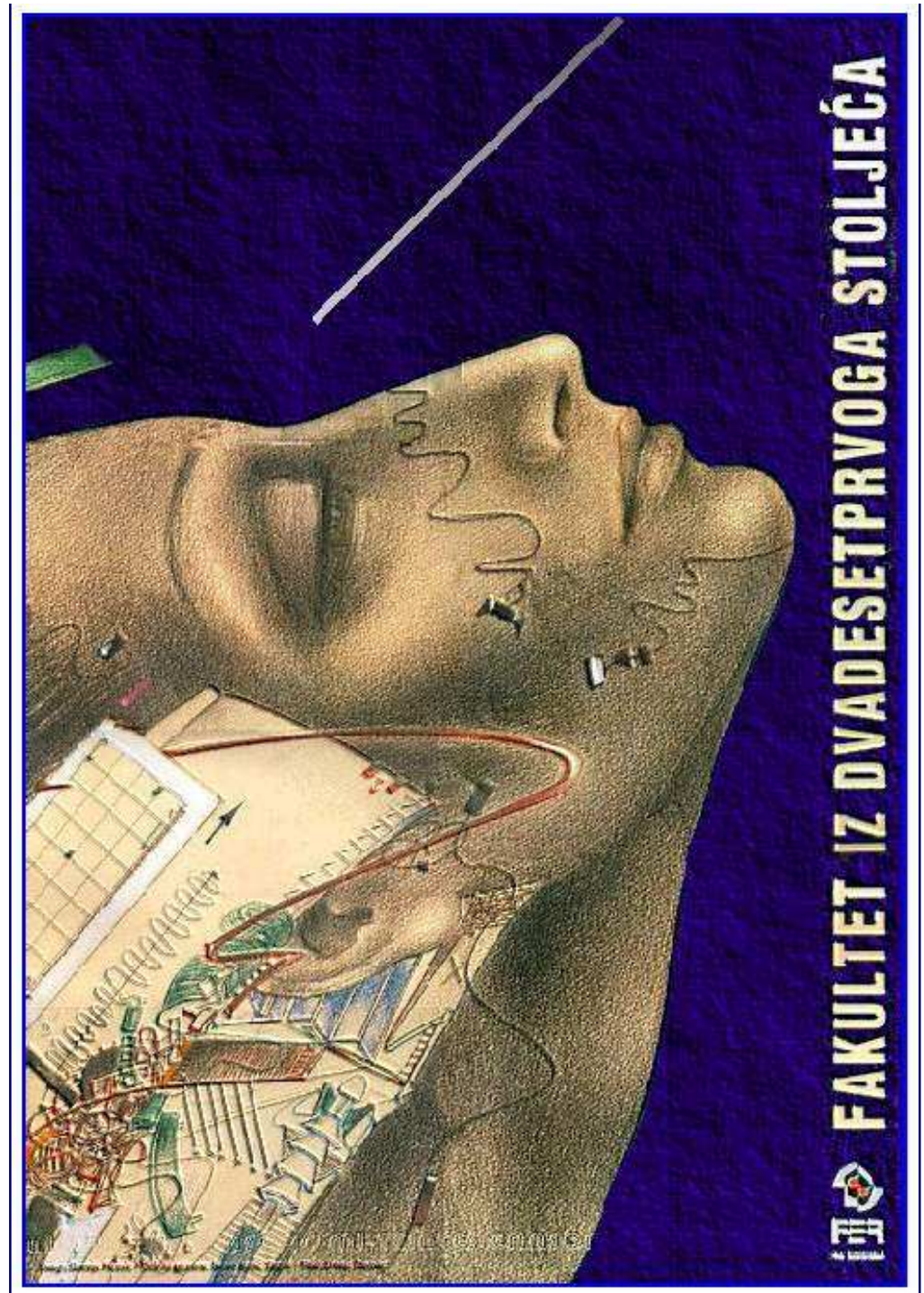


# Napredni modeli i baze podataka

Predavanja

## 11. Neki izvori podataka na Internetu

Prosinac 2008.



# Uvod

---

- Ljudi koriste Internet svakodnevno kao izvor podataka i informacija
- Koji su postojeći izvori korisni i kako ih možemo upotrijebiti prilikom izrade vlastitih aplikacija?

