Pronalaženje informacija na Internetu i metapodaci

Miroslav Milinović

miro@srce.hr

Zagreb, studeni 2003.

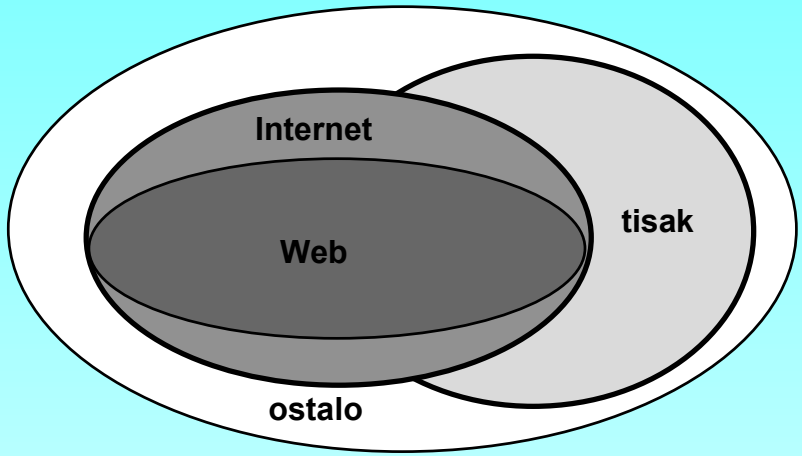
Sadržaj

- Dio 1. - Pronalaženje informacija na Internetu
  - Internetski prostor informacija
  - pretraživanje Weba
  - tražilice
  - tematski katalozi
- Dio 2. - Metapodaci
  - motivi i koncepti
  - Dublin Core
  - RDF

Dio 1. - Pronalaženje informacija na Internetu

Internetski prostor informacija

- NIJE UREĐEN - unificiran
- Mnoštvo tema
- Različiti izvori informacija
- Različiti formati



- 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
  - **...**

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

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
  - 2003. (?)
    - ≈ 5 milijardi Web stranica
  - mwp@SRCE.hr (2002.)
    - Hrvatski Web prostor (.hr TLD) ≈ 320 GB (≈ 6 milijuna resursa)

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.

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

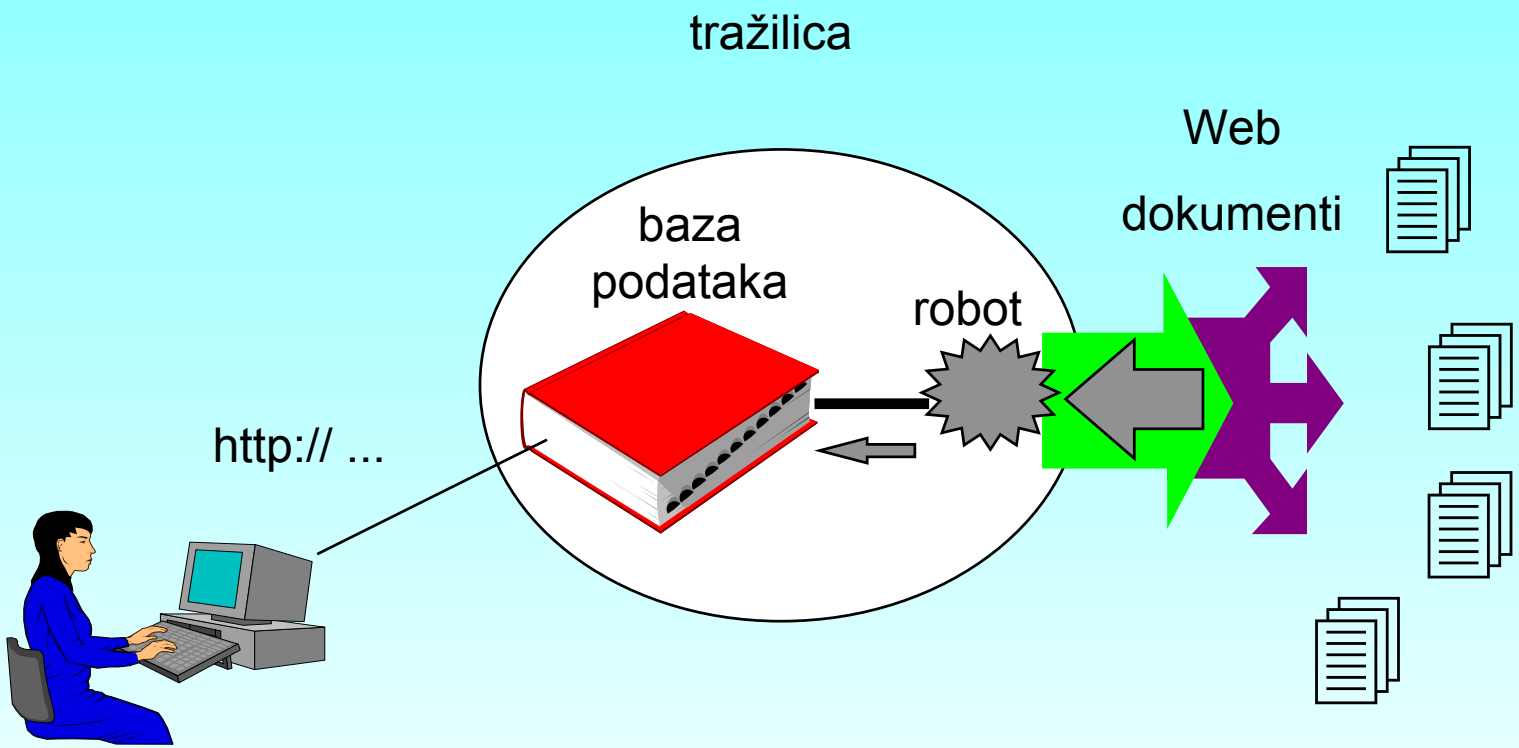
- **Tražilice (pretraživački mehanizmi)** (*search engines*)
  - tražilice *(search engines)*
  - metatražilice *(metasearch engines, unified search interfaces)*
- **Tematski katalozi** (*subject catalogs, subject directories, ...*)
  - u pravilu pretraživi (searchable indexes, searchable catalogs)
- **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)
  - sudjeluju u gradnji/održavanju pretražive kolekcije podataka (baze 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?



