slikovni, audio i video zapis
- elektronička izdanja novina, časopisa, knjiga, ...
- katalozi, organizirane kolekcije informacija
- baze podataka
- javno dostupna programska podrška
- **...**
- *smeće*

Mrežni izvori informacija (resursi)

- **Informacije se publiciraju pomoću različitih mrežnih usluga i servisa:**
  - Web
  - FTP arhive
  - distribucijske (mailing) liste
  - mrežne novine (USENET)
  - elektronička pošta
  - imenički servisi (LDAP, ...)
  - baze podataka dostupne putem mreže
  - **...**

Web informacijski prostor

- pretraživi (*publicly indexable*) Web
  - veljača 1999., *Lawrence and Giles, NEC Institute*
    - 800 milijuna stranica, 15 (6) TB informacija
    - sadržaj: 83% com, 6% sci/edu, 1.5% porn
    - 60% Weba je indeksirano / katalogizirano
  - siječanj 2000., *Inktomi & NEC Institute*
    - više od 1 milijarde Web stranica
    - top-level domene: 55% .com, 8% .net, 4% .org, 1% .gov

Web informacijski prostor

- 40% od 800 milijuna stranica su duplikati

  FAST, 2000.

- 30% Web stanica su kopije

  Shivakumar and Garcia-Molina, 1998.

- “Deep” Web
  - 400 do 550 puta veći od “surface” Weba
  - 7500 TB podataka

  The Deep Web: Surfacing Hidden Value; BrightPlanet.com, srpanj 2000.

Web informacijski prostor

- 85% korisnika rabi pretraživačke mahanizme ili tematske kataloge kako bi pronašli informacije

  Steve Lawrence, Lee Giles , Nec Institute, veljača 1999.

- korisnici smatraju da je Internet važan izvor informacija
  - 2/3 korisnika smatra da je Internet važan ili vrlo važan izvor informacija
  - 53%(47%) smatra TV (radio) jednako važnim

    Center for Communication Policy, UCLA, kolovoz 2000.

Problemi?

- velika očekivanja korisnika
- alati i mehanizmi
  - još uvijek nedovoljno dobri
  - u stalnom razvoju
- informacijski prostor nije (dobro) organiziran
- nepouzdana:
  - kvaliteta informacija
  - integritet informacija
  - povjerenje u izvor informacija

Znate li ...

- **tko je bila prva žena pilot u nekoj komercijalnoj avio-kompaniji? Možete li pronaći njenu sliku (traži se točna URL adresa)?**

  - **Odgovor:** Helen Richey; da (http://iswap.org/images/richey.jpg)
  - **Put:** Rabimo **Northern Light** s upitom **"first woman airline pilot"**. Jedan od prvih 10 odgovora je i link na *ISAfaqs.html* Web stranicu.
  - **URL:** http://iswap.org/ISAfaqs.html

Identifikacija mrežnih resursa

- **URI** - Uniform Resource Identifier (RFC 2396)
  - **URL -** Uniform Resource Locator (RFC 1630, RFC 1738)
    - određuje: način pristupa, adresu računala, naziv datoteke **...**
    - **protocol://host_name[:port_num][/path][/file_name]**
    - PURL - Persistent URL
  - **URN -** Uniform Resource Name (RFC 1737, RFC 2141)
- **URC** - Uniform Resource Characteristics
  - podaci o mrežnom resursu
  - metadata = podaci o podacima

Meta podaci (metadata)

- podaci o mrežnim resursima
- mogu se rabiti u različite svrhe:
  - pronalaženje informacija
  - rangiranje/vrednovanje sadržaja
  - zaštita autorskih prava
  - zaštita privatnosti
  - ...