# Uvod

---

- Korištenjem otvorenih sučelja i strukturiranih podataka osigurava strojevima tj. aplikacijama čitanje i razumijevanje podataka na Internetu
- Time se otvara mogućnost korištenja Interneta kao jedne velike "baze podataka"
- Ova je tema usko vezana za **semantički web** i pojmove iz te domene kao što je mreža podataka (*web of data*) i povezani podaci (*linked data*)

# Pregled

---

- Neke javno dostupne baze podataka
- CiteSeer
- freedb
- MusicBrainz
- Flickr
- Google
  - Napredno pretraživanje
  - Google Base
- Freebase
- Wikipedia
  - MediaWiki
  - DBpedia
  - WikiXMLDB
- Mashup aplikacije
  - Yahoo! Pipes

# Primjeri

---

- Primjeri popularnih javno dostupnih baza podataka:
  - All Media Guide (Allmusic, Allgame, Allmovie), BiblioPage, IMDb, Library of Congress, National Library of Medicine, TV.com, itd.
- Primjeri znanstvenih baza podataka dostupnih na Internetu:
  - CiteSeer, OvidSP, Scopus, Science Direct, ISI Web of Knowledge, Engineering Village, Wiley Interscience, SpringerLink, Scopus, itd.
  - Detaljnije se može pogledati na stranicama instituta Ruđer Bošković: <http://www.online-baze.hr/>

# CiteSeer

---

- *CiteSeer* je besplatna digitalna knjižnica i tražilica znanstvene literature koja se prvenstveno fokusira na područje računalne i informatičke znanosti
- Trenutačno se koristi nova verzija *CiteSeer<sup>x</sup>*
- Osim literature *CiteSeer* nastoji ponuditi resurse poput algoritama, podataka, metapodataka, usluga, tehnika i softwarea
- Prvi su ponudili automatsko indeksiranje i povezivanje citata (*Autonomous Citation Indexing*)

# CiteSeer – posebnosti (features)

---

- Statistike citata svih članaka
- Pregledavanje baze podataka koristeći linkove citata
- Može prikazati kontekst citata odabranog članka
- Automatske obavijesti o novom citiranju odabranog članka
- Pronalazi tematski povezane dokumente na temelju citata i analize riječi
- Potpuno indeksiranje teksta i citata članaka
- Automatski dohvaća nove članke s Interneta
- Automatski izvlači metapodatke iz članaka

# CiteSeer i ostali

---

- Osim *CiteSeer*-a postoje i slične digitalne knjižnice znanstvenih članaka kao što su ACM i IEEE koje su jako popularne i kvalitetne, ali nisu besplatne
- *Google Scholar* je tražilica koja indeksira znanstvenu literaturu od različitih izvora, vezanu za različite discipline i područja
- Hrvatska znanstvena bibliografija ([bib.irb.hr](http://bib.irb.hr))



## *freedb*

---

- *freedb* je baza podataka Internetu za dohvat informacija o glazbenom CD-u
- Disk ID se dobije preko izračuna hash funkcije tablice sadržaja CD-a, a služi kao primarni ključ u bazi podataka
- Iz baze se mogu dohvatiti podaci o imenu glazbenika/grupe, imenu albuma, popisu pjesama, godini izdanja, žanru itd.
- Korisnici sami unose podatke o novim albumima

# MusicBrainz

---

- Slično kao i *freedb*, *MusicBrainz* je započeo kao baza glazbenih metapodataka, ali se razvija u strukturiranu "Wikipediju za glazbu"
- Za razliku od *freedb* postoji web sučelje preko kojeg se podaci mogu detaljno pregledavati i uređivati
- Za identifikaciju glazbe koristi se "akustički otisak prsta" preko servisa *MusicDNS*
- Podaci se opisuju koristeći RDF i XML tehnologije

# *Flickr*

---

- *Flickr* je najpopularnija web stranica za udomljavanje slika koja je postala poznata zahvaljujući svojim organizacijskim alatima koji omogućuju označavanje slika
- *Flickr Services* - programsko sučelje
- *Flickr API* je potpuno otvoren, tako da svatko može koristiti sadržaj (slike, video, oznake, grupe itd.) iz *Flickr*-ove baze podataka u raznim programima i aplikacijama