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/**)
(http://altavista.hinet.hr/)

Tražilice
Osnovni procesi

- **prikupljanje resursa (gathering)**
- **izgradnja baze podataka (indexing)**
- **posluživanje korisničkih upita (searching)**

Tražilice
Prikupljanje resursa

- valja razmisliti o:
  - samo HTTP ili i drugi protokoli/resursi (FTP, Usenet, …)
  - koristimo li cacheve za dohvat resursa (ICP?)
  - distribuirano prikupljanje
  - popis resursa (gdje početi i gdje završiti)
  - kojom učestalošću obilaziti (iste) resurse
  - potrebni mrežni i računalni resursi

Tražilice
Indeksiranje

- valja razmisliti o:
  - koje formate dokumenata indeksiramo (samo HTML ili …)
  - puni tekst ?
  - metapodaci ?
  - višejezičnost
  - uporaba podataka iz drugih baza (indeksa)
  - potrebni mrežni i računalni resursi

Tražilice
Pretraživanje

- valja razmisliti o:
  - sintaksa i mogućnosti pri postavljanju upita
    - metode za pretraživanje baze i dohvat rezultata
  - način ispisivanja/predstavljanja rezultata upita
  - korisničko sučelje
  - višejezičnost
  - potrebni mrežni i računalni resursi

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 oznaka
- korisna URL adresa:
  - http://www.robotstxt.org/
  - http://www.searchenginewatch.com/webmasters/spiderchart.html

Robot Exclusion Protocol

- može rabiti samo osoba s pravom pisanja u odgovarajućem direktoriju (webmaster)
- robot.txt datoteka
  - posebna sintaksa
  - u početnom direktoriju Web poslužitelja
  - URL: http://hostname/robots.txt
- primjer:

```
User-agent: *
Disallow: /archives/
Disallow: /radni/
```

ROBOT META oznaka

- može rabiti autor Web stranice prema potrebi
- `<META NAME="ROBOTS" CONTENT="content">`

  content = ALL | NONE | directive ["," directive]

  directive = index | follow

  index = "INDEX" | "NOINDEX"

  follow = "FOLLOW" | "NOFOLLOW"
- default: INDEX, FOLLOW
- primjer:

```
<meta name="robots" content="index,nofollow">
```

