

UMREŽAVANJE PODATAKA NA WEB-U

JELENA JOVANOVIĆ

EMAIL: JELJOV@GMAIL.COM

WEB: [HTTP://JELENAJOVANOVIĆ.NET](http://JELENAJOVANOVIĆ.NET)



KRATKO PODSEĆANJE:

GIGANTSKI GLOBALNI GRAF I

UMREŽENI PODACI NA WEB-U (WEB PODATAKA)

GIGANTSKI GLOBALNI GRAF (1)

International Information Infrastructure (III)

- graf/mreža računara poznata kao *Internet* ili *Net*
- *"It isn't the cables, it is the computers which are interesting"*

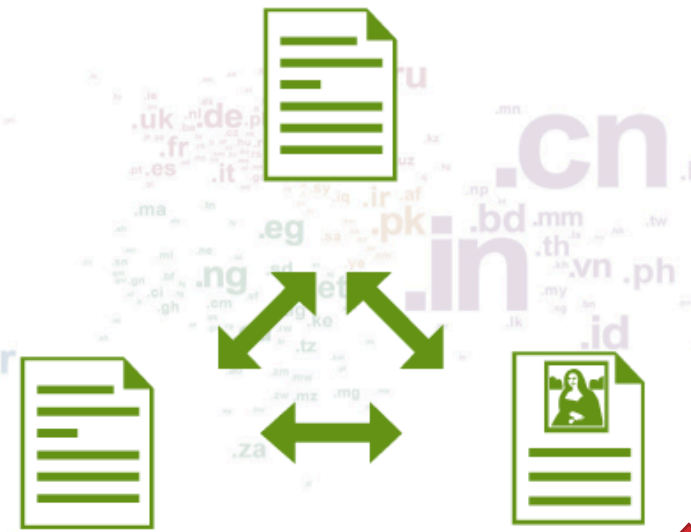
World Wide Web (WWW)

- graf/mreža dokumenata poznata kao *Web*
- *"It isn't the computers, but the documents which are interesting"*

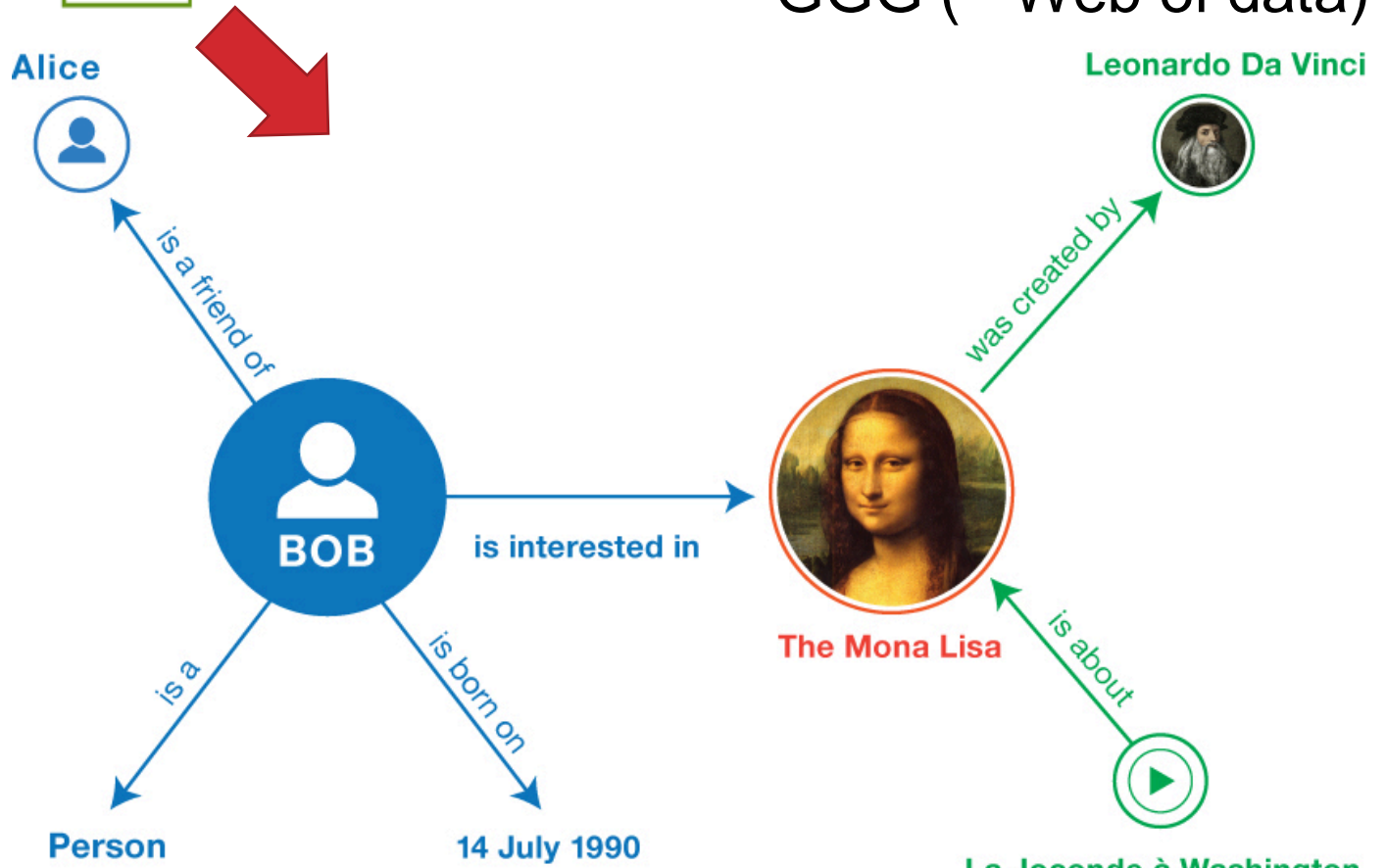
Gigantic Global Graph (GGG)