# Google napredno pretraživanje

---

- Google podržava napredne operatore kojima se odabire tip pretrage ili se modificira na različite načine
- Osnovni operatori:
  - " " Upit s izrazom unutar navodnika traži točno taj izraz (npr. "Alisa u zemlji čudesa")
  - \* Zamijenjuje jednu ili više riječi (npr. "Alisa \* čudesa")
  - | Jedna, druga ili obje riječi (npr. hotel Tahiti | Hawaii)
  - + Uključuje riječ u pretragu (npr. Star Wars +"I")
  - Isključuje riječ iz pretrage (npr. Hurt -ringtone)
  - ~ Pretraživa i sinonime riječi, te riječi s drugačijim završetkom (npr. ~car )
  - .. Koristi se za raspon brojeva (npr. Olimpijada 2000..2008)

# Google napredno pretraživanje

---

- Alternativni tipovi upita:

**cache:** Prikazuje verziju web stranice koja se nalazi u *Google* priručnoj memoriji. Ako se dodatno navedu još neke riječi, bit će označene bojom (npr. **cache:www.fer.hr student**)

**link:** Vraća sve stranice na kojima se nalazi link na navedenu stranicu (npr. **link:www.google.com**)

**related:** Vraća web stranice koje su slične navedenoj stranici (npr. **related:www.zpr.fer.hr**)

**info:** Vraća informacije koje Google posjeduje o nekoj web stranici (npr. **info:www.fer.hr**)

# Google napredno pretraživanje

---

- Modifikatori upita:

- site:** Ograničava pretragu na navedenu domenu (npr. **help site:www.google.com**)
- allintitle:** Vraća samo one stranice kojima su navedene riječi u naslovu (npr. **allintitle: search google**)
- intitle:** Vraća one stranice kojima je riječ iza operatora u naslovu. Druge navedene riječi mogu se pojaviti bilo gdje na stranici (npr. **intitle:search google**)
- allinurl:** Vraća stranice kojima su navedene riječi unutar URL-a (npr. **allinurl: fer zpr**)
- inurl:** Vraća stranice kojima je jedna riječ iza operatora unutar URL-a (npr. **inurl:fer zpr**)
- filetype:** traži podatke zadanog formata (npr. **diplomski java filetype:pdf**)

# Google napredno pretraživanje

- Primjer – Pretraga glazbe koristeći Google



**imagine** intitle:"index of" "parent directory" "size" "last modified" "description" (mp3|flac|aac|ape|ogg)

-inurl:(jsp|php|html|aspx|htm|cf|shtml|lyrics-realm|mp3-collection) -site:.info

- Specijalizirani informativni upiti:

- **define:**, **weather:**, **movie:**, **stocks:**, **phonebook:**, itd.

## Index of /102-john\_lennon-imagine.mp3

<u>Name</u>	<u>Last modified</u>	<u>Size</u>	<u>Description</u>
 <a href="#">Parent Directory</a>	10-Aug-2007 18:13	-	
 <a href="#">102-john lennon-imagine.m..&gt;</a>	22-Jan-2006 12:4	4.6M	

# Google Base

---

- *Google Base* omogućuje da korisnici mogu dodati bilo kakav tip sadržaja, opisati ga atributima, a ovisno o važnosti sadržaja, *Google* će ga uvrstiti u pretragu
- Npr. kod recepta za neko jelo, atributi su sastojci sa zadanim količinama
- Ako se sadržaj već nalazi na webu, *Google* će ga povezati linkovima



# Google Base

---

- *Google Base Data API* služi za programski pristup, a omogućuje:
  - Upravljanje strukturiranim podacima:
    - dodavanje novih objekata
    - izmjenu i brisanje postojećih
  - API je podržava bogati upitni jeziku

# Freebase

---

- Freebase je visoko strukturirana otvorena baza podataka koja sadrži međusobno povezane podatke (*cross-linked data*) izvučene s drugih izvora kao što je *Wikipedija* ili *MusicBrainz*, ali i podatke koje su individualno unijeli korisnici
- Koristi posebnu infrastrukturu koja je razvijena unutar kompanije (*Metaweb*) temeljena na modelu grafa
- Otvoreno programsko sučelje (API) omogućuje programerima i aplikacijama pristup svim podacima
- Freebase koristi MQL (*Metaweb Query Language*) upitni jezik razvijen posebno za *Freebase*

# Freebase

---

- *Freebase* i *Wikipedija* razlikuju se po tipovima i po organizaciji podataka. Kod *Wikipedije* podaci su organizirani u članke namijenjene ljudima, dok *Freebase* sadrži liste činjenica i statistika koje se lako koriste u drugim aplikacijama i web stranicama
- *Google Base* je sličan *Freebaseu*, ali s različitim posebnostima i različitim tipovima informacija koji su pokriveni. *Google Base* je orijentiran na objekte (*things*) i događaje (*events*) unutar neke kategorije. Tu kategoriju unosi korisnik, a drugim korisnicima omogućuje čitanje, ali ne i izmjenu. *Freebase* je zajednička i svi korisnici mogu mijenjati podatke.