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 (Keyword Truncation) - *, ?, %

  alumi*um

  comput*
- kaskadno pretraživanje (Infoseek)
- kontrola resursa

  `title:"Internet training"` (AltaVista, HotBot, …)

  `host:www.fer.hr` (AltaVista)

  `image:slika.jpg` (AltaVista)

  `related:` (Google)

Tražilice
Mogućnosti kod postavljanja upita (3)

- *natural language searching* (Ask Jeeves! - **http://www.ask.com/**)

- drugačiji pristupi:
  - Ditto.com - **http://www.ditto.com/**
  - (Oingo - **http://www.oingo.com/**)

- korisna URL adresa:
  - http://www.searchenginewatch.com/

Tražilice
Važne odlike

- Baza podataka (veličina, ažurnost, složenost) / rujan 2003.
  - Google – 3,3 milijardi Web stranica
  - AllTheWeb (FAST) – 3,2 milijardi Web stranica
  - INKTOMI - 3 milijarde Web stranica
  - AltaVista - 1 milijarda 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
  - metapodacima
  - 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.search.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 (2000.)
  - Open Directory - ≈60000 urednika, 3,8 milijuna Webova (2003.)
  - LookSmart - 200 urednika, 2,5 milijuna Webova (2001.)
- 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

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 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/**
  - ...

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
- **specijalizirana sučelja (za neko područje)**
  - nude kvalitetne informacije (ako postoje i znamo za njih)

Dio 2. - Metapodaci

Problemi?

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

Web: informacijski servis

- jednostavno publiciranje
  - manje barijera / lakši pristup
  - brzo i efikasno publiciranje (posebno za dinamičke izvore informacija)
- informacije su distribuirane
- upravljanje informacijskim prostorom je teško, mogućnosti su ograničene
- novi odnosi između autora, izdavača, distributera, posrednika i korisnika (potrošača)

Web ⇔ knjižnica

- pretraživački sustavi:
  - motivirani su oglašavanjem (?)
  - doseg im je omeđen i nepredvidiv
  - recall .vs. precision
  - index spam (metadata spam)
- otvoreni problemi s resursima:
  - kvaliteta, integriret, pouzdanost
  - arhiviranje
  - autorska prava

Metapodaci

- strukturirani podaci o podacima
  - pomažu u stvaranju reda u Web informacijskom prostoru
  - omogućuju automatsko pronalaženje / upravljanje informacijama
- mogu se rabiti u različite svrhe:
  - pronalaženje informacija (resource discovery)
  - vrednovanje sadržaja (content rating)
  - upravljanje pravima (rights management)
  - sigurnost i autentifikacija (security and authentication)
  - …

Izazovi

- osigurati potrebnu raznolikost (uporabe) metapodataka
- funkcionalnost $\Leftrightarrow$ jednostavnost
- proširivost $\Leftrightarrow$ interoperabilnost
- kreiranje i uporaba: ljudskim radom i strojno
- ispuniti specifične potrebe

metadata = pidgin language of the Internet

Modularnost i interoperabilnost

- Modularnost
  - kocept “lego kocaka”
  - omogućuje distriburirano upravljanje
  - odgovornost je na pravom mjestu
- Interoperabilnost zahtjeva dogovore o:
  - semantici (značenju pojedinih elemenata)
  - strukturi (razumljiva ljudima, čitljiva strojevima)
  - sintaksi (gramatici)

Zapisivanje metapodataka

- **osnovni načini pohrane:**
  - metapodaci uloženi (embedded metadata) u Web stanicu - HTML dokument (npr. HTML META tag)
  - metapodaci povezani s Web stranicom (HTTP header)
  - metapodaci dostupni preko treće strane po posebnom zahtjevu (eksplicitni HTTP GET)

- **sintaktička rješenja za zapisivanje:**
  - HTML - trenutno najrasprostranjenije i najjednostavnije rješenje (META oznaka)
  - RDF - temelji se na XML-u; posebno namjenjen za metapodatke
  - XML

HTML META oznaka

- omogućuje definiranje para *značajka - vrijednost* koji se veže uz dokument
- osnovni oblik META oznake:
  - par atributa **NAME** i **CONTENT**

```
<META NAME="value" CONTENT="value">
```