- graf/mreža entiteta (resursa) i podataka koji ih opisuju
- *"It's not the documents, it is the things they are about which are important"*

WWW (= Web of documents)



GGG (= Web of data)



KARAKTERISTIKE UMREŽENIH PODATAKA NA WEB-U

Web podataka podrazumeva podatke koji imaju:

- jasno definisanu strukturu (= strukturirani podaci), i
- eksplicitno definisano značenje (= semantika podataka je data na način da se može direktno procesirati od strane programa)

UMREŽENI PODACI NA WEB-U

Dva osnovna pojavna oblika ovih podataka:

- podaci umetnuti u Web stranice
- podaci sadržani u bazama podataka i dostupni na Web-u kroz različite oblike programskih interfejsa
 - RESTful APIs (npr. [Knowledge Graph API](#)),
 - SPARQL query endpoints (tema sledećeg časa)

PODACI UMETNUTI U WEB STRANICE

OSNOVNA IDEJA:

OMOGUĆITI PROGRAMIMA RAZUMEVANJE

SADRŽAJA *Web* STRANICA

TIME ŠTO ĆE SE TI SADRŽAJI OPISATI

NA MAŠINSKI RAZUMLJIV NAČIN

(STRUKTURIRANIM I SEMANTIČKI OPISANIM PODACIMA)

PRIMER

Linkovi predstavljeni na isti način, bez obzira na smisao/značenje relacije

'Tradicionalna' Web stranica

```
Jane Doe
Professor
20341 Whitworth Institute
...
Graduated from <a href="http://www.umbc.edu/">UMBC</a>
...
Research associates:
<a href="http://www.xyz.edu/students/alicejones.html">Alice Jones</a>
...
```

Web stranica sa umetnutim strukturiranim podacima

```
<div vocab="http://schema.org/" typeof="Person">
  <span property="name">Jane Doe</span>
  <span property="jobTitle">Professor</span>
  ...
  Graduated from <a href="http://www.umbc.edu/"
    property="alumniOf">UMBC</a>
  ...
  Research associates:
    <a href="http://www.xyz.edu/students/alicejones.html"
      property="colleague">Alice Jones</a>
  ...
</div>
```

Entitetima je definisan tip i opis

<https://schema.org/jobTitle>

Linkovima pridruženo značenje

<https://schema.org/colleague>

PODACI UMETNUTI U WEB STRANICE

Veliki broj Web stranica već sadrži strukturirane podatke sa eksplicitno definisanim značenjem

Na primer:

- filmovi na [RottenTomatoes.com](https://www.RottenTomatoes.com)
- dešavanja na [Ticketmaster.com](https://www.Ticketmaster.com)
- proizvodi na [BestBuy.com](https://www.BestBuy.com)
- recepti na [AllRecipes.com](https://www.AllRecipes.com)

EKSTRAKCIJA PODATAKA IZ WEB STRANICA

Structured data testing tool

- <https://search.google.com/structured-data/testing-tool>
- Omogućuje Web administratorima uvid u podatke koji su dostupni programima koji pristupaju datoj Web stranici
- Ali, ne omogućuje direktan, programski pristup tim podacima, tj. njihovu ekstrakciju iz date Web stranice

EKSTRAKCIJA PODATAKA IZ WEB STRANICA

Microdata Distiller alat

- <http://www.w3.org/2012/pyMicrodata/>
- omogućuje programski pristup podacima umetnutim u Web stranice
- osnovni benefit: jednostavna ekstrakcija podataka iz Web stranice – bez screen scraping-a ili nekih sličnih pristupa – radi korišćenja tih podataka u vašem programu
- može se pozivati kao RESTful servis ili preuzeti i instalirati na lokalnoj mašini

DODAVANJE PODATAKA U WEB STRANICE

Za dodavanje podataka u Web stranice, potrebni su nam:

- RDFS vokabulari koji će omogućiti opisivanje sadržaja Web stranica u mašinski razumljivom formatu
- Način da proširimo HTML jezik tako da mašinski razumljivi opisi podataka budu sastavni deo Web stranice

Da bi odgovorili na 1. zahtev, možemo koristiti Schema.org ili neki drugi RDFS vokabular

Da bi odgovorili na 2. zahtev, možemo koristiti RDFa, Microdata ili JSON-LD – W3C preporuke za proširenje HTML jezika mašinski razumljivim opisima podataka

SCHEMA.ORG

Vokabular za opisivanje podataka u mašinski razumljivom obliku; trenutno, najzastupljeniji vokabular na Web-u

Inicijativa potekla od velikih Web kompanija: Google, Yahoo, Microsoft (Bing), Yandex

Dalje se razvija kao community effort u okviru Web konzorcijuma:

<https://www.w3.org/community/schemaorg/>

Inicijalno omogućavao opis malog broja osnovnih tipova sadržaja, vremenom se taj broj značajno uvećao

- Lista svih tipova koje Schema.org trenutno podržava:

<http://schema.org/docs/full.html>

SCHEMA.ORG

Preporuke:

- Pogledati keynote R. Guha-e – lidera [W3C WebSchemas](http://www.w3.org/2001/sw/) grupe – na temu strukturiranih podataka na Web-u, Schema.org, kao i razvoja i značaja otvorenih tehnologija za Web podataka:
http://videlectures.net/iswc2013_guha_tunnel/
- Takođe, interesantan i koristan može biti i članak “Schema.org: Evolution of Structured Data on the Web”:
<http://queue.acm.org/detail.cfm?id=2857276>

PROŠIRENJE HTML-A MAŠINSKI RAZUMLJIVIM OPISIMA PODATAKA

W3C preporuke (de-facto standardi) za dodavanje strukturiranih podataka u HTML stranice:

- RDFa
- Microdata
- JSON-LD

RDFA

Jane Doe Professor

20341 Whitworth Institute 405 N. Whitworth Seattle, WA 98052

(425) 123-4567

jane-doe@xyz.edu

Research associates: [Alice Jones](#) [Bob Smith](#)