# Wikipedija

---

- Wikipedija trenutčno sadrži preko 10 milijuna članaka na više od 250 jezika i svakako je zanimljiv izvor podataka
- Ali da bi se Wikipedija koristila kao izvor podataka potrebno je omogućit programski pristup podacima. To je moguće na više načina, a najpopularniji su:
  - *MediaWiki API*
  - *DBpedia*
  - *WikiXMLDB*

# *MediaWiki API*

---

- *MediaWiki* je wiki programski paket koji je originalno bio namijenjen za Wikipediju, ali danas ga koriste i ostali projekti organizacije *Wikimedia*
- *MediaWiki API* je programsko sučelje za *MediaWiki* programski paket, a time i službeni API za Wikipediju
- Još je u aktivnom razvoju, te je taj relativno kasni početak omogućio procvat neslužbenih sučelja i drugih projekata vezanih uz dohvat podataka s Wikipedija kao što je *DBpedia* i *WikiXMLDB*

# ***MediaWiki API***

---

- Cilj ovog programskog sučelja je pružiti direktan pristup *MediaWiki* bazama podataka
- Korisnički programi mogu koristiti API za prijavu, dohvat podataka i promjenu sadržaja
- API kao ulaz prima upitni znakovni niz
- Izlaz je moguć u raznim izlaznim formatima (XML, JSON, YAML...)

# MediaWiki API - primjer

---

- Dohvati listu kategorija u kojima se nalazi Albert Einstein:

```
api.php ? action=query & titles=Albert%20Einstein & prop=categories
```

- Rezultat vraćen u XML-u:

```
<api>
  <query>
    <pages>
      <page pageid="736" ns="0" title="Albert Einstein">
        <categories>
          <cl ns="14" title="Category:1879 births" />
          <cl ns="14" title="Category:1955 deaths" />
          <cl ns="14" title="Category:Albert Einstein" />
          ...
        </categories>
      </page>
    </pages>
  </query>
</api>
```

# Semantic MediaWiki (SMW)

---

- Dodatak *MediaWiki* programskom paketu koji omogućuje kodiranje semantičkih podataka unutar wiki stranica
- Ti podaci se mogu koristiti kod semantičkih pretraga i kod združivanja (aggregation) stranica, te se mogu prikazati putem RDF-a

- Primjer:

The population of `[[city::Zagreb]]` in  
`[[year::2006]]` was `[[population::784,900]]`.