  - atributom **NAME** definira se značajka resursa (npr. autor ili naslov)
    - vrijednosti NAME atributa nisu standardizirane
    - DUBLIN CORE - prijedlog standarda
    - najčešće vrijednosti koje se rabe: AUTHOR, KEYWORDS, DESCRIPTION, TITLE
  - atributom **CONTENT** definira se vrijednost značajke definirane NAME atributom

META oznaka - primjeri

```
<HEAD>
<TITLE>Naslov</TITLE>
<META name="description" content=”Opis sadrzaja">
<META name="keywords" content=“kljucne rijeci">
</HEAD>
```

```
<HEAD>
<TITLE>Naslov</TITLE>
<META name=“DC.Creator” content=“December, John”>
</HEAD>
```

Dublin Core (DC)

- predloženi standard za metapodatke (RFC 2413)
- namjenjen opisivanju resursa radi olakšanog pronalaženja informacija
- sastoji se od temeljnog skupa od 15 elemenata (značajki kojima se opisuje mrežni resurs)
- elementi su definirani temeljem ISO 11179 standarda
- elementi nisu obvezni i mogu se ponavljati
- DC predviđa mogućnost profinjenja elementa (qualifiers)
- temeljni skup elemenata je proširiv
- moguća je nadopuna drugim metapodacima
- specifikacija (ver 1.1)
  - http://dublincore.org/documents/dces/

DC elementi (DCMES)

- Title
- Creator (author)
- Subject (keywords)
- Description
- Publisher
- Other Contributor
- Date
- Resource Type
- Format
- Resource Identifier
- Source
- Language
- Relation
- Coverage
- Rights Management

DC kvalifikatori (qualifiers)

- profinjenje osnovnih elemenata
  - profinjuju značenje pojedinog elementa
- definiraju *encoding schemas*
  - određuju pravila ili algoritme za izvrednjavanje ili interpretiranje podataka
  - kontrolirani rječnici, klasifikacijske sheme, liste

- specifikacija:
  - http://dublincore.org/documents/dcmes-qualifiers/

DC u HTML-u

- META oznaka:
  - sintaksa:

<META name="PREFIX.ELEMENT_NAME" content="ELEMENT_VALUE">

  - primjer:

<META name="DC.Creator" content="December, John">

DC u HTML-u (2)

- LINK oznaka
  - sintaksa:

<LINK rel="schema.PREFIX" href="LOCATION_OF_DEFINITION">

  - primjer:

<LINK rel="schema.DC" ref="http://dublincore.org/documents/dces/">

DC u HTML-u (3)

- ponavljanje elemenata
  - primjer:

```
<META name="DC.Creator" content="Green, John">
<META name="DC.Creator" content="Brown, Fred">
```

DC u HTML-u (4)

- dodatne mogućnosti (profinjenje):
  - sintaksa:

```
<META
      name="PREFIX.ELEMENT_NAME.SUBELEMENT_NAME"
      content="ELEMENT_VALUE"
      scheme="SCHEME"
      lang="LANGUAGE">
```

  - primjeri:

```
<META name="DC.Date.Created" content="2000-08-01"
   scheme="ISO8601">
<META name="DC.Relation.isPartOf"
   content="http://www.somewhere.net">
```

DC Metadata Initiative (DCMI)

- Formalna podrška razvoju DC-a
- Radne skupine
- Domain-specific initiatives
  - DC-Education, DC-Libraries, DC-Government, …
- DC-Registry
  - http://www.schemas-forum.org/
- Redoviti sastanci / radionice
- http://dublincore.org/