Prikaz u browser-u

HTML source

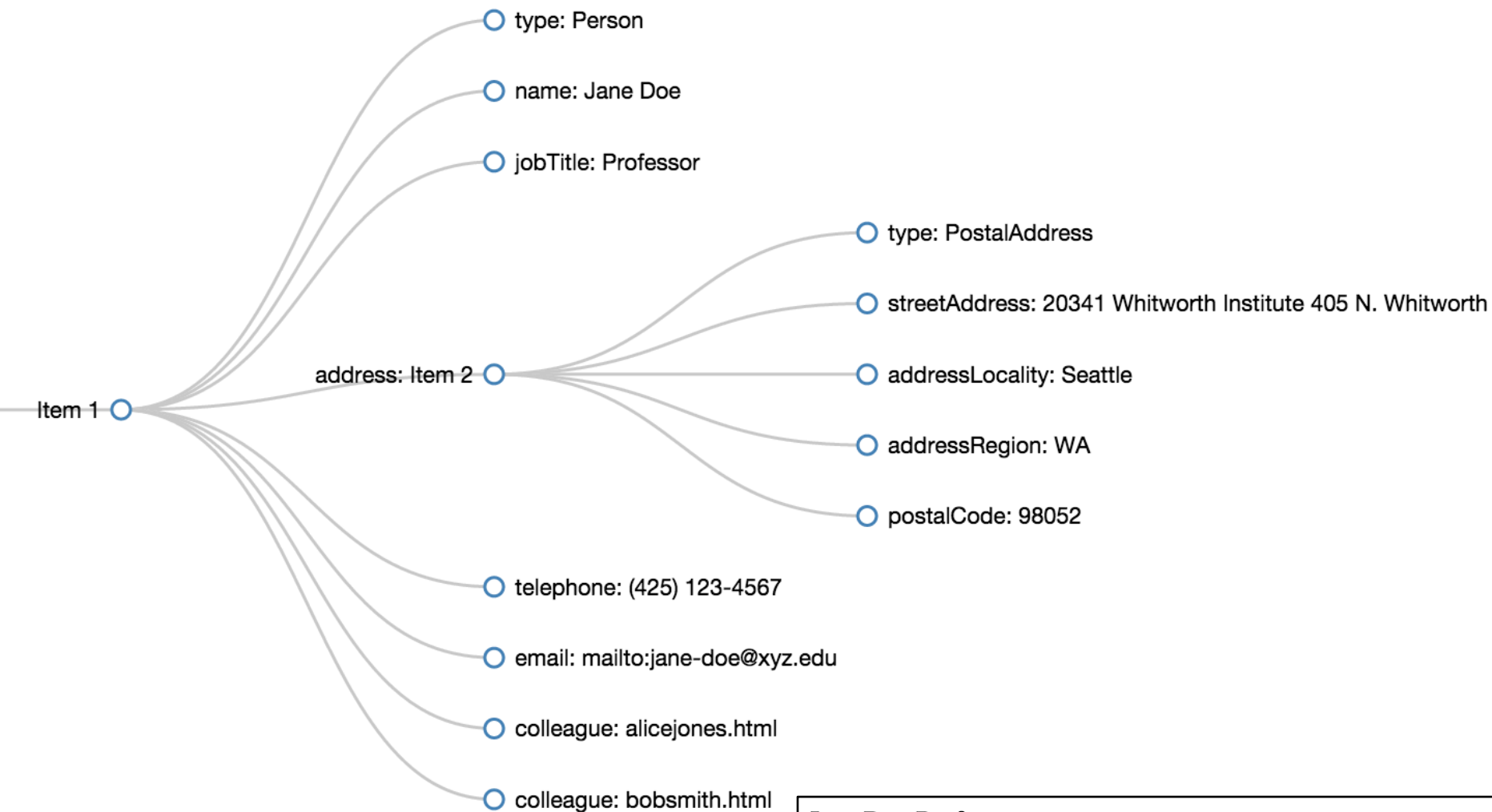
(sa umetnutim

RDFA podacima)



```
<div vocab="http://schema.org/" typeof="Person">
  <span property="name">Jane Doe</span>
  <span property="jobTitle">Professor</span>
  <div property="address" typeof="PostalAddress">
    <span property="streetAddress">
      20341 Whitworth Institute
      405 N. Whitworth
    </span>
    <span property="addressLocality">Seattle</span>,
    <span property="addressRegion">WA</span>
    <span property="postalCode">98052</span>
  </div>
  <span property="telephone">(425) 123-4567</span><br/>
  <a href="mailto:jane-doe@xyz.edu" property="email">
    jane-doe@xyz.edu</a><br/>
  Research associates:
  <a href="http://www.xyz.edu/students/alicejones.html" property="colleag
    Alice Jones</a>
  <a href="http://www.xyz.edu/students/bobsmith.html" property="colleagu
    Bob Smith</a>
</div>
```

GRAF PODATAKA IZ PRETHODNOG PRIMERA



Jane Doe Professor
20341 Whitworth Institute 405 N. Whitworth Seattle, WA 98052
(425) 123-4567
jane-doe@xyz.edu
Research associates: [Alice Jones](#) [Bob Smith](#)