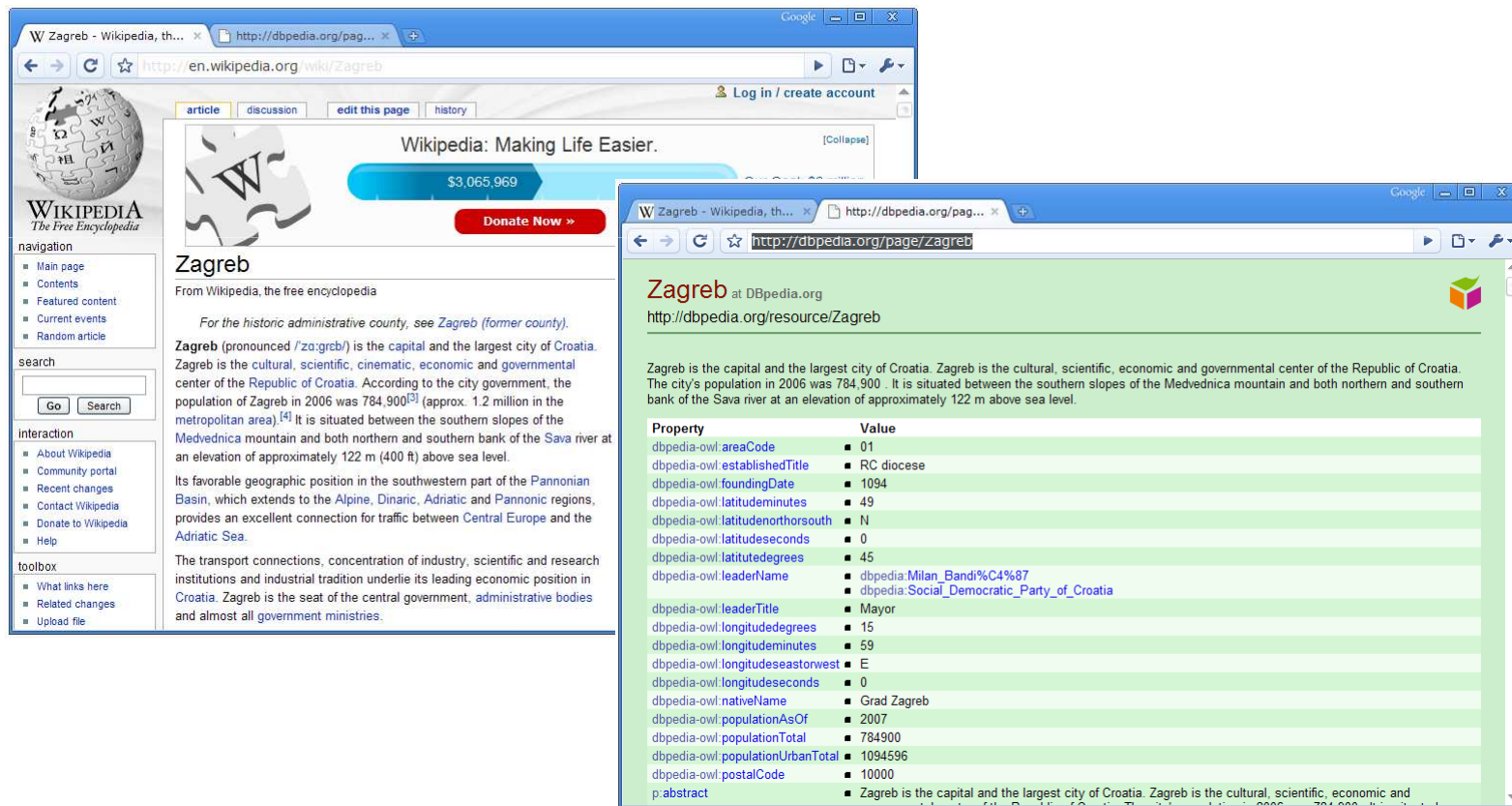
# ***DBpedia***

---

- *DBpedia* je zajednica na Internetu kojoj su ciljevi:
  - Izvući strukturirane podatke na Wikipediji
  - Pružiti te podatke na Internetu pod slobodnom licencom
  - Povezati podatke s Wikipedije s drugim skupovima podataka na Internetu
- *DBpedia dataset*
  - Trenutačno sadrži 2.6 milijuna objekata (213,000 ljudi, 328,000 mjesta, 57,000 glazbenih albuma, 30,000 filmova itd.)
  - Za svaki objekt postoji etiketa i kratki opis u 14 različitih jezika
  - 609,000 linkova na slike, 3,150,000 linkova na vanjske web stranice
  - Međusobno je povezan na RDF razini s drugim slobodnim skupovima podataka na Internetu (Freebase, OpenCyc, UMBEL, GeoNames itd.)

# DBpedia - primjer

- <http://en.wikipedia.org/wiki/Zagreb>
- <http://dbpedia.org/page/Zagreb>



The image shows two overlapping browser windows. The background window displays the Wikipedia page for Zagreb, featuring the Wikipedia logo, navigation links, and a search bar. The foreground window displays the DBpedia page for Zagreb, which includes a table of properties and values for the city.

**Zagreb** at DBpedia.org  
<http://dbpedia.org/resource/Zagreb>

Zagreb is the capital and the largest city of Croatia. Zagreb is the cultural, scientific, economic and governmental center of the Republic of Croatia. The city's population in 2006 was 784,900 . It is situated between the southern slopes of the Medvednica mountain and both northern and southern bank of the Sava river at an elevation of approximately 122 m above sea level.