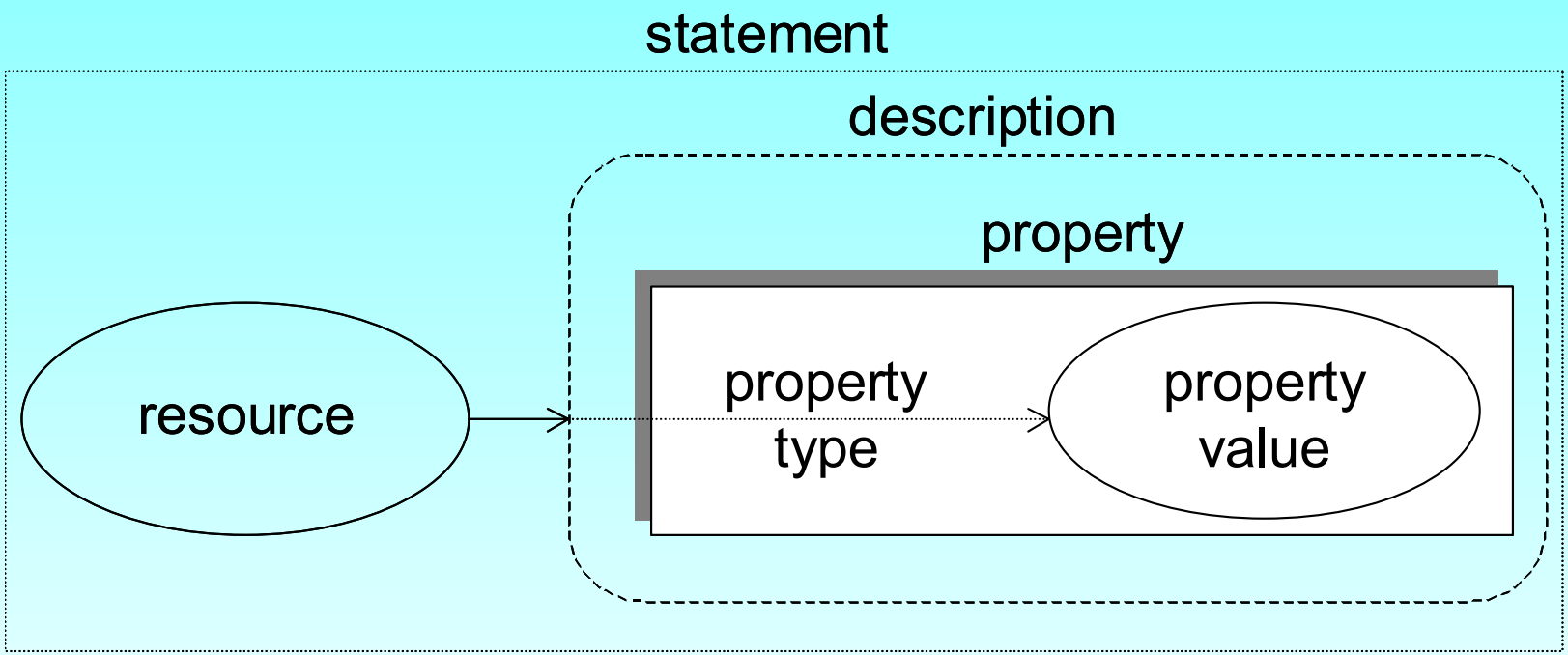
Resource Description Framework

- W3C standard (http://www.w3.org/RDF/)
- generalizirani pogled na metapodatke
- opće namjene
- strukturirani, strojno “razumljivi” metapodaci
- metapodatkovni rječnici (sheme) mogu se razvijati decentralizirano (bez središnjeg nadzora)
- podrška za autentikaciju metapodataka (i povjerenje u njihovu kvalitetu)