Meta podaci (2)

- povezivanje s dokumentom:
  - uloženi (embedded) npr. HTML META tag
  - povezani s dokumentom (HTTP header)
  - dostupni preko treće strane (eksplicitni HTTP GET)
- načini zapisivanja (sintaksa):
  - HTML (META tag)
    - `<META NAME=“value” CONTENT=“value”>`
    - najčešće korištene vrijednosti NAME atributa:
      DESCRIPTION, KEYWORDS, TITLE, AUTHOR
  - XML
  - RDF (Resource Description Framework**)**

Meta podaci (3)

- aktualno stanje:
  - posebna pažnja usmjerena je na Web:
    - W3C: http://www.w3.org/Metadata/
    - Dublin Core: http://dublincore.org/
  - sustavi za pretraživanje Weba koriste meta podatke, ali ne bez poteškoća
  - nema pravog standarda, ali Dublin Core je dobar kandidat
  - rabite HTML META tag s oprezom

Meta podaci (4)

- oko 800 milijuna Web stranica
- 15 TB (6 TB) podataka
- jednostavni HTML META tag - 34%
- Dublin Core standard - 0,3 %
- 123 različita oblika META taga

Steve Lawrence, Lee Giles (Nec Institute, February 1999)

Sustavi za pretraživanje

- mnoštvo različitih sustava (alata)
- većinom su specijalizirani za pretraživanje određenih resursa
- (gotovo) svi alati imaju Web sučelje
- doseg pretraživanja je globalni ili lokalni
- nema savršenog niti sveobuhvatnog alata
- opterećeni su problemom ažurnosti i/ili kvalitete
- postoje alati koji se temelje na Webu, ali ne pretražuju Web resurse