Property	Value
dbpedia-owl:areaCode	01
dbpedia-owl:establishedTitle	RC diocese
dbpedia-owl:foundingDate	1094
dbpedia-owl:latitudeminutes	49
dbpedia-owl:latitudenorthorsouth	N
dbpedia-owl:latitudeseconds	0
dbpedia-owl:latitudedegrees	45
dbpedia-owl:leaderName	dbpedia:Milan_Bandi%C4%87 dbpedia:Social_Democratic_Party_of_Croatia
dbpedia-owl:leaderTitle	Mayor
dbpedia-owl:longitudedegrees	15
dbpedia-owl:longitudeminutes	59
dbpedia-owl:longitudeseastorwest	E
dbpedia-owl:longitudeseconds	0
dbpedia-owl:nativeName	Grad Zagreb
dbpedia-owl:populationAsOf	2007
dbpedia-owl:populationTotal	784900
dbpedia-owl:populationUrbanTotal	1094596
dbpedia-owl:postalCode	10000
p:abstract	Zagreb is the capital and the largest city of Croatia. Zagreb is the cultural, scientific, economic and

# DBpedia

---

- DBpedia skupu podataka se može pristupiti na tri načina:
  - *SPARQL Endpoint* - omogućuje postavljanje upita koristeći SPARQL upitni jezik. Također se može koristiti SNORQL query explorer
  - Pošto je DBpedia skup podataka u obliku povezanih podataka (linked data) moguće ga je pregledavati pomoću preglednika za semantički web (npr. Tabulator)
  - sami podaci se mogu i skinuti s DBpedia.org

# DBpedia – SPARQL primjer

- Dohvati ime poznatih osoba rođenih u Zagrebu:

```
PREFIX dbpedia2: <http://dbpedia.org/property/>
```

```
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
```

```
SELECT ?name WHERE {
```

```
?person dbpedia2:birthPlace <http://dbpedia.org/resource/Zagreb>.
```

```
?person foaf:name ?name.
```

```
}
```

- Rezultat upita:



# WikiXMLDB

---

- Omogućuje postavljanje upita Wikipediji pomoću XQuery upitnog jezika
- Sadržaj s Wikipedije je parsiran u pravilno strukturirani XML prikaz i pohranjan u *Sedna* XML bazu podataka
- Pristup podacima je omogućen preko *XQuery* web sučelja

# WikiXMLDB - primjer

---

- Dohvati sve hrvatske pisce koji imaju neke veze sa Zagrebom (tj. na njihovoj stranici na Wikipediji se nalazi link za Zagreb):

```
declare default element namespace
    "http://www.mediawiki.org/xml/export-0.3/";
```

```
(index-scan('article-by-link','Zagreb','EQ')
intersect
index-scan('article-by-cat','Category:Croatian
    writers','EQ'))/title/text()
```

- Rezultat:

Ivo Andrić, Miroslav Krleža, August Šenoa, Antun Branko Šimić, Vladimir Nazor, Ljudevit Gaj, Silvije Strahimir Kranjčević...

# ***Mashup aplikacije***

---

- kombinirajući podatke iz raznih (vanjskih) izvora stvara se potpuno nova usluga
- Sadržaj korišten u Mashup aplikacijama se tipično dohvaća preko javnih sučelja (API), mrežnih usluga (Web Services), Web feeds (RSS, Atom) i na razne druge načine izvlačenja i dohvata podataka (npr. Screen scraping)

# Mashup aplikacije

---

- Mogu se podijeliti na sljedeće tipove:
  - Potrošački (Consumer Mashup)
  - Poslovni (Business Mashup)
  - Podatkovni (Data Mashup)
- Primjena Mashup aplikacija je raznolika kao i sam Internet, ali se mogu podijeliti u sljedeće kategorije:
  - Mapiranje
    - različiti setovi podataka se mogu predložiti grafički koristeći karte (npr. Google Maps, Yahoo Maps itd.)
  - Foto i video
    - popularni zbog raznih servisa kao što je Flickr i YouTube, koji pohranju velike količine slika, odnosno video dokumenata
    - veliki broj metapodata vezan za pohranjeni sadržaj
  - Pretraga i kupovina
    - popularnost znatno skočila nakon što su Amazon i eBay ponudili sučelje za programski dohvat njihovih podataka
  - Vijesti
    - izvori vijesti kao BBC, Reuters, New York Times i drugi već odavno koriste tehnologije kao što je RSS ili Atom da bi ponudili vijesti vezane za različite teme



# ***Mashup aplikacije***

---

- *Mashup* aplikacija se sastoji od tri dijela koji su logički i fizički razdvojeni:
  - Pružatelji sadržaja/API
    - Da bi olakšali korištenje sadržaja, izlažu ga putem različitih tehnologija
    - Mnogi još nemaju otvoreni API, pa se podaci dohvaćaju metodama izvlačenja podataka
  - *Mashup* stranica
    - Tu se nalazi Mashup logika. Ali se ne mora tu izvršavati, već to može i na strani korisnika.
  - Korisnik tj. internet preglednik
    - Mjesto gdje se aplikacija i rezultati prikazuju grafički i gdje se odvija interakcija s korisnikom