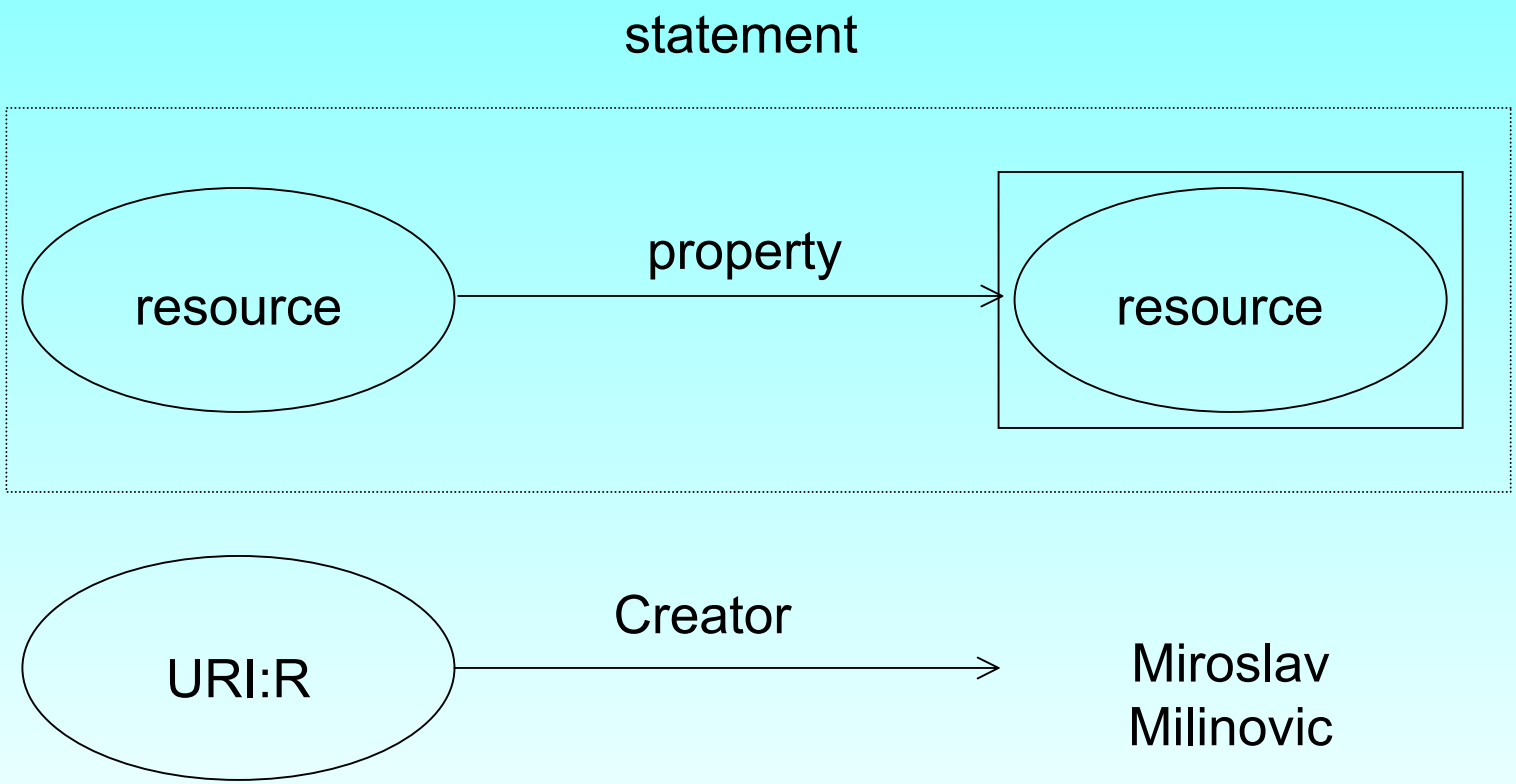
RDF koncept

- model podataka
- sheme
- sintaksa (u XML)

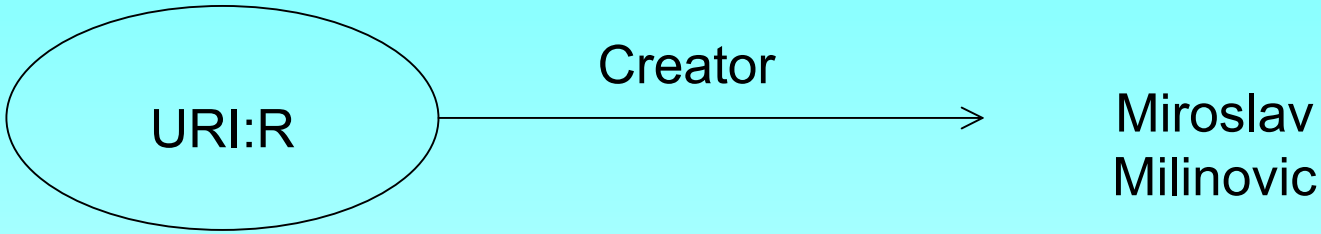
RDF model podataka (1)



RDF model podataka (2)



RDF (XML) sintaksa




```
<RDF xmlns =“http://www.w3.org/TR/WD-rdf-syntax#”
    xmlns:dc=“http://dublincore.org/documents/dces/”>
    <Description about =“URI:R”>
        <dc:Creator>Miroslav Milinovic</dc:Creator>
    </Description>
</RDF>
```