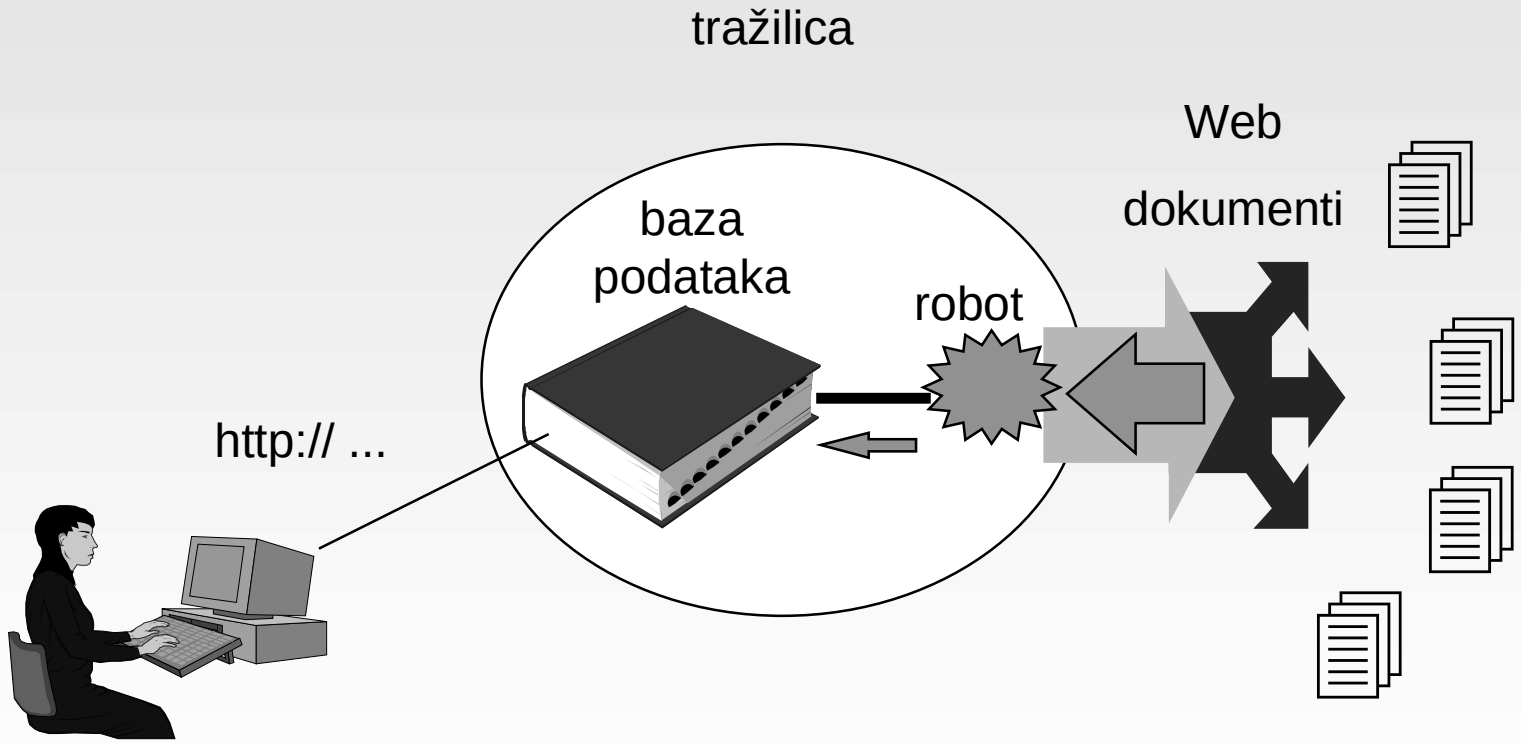
Sustavi za pretraživanje Weba

- **Tražilice (pretraživački mehanizmi)** (*search engines*)
  - tražilice *(search engines)*
  - metatražilice *(metasearch engines, unified search interfaces)*
- **Tematski katalozi** (*subject catalogs, subject directories, ...*)
  - u pravilu pretraživi (searchable indexes, searchable catalogs)
- **Ostali sustavi**:
  - višestruka sučelja *(multiple search interfaces)*
  - specijalizirana sučelja (*information gateways*)
  - **...**
- **Portali**

Tražilice
Što su i kako rade?

- automatizirani sustavi
- prikupljaju informacije o mrežnim resursima i omogućuju pretraživanje prikupljenih informacija
- posebni programi - roboti (*robot, crawler, spider*)
  - dohvaćaju dostupne mrežne resurse (Web stranice)
  - grade/održavaju pretraživu kolekciju podataka (bazu podataka)
- sustav za pretraživanje (baze podataka)
  - Web sučelje omogućuje korisniku postavljanje upita
  - posebna pravila za postavljanje upita
  - ispis rezultata pretraživanja (*hits*)

Tražilice
Što su i kako rade?



Roboti

- mogu jako opteretiti i mrežu i računalo (poslužitelj)
  - vodite brigu o robotima, ali i o tuđim resursima
- postoje pravila ponašanja (etika) za robote:
  - robot exclusion protocol
  - ROBOT META tag
- korisne URL adrese:
  - http://info.webcrawler.com/mak/projects/robots/robots.html
  - http://www.searchenginewatch.com/webmasters/spiderchart.html

Tražilice
Primjeri

GO.com (InfoSeek) - **http://www.go.com/**
Lycos Search - **http://www.lycos.com/**
Alta Vista - **http://www.altavista.com/**
excite! NetSearch - **http://www.excite.com/**
Google - **http://www.google.com/**
HotBot - **http://hotbot.lycos.com/**
WebCrawler - **http://www.webcrawler.com/**
Nothern Light Search - **http://www.northernlight.com/**
FAST - **http://www.alltheweb.com/**
Raging Search - **http://ragingsearch.altavista.com/**

tražilice lokalnog dosega

http://cross.carnet.hr/

Tražilice
Postavljanje upita

- Sintaksa upita i spektar mogućnosti ovisi o alatu
  - postoji standardni spektar mogućnosti (uporaba malih i velikih slova, fraze, kontrola ključnih riječi, …)
- Moguć je izbor resursa koje pretražujemo
  - Web ili neki drugi resursi; čitavi dokumenti ili samo naslovi, ...
- Korisno je pri prvom susretu s nekim alatom pročitati raspoložive upute