# Mashup aplikacije

---

- Tehnički problemi
  - Problemi kod integracije podataka s različitih izvora
  - Kvaliteta podataka upitna (nekonzistentni i nepotpuni podaci)
  - Onečišćenje podataka (*data pollution*) – pogotovo kod javnih servisa (*Flickr, Wikipedia* itd.)
- Društveni problemi
  - Intelektualno vlasništvo podataka
  - Privatnost podataka
  - Sigurnost

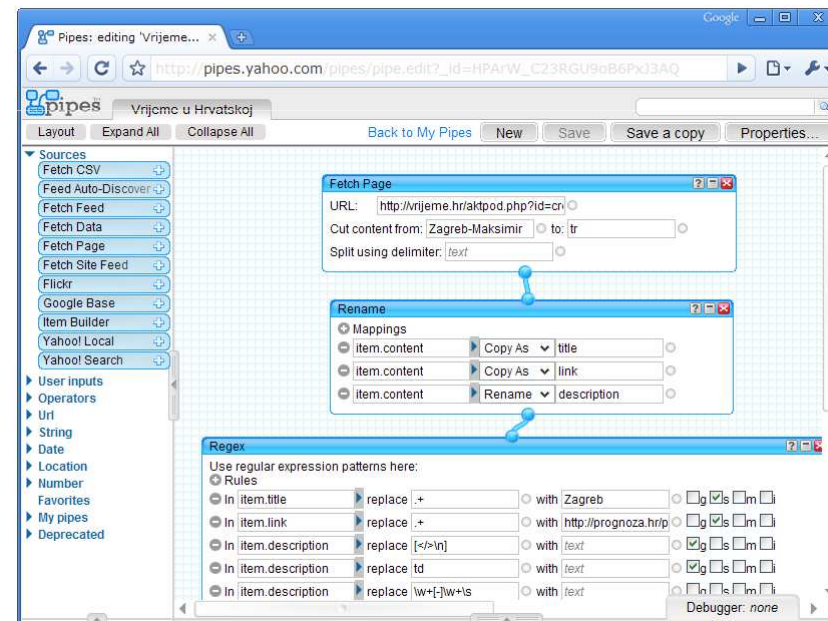
# Mashup aplikacije u stvarnom svijetu

---

- Mashup aplikacije su još u ranom stadiju razvoja i koriste se više za igru, nego kao neke ozbiljne aplikacije
- Nedostatak robusnih standarda, protokola i modela
- Popularni *Mashup* editori:
  - *Dapper*
  - *Yahoo! Pipes*
  - *Google Mashup Editor*
  - *IBM Mashup Center*
  - *Microsoft Popfly*
  - *Intel Mash Maker*

# Yahoo! Pipes

- Web aplikacija koja kroz grafičko sučelje omogućuje stvaranje novih aplikacija ili usluga, korištenjem vanjskih *web feed*-ova, web stranica i drugih usluga, te njihovo objavljivanje u raznim formatima
- Napravljeni po uzoru na *Unix pipes*. Jednostavne naredbe je moguće zajedno kombinirati i tako dobiti traženi rezultat