RDF kontejneri (containers)

- više vrijednosti za istu značajku (property)
- 3 tipa kontejnera:
  - **Bag**
    - grupa (unordered grouping)
  - **Sequence**
    - uređena grupa (ordered grouping)
  - **Alternatives**
    - alternativne vrijednosti

RDF kontejneri - primjer

- grupa

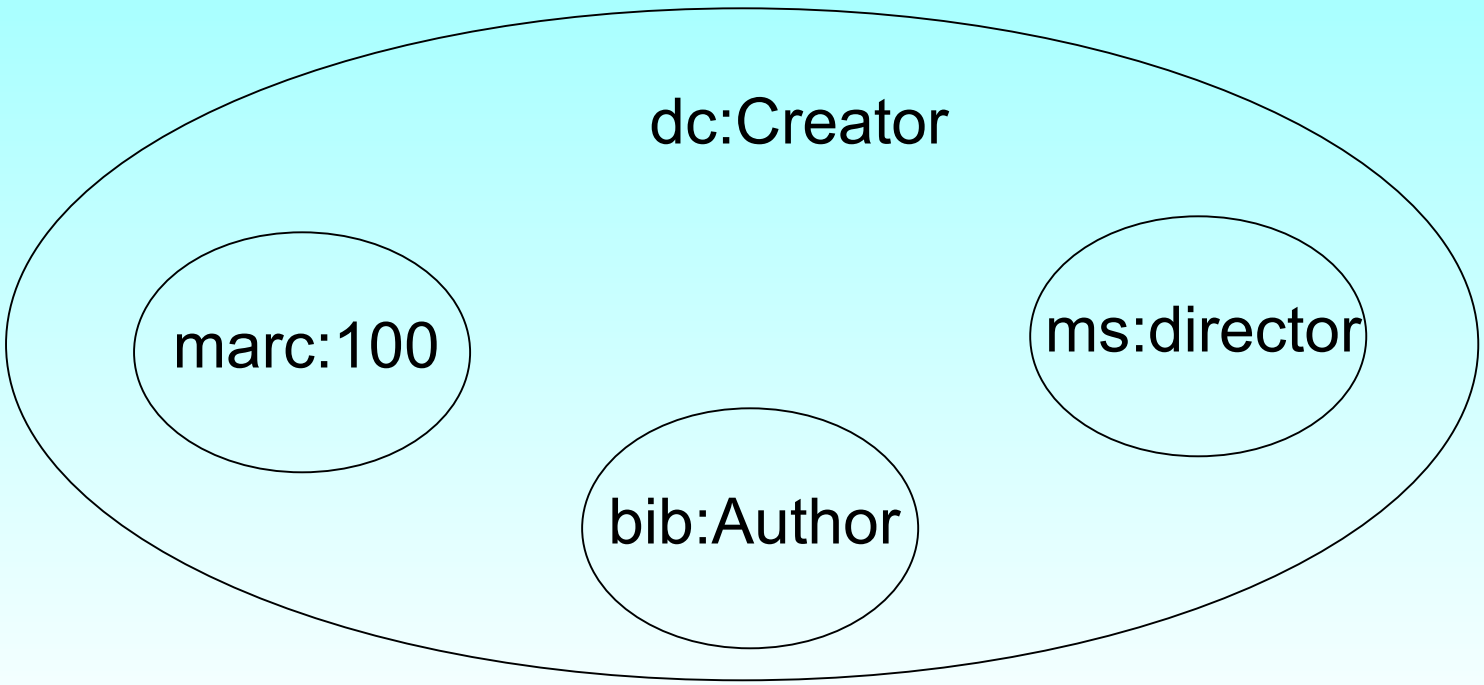
```
<BIB:Author>
   <Bag>
            <li> Miro </li>
            <li> Ivan </li>
   </Bag>
</BIB:Author>
```

- uređena grupa

```
<BIB:Author>
   <Seq>
            <li> Miro </li>
            <li> Ivan </li>
   </Seq>
</BIB:Author>
```

RDF scheme

- deklaracije rječnika (sadrže definicije značajki)
- omogućuju interoperabilnost



Ostali standardi

- P3P - Platform for Privacy Preferences Project
- PICS - Platform for Internet Content Selection
- DSig - Digital Signatures
- CC/PP - Composite Capabilities/Preference Profiles
- …
- korisna adresa:
  - http://www.w3.org/Metadata/