Tražilice

Mogućnosti kod postavljanja upita

- uporaba malih i velikih slova
  John December
  island
- uporaba fraza
  “John December”
  “NASA Space shuttle program”
- uporaba logičkih operatora (AND, OR, NOT)
  vegetables AND green
  fruit NOT apple
- kontrola ključnih riječi (+, -)
  +film +noir -”pinot noir”
  +python -monty

Tražilice
Mogućnosti kod postavljanja upita (2)

- susjednost - proximity search
  `Internet NEAR training`
- uporaba dijelova (korijena) riječi (truncation; stemming)
  – wildchars: *, ?, %
  `alumi*um`
  `comput*`
- kontrola resursa
  `title:"Internet training"` (AltaVista, HotBot, …)
  `host:www.fer.hr` (AltaVista)
  `cache: www.carnet.hr` (Google)
- kaskadno pretraživanje (refine); related; clustering

Tražilice

Mogućnosti kod postavljanja upita (3)

- *natural language searching* (Ask Jeeves! - **http://www.ask.com/**)
- novi pristupi:
  - Ditto.com - **http://www.ditto.com/**
  - Simpli.com - **http://www.simpli.com/**
  - Oingo - **http://www.oingo.com/**
- korisna URL adresa:
  - http://www.searchenginewatch.com/

Tražilice
Važne odlike

- Baza podataka (veličina, ažurnost, složenost) / kolovoz 2001.
    - Google - 1000 milijuna Web stranica (1300 ?)
    - INKTOMI - 500 milijuna Web stranica
    - AltaVista - 550 milijuna Web stranica
    - FAST - 625 milijuna Web stranica
- Mogućnosti postavljanja (složenih) upita
- Brzina rada (odziv)
- Rangiranje rezultata (*ranking*)
- Kvaliteta i mogućnost kontrole ispisa
- Dodatne mogućnosti

  (kaskadno pretraživanje/profinjavanje upita, ...)

Tražilice
Rangiranje rezultata

- kriteriji se temelje na:
  - frekvenciji i položaju (npr. u naslovu) ključnih riječi
  - meta podacima
  - popularnosti
  - analizi linkova (relevantnost)
- plaćeno oglašavanje vs. objektivno rangiranje

Tražilice
Prednosti i mane

- Prednosti:
  - veliki opseg
  - efikasno pretraživanje i pristup informacijama
  - automatiziran rad
- Mane:
  - nema kontrole kvalitete
  - nema klasifikacije
  - rezultati mogu biti izvan konteksta (npr. “space”)
  - sadrže i zastarjele i nepostojeće URL adrese
  - sadrže i smeće

Tražilice

Metatražilice

- ***metasearch engines, unified search interfaces***
- omogućuju korisniku da putem unificirane forme postavi jedan upit kojeg zatim distribuiraju odabranim tražilicama
- kod postavljanja upita treba koristiti samo sintaksu koju poznaje tražilica
- korisnik dobiva zbirni rezultat pretraživanja
- nemaju vlastite baze podataka niti robot program

Tražilice
Metatražilice (2)

- **primjeri:**

  All4one - http://all4one.com/

  Mamma - http://www.mamma.com/

  MetaCrawler - http://www.metacrawler.com/

  SavvySearch (CNET Search.com) - http://www.savvysearch.com/

Tražilice
Metatražilice (3)

- **važne odlike**:
  - broj i izbor povezanih tražilica
  - brzina rada (odziv)
  - rangiranje rezultata
  - način udruživanja rezultata (*results merging*)
  - kvaliteta ispisa
  - mogućnost kontrole ispisa
  - dodatne mogućnosti

Tražilice
Metatražilice (4)

- imaju sve prednosti i mane običnih tražilica
- **dodatna prednost:**
  - pojednostavljuju pristup i pretraživanje
- **dodatne mane:**
  - unificiranjem upita gube se dodatne mogućnosti postavljanja složenijih upita i kontrole ispisa
  - sporije pretraživanje

Tematski katalozi

Što su i kako rade?