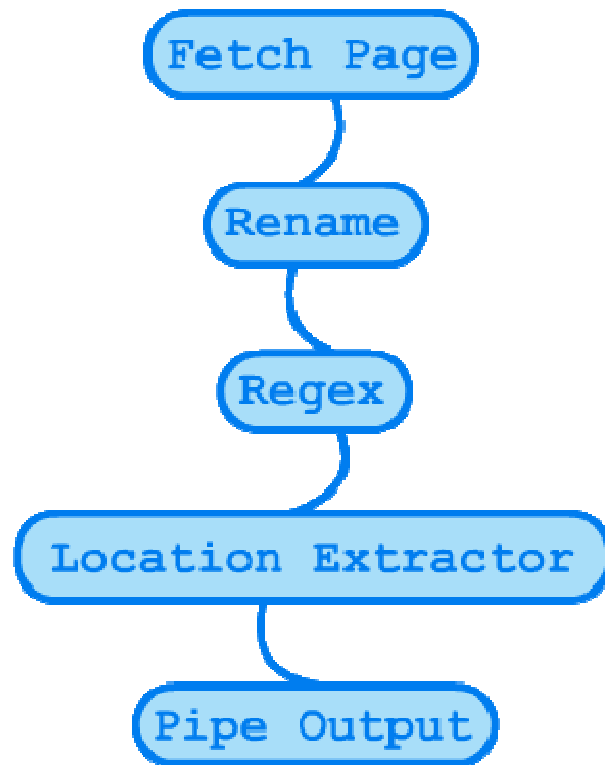
# Yahoo! Pipes - primjer

---

- Sa stranica DHMZ-a uzimaju se aktualni podaci o vremenu za Zagreb i Dubrovnik, te se grafički prikazuju na karti ([link](#))
- Stranice DHMZ-a nemaju otvoreni API ili *feed*, tako da se podaci najprije moraju izvući iz HTML kôda i generirati *web feed*
- *Feed* se prosljeđuje dalje i svakom skupu podataka se dodjeljuje prostorno obilježje

# Yahoo! Pipes - primjer

---



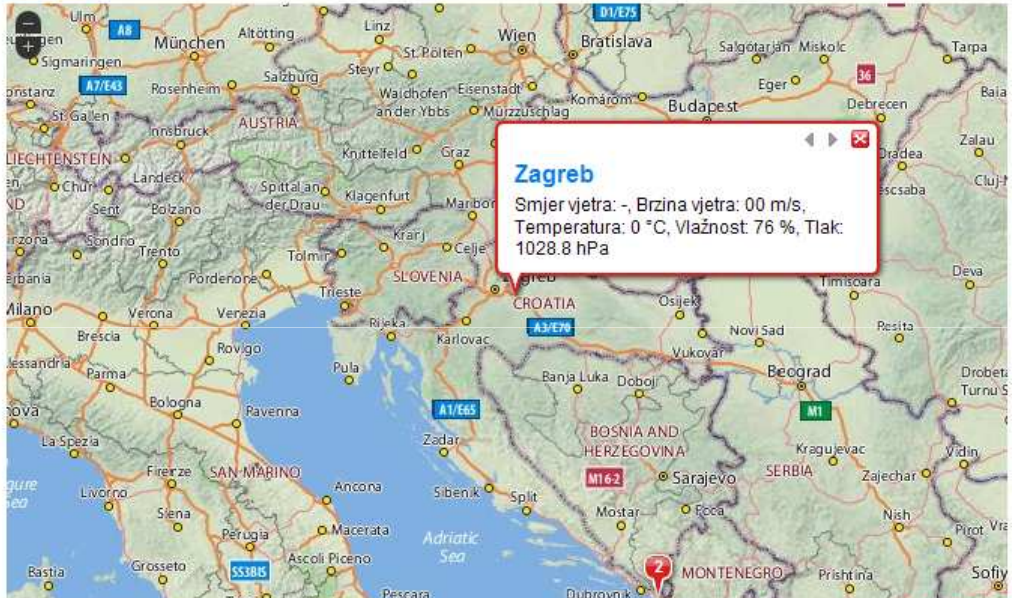
- Dohvaća se točno određeni dio HTML kôda stranice
- Stvaraju se standardna RSS polja
- Modul za operacije s regularnim izrazima. Pomoću njega se iz HTML kôda izvlače traženi podaci, te se spremaju u predviđena RSS polja
- *RSS feed-u* se pridružuje prostorno obilježje
- Izlaz

# Yahoo! Pipes - primjer

Use this Pipe

Get as a Badge MY YAHOO! Google Results by Email or Phone More options

Map List 2 items



**Zagreb**  
Smjer vjeta: -, Brzina vjeta: 00 m/s,  
Temperatura: 0 °C, Vlažnost: 76 %, Tlak:  
1028.8 hPa

Map Sat Hyb Data © 2008 NAVTEQ

Use this Pipe

Get as a Badge MY YAHOO! Google Results by Email or Phone More options

Map List 2 items

**Zagreb**  
Smjer vjeta: -, Brzina vjeta: 00 m/s, Temperatura: 0 °C, Vlažnost: 76 %, Tlak: 1028.8 hPa

**Dubrovnik**  
Smjer vjeta: NW, Brzina vjeta: 01 m/s, Temperatura: 7 °C, Vlažnost: 51 %, Tlak: 1024.2 hPa