Metapodaci u praksi

- standardi (polako) dozrijevaju
- istraživanja: DC-DOT, Reggie, CORC, MAENAD, Nordic metadata project, SAFARI, …
- alati:
  - su tu, ali još u razvoju
  - nedostaje veza s ostalim (popularnim) programima
  - traži se bolja konfigurabilnost (sheme, jezici, formati, ...)
  - W3C RDF Validation Service: http://www.w3.org/RDF/Validator/

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

Trenutno stanje

- nema pravog standarda (?)
  - Dublin Core je dobar kandidat
- HTML ima META oznaku
  - treba rabiti uz nužan oprez
- pretraživački mehanizmi rabe metapodatke (?)
- korisne adrese:

  W3C: http://www.w3.org/Metadata/

  Dublin Core: http://dublincore.org/

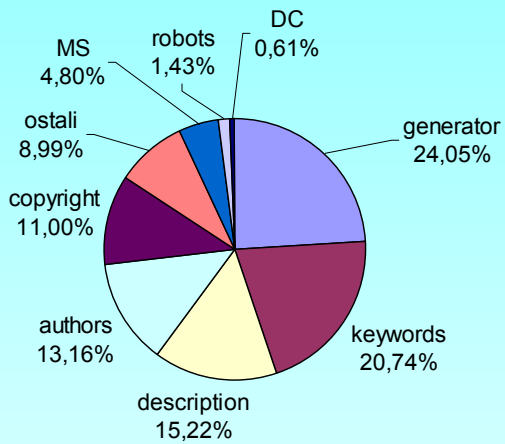
Trenutno stanje (2)

- oko 800 miliona Web stranica
- 15 TB (6 TB) podataka
- običnu HTML META oznaku ima 34% stranica
- Dublin Core standard poštuje 0,3 % stranica
- META oznaka se šaroliko rabi - 123 različita tipa

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

Metapodaci u hrvatskom Web prostoru

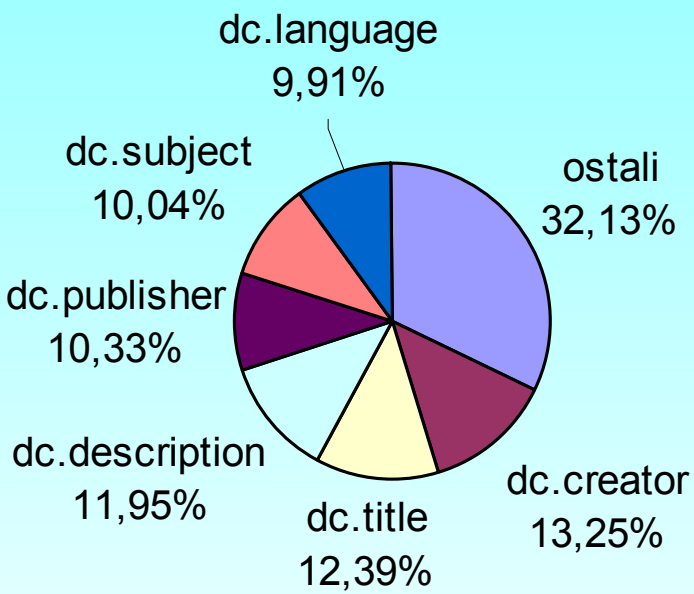
- 31% HTML resursa ima META oznaku
- 744 različite vrijednosti NAME atributa META oznake
- nebriga autora (?)
- udio “standarda”:
  - Dublin Core – 0,09%
  - alati za izradu Web stranica – 25%
  - tražilice – 19,7%
  - ROBOTS META oznaka – 1,35%



izvor: http://www.srce.hr/mwp/

Dublin Core
u hrvatskom Web prostoru

Frekvencija uporabe različitih DC elemenata



izvor: http://www.srce.hr/mwp/

O čemu je bilo riječi?

- Dio 1. - Pronalaženje informacija na Internetu
  - Internetski prostor informacija
  - pretraživanje Weba
  - tražilice
  - tematski katalozi
- Dio 2. - Metapodaci
  - motivi i koncepti
  - Dublin Core
  - RDF