- tematski organizirane kolekcije podataka o odabranim mrežnim resursima

  (odabrani resursi klasificirani po temama)
- sadrže URL adrese mrežnih resursa
- mogu sadržavati i nazive resursa, sažetke, ...
- ne održavaju se automatski (programski) već se temelje na radu urednika

Tematski katalozi
Što su i kako rade?

- klasificiranje resursa se odvija prema hijerarhijskoj shemi tema (područja)
- način klasificiranja nije unificiran
  (UDC, Dewey, proizvoljan ...)
- postoji mogućnost pretraživanja kataloga
- neki tematski katalozi povezani su s tražilicama

Tematski katalozi
Primjeri

Yahoo - **http://www.yahoo.com/**

LookSmart - **http://www.looksmart.com/**

EINet Galaxy - **http://galaxy.einet.net/**

Magellan - **http://magellan.excite.com/**

NetGuide - **http://www.netguide.com/**

About.com - **http://www.about.com/**

Open Directory - **http://dmoz.org/**

katalozi lokalnog opsega:

WWW.HR - **`http://www.hr/wwwhr/`**

Tematski katalozi
Važne odlike

- veličina (broj klasificiranih resursa)
  - Yahoo - >100 urednika, 1,8 milijuna Webova
  - Open Directory - 36000 urednika, 2,6 milijuna Webova
  - LookSmart - 200 urednika, 2,5 milijuna Webova
- tematsko stablo - način klasifikacije
- dodatne informacije o resursima
- rangiranje resursa
- mogućnost pretraživanja
- veze s tražilicama
- dodatne mogućnosti

Tematski katalozi
Prednosti i mane

- Prednosti:
  - klasifikacija resursa po temama (područjima)
  - mogućnost internog pretraživanja kataloga
  - nema “smeća”
- Mane:
  - manualno održavanje
  - pojedine dijelove kataloga ne uređuju profesionalci
  - sadrže i zastarjele informacije

Ostali sustavi

Višestruka sučelja (*multiple search interfaces*)

- jednostavna sučelja koja korisniku omogućuju da na jednom mjestu odabere tražilicu koju će rabiti
- nemaju vlastite baze podataka niti robot program
- primjeri:
  - All-in-One - **`http://www.albany.net/allinone/`**
  - Easy Searcher - **`http://www.easysearcher.com/`**

Ostali sustavi

Specijalizirana sučelja (*information gateways*)

- **prednosti:**
  - korektno klasificiran sadržaj uvijek u kontekstu
  - moguće pretraživanje
- **mane:**
  - vezani uz jednu temu (područje)
  - manualno održavanje
- **primjeri:**

  OMNI - **`http://www.omni.ac.uk/`**

  SOSIG - **`http://sosig.ac.uk/`**

Ostali sustavi

- **Imenički servisi utemeljeni na Webu**
  - White pages & Yellow pages
  - ne rabe niti LDAP niti neki drugi protokol specifičan za imeničke servise
- **Web alati za pretraživanje ne-Web resursa**
  - USENET (http://www.deja.com/usenet/)
  - FTP search (http://ftpsearch.lycos.com/)
  - distribucijske (mailing) liste (http://www.liszt.com)
  - . . .

Ostali sustavi

- **pretraživanje kolekcija (baza) podataka**

  Inivisible Web - **http://www.invisibleweb.com/**

  Lycos Seach. DB - **http://dir.lycos.com/Reference/Searchable_Databases/**

  INFOMINE - **http://infomine.ucr.edu/**

  Terraserver - **http://terraserver.com/**

- **i ...**
  - rječnici, enciklopedije, vodiči, pretražive kolekcije multimedijalnih sadržaja, ….

⟹ **PORTALI**

Portali

- ulaz u informacijski prostor Interneta
- hibridni alat - pravo rješenje
- **nude pristup** (svim) **mrežnim servisima na jednom mjestu**
- temelje se na tražilici i/ili tematskom katalogu
- **nude kvalitetne informacije**
- **nude personalizirano sučelje**
- opći ili specijalizirani (tema ili interesna skupina)
  - http://cnn.com/
  - http://www.excite.com/
  - http://www.yahoo.com/
  - http://www.ihlth.com/
  - http://www.digitalessays.com/
  - ...

Sustavi za pretraživanje Weba
Zaključak

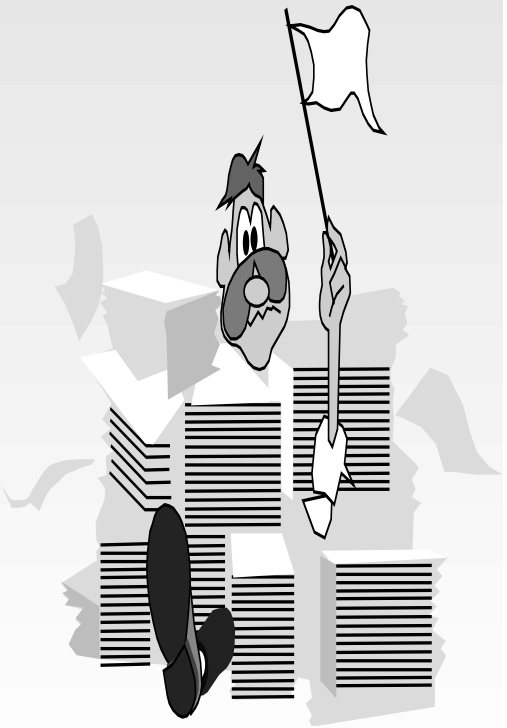
- svaka grupa alata ima svojih prednosti i mana
- orijentirani su na tekst dokumenta
  (slikovni i zvučni zapis nije moguće pretraživati po sadržaju)
- očekuje se da obuhvaćaju i ne-Web resurse
- temeljne brige:
  - kako biti ažuran
  - kako očuvati kvalitetu (precision .vs. recall)
  - kako odijeliti “mrežno smeće” od kvalitetne informacije
- budućnost je u “suradnji među alatima”
- pobjednik: **PORTAL**
- korisna adresa: http://searchenginewatch.com/

Pretraživanje Web resursa
Izbor alata

- **PORTALI !**
- **tematski katalozi**
  - kad nemamo (dobre) ključne riječi odnosno jasnu ideju što tražimo
- **tražilice**
  - kad imamo precizne ključne riječi i jasnu ideju što tražimo
- **višestruka sučelja**
  - korisna jer daju pregled raspoloživih alata
- **specijalizirana sučelja (za neko područje)**
  - nude kvalitetne informacije (ako postoje i znamo za njih)

Pretraživanje Web resursa
Kako pretraživati?

- dobar izbor ključnih riječi je presudan
- biti usmjeren k cilju (Ne lutati!)
- treba se koncentrirati na temu, a ne na postavljanje uputa
- ići k cilju postepeno (profinjavati upite)
- upoznati alat (Pročitajte HELP i FAQ!)
- biti fleksibilan i probati više različitih (tipova) alata
- graditi vlastite kolekcije zanimljivih mjesta na mreži



O čemu je bilo riječi?

- Internetski prostor informacija
- Mrežni izvori informacija (resursi)
- Identifikacija mrežnih resursa
- Meta podaci (metadata)
- Pretraživanje mrežnih resursa (posebno Weba)

“Pity the poor fanatic! When he loses sight of his objective he redoubles his efforts!” (Einar Stefferud)