MICRODATA

```
<div itemscope itemtype="http://schema.org/Person">
  <span itemprop="name">Jane Doe</span>
  <span itemprop="jobTitle">Professor</span>
  <div itemprop="address"
    itemscope itemtype="http://schema.org/PostalAddress">
    <span itemprop="streetAddress">
      20341 Whitworth Institute
      405 N. Whitworth
    </span>
    <span itemprop="addressLocality">Seattle</span>,
    <span itemprop="addressRegion">WA</span>
    <span itemprop="postalCode">98052</span>
  </div>
  <span itemprop="telephone">(425) 123-4567</span><br>
  <a href="mailto:jane-doe@xyz.edu" itemprop="email">
    jane-doe@xyz.edu</a><br>
  Research associates:
  <a href="http://www.xyz.edu/students/alicejones.html"
    itemprop="colleague">Alice Jones</a>
  <a href="http://www.xyz.edu/students/bobsmith.html"
    itemprop="colleague">Bob Smith</a>
</div>
```

Isti primer u Microdata notaciji; usuštini, sve je isto, razlika je samo u nazivima HTML atributa

JSON-LD

```
<script type="application/ld+json">
{
  "@context": "http://schema.org",
  "@type": "Person",
  "address": {
    "@type": "PostalAddress",
    "addressLocality": "Seattle",
    "addressRegion": "WA",
    "postalCode": "98052",
    "streetAddress": "20341 Whitworth Institute 405 N. Whitworth"
  },
  "colleague": [
    "http://www.xyz.edu/students/alicejones.html",
    "http://www.xyz.edu/students/bobsmith.html"
  ],
  "email": "mailto:jane-doe@xyz.edu",
  "image": "janedoe.jpg",
  "jobTitle": "Professor",
  "name": "Jane Doe",
  "telephone": "(425) 123-4567",
  "url": "http://www.janedoe.com"
}
</script>
```

Za razliku od RDFa i Microdata, JSON-LD podaci se ne umeću u deo HTML stranice koji se prikazuje korisniku (body), već je deo sekcije namenjene za programe (head)

RDFa, MICRODATA, JSON-LD

RDFa:

- Specifikacija: <http://www.w3.org/TR/xhtml-rdfa-primer/>
- Relevantne informacije, primeri, dodatni materijali: <http://rdfa.info/>

Microdata:

- Specifikacija: <http://dev.w3.org/html5/md/>

JSON-LD:

- Specifikacija: <http://www.w3.org/TR/json-ld/>
- Relevantne informacije, primeri, dodatni materijali: <http://json-ld.org/>

JOŠ NEKI RELEVANTNI VOKABULARI: SCHEMA ACTIONS

- Skup Schema.org klasa i atributa namenjenih
 - (maš. čitljivom) opisu akcija koje neki Web sajt ili aplikacija omogućuje svojim korisnicima, i kako se te akcije mogu programski inicirati
 - integrisanju podataka o akcijama korisnika na različitim sajtovima na Web-u
- Za više informacija, pogledati:
 - članak (<http://goo.gl/9zkeUK>) koji objašnjava značaj Schema Actions, kao i članak koji ilustruje primenu u domenu muzike (<http://goo.gl/xPRpQz>)
 - opis Schema Actions i primere načina korišćenja: <https://schema.org/docs/actions.html>

JOŠ NEKI RELEVANTNI VOKABULARI: GOODRELATIONS

- <http://www.heppnetz.de/projects/goodrelations/>
- Vokabular za opisivanje proizvoda, ponuda, prodavnica i sl.
- Ima široku primenu u domenu elektronske trgovine
 - Npr. Kmart.com, Sears.com, BestBuy.com
- Razvijeni su brojni alati koji omogućuju jednostavno opisivanje podataka ovim vokabularom
 - pogledati: <http://wiki.goodrelations-vocabulary.org/Tools>
- Ovaj vokabular je takođe sastavni deo Schema.org
 - <http://schema.org/Product> ; <http://schema.org/Offer> ...

JOŠ NEKI RELEVANTNI VOKABULARI: OPEN GRAPH PROTOCOL (OGP)

- <http://ogp.me/>
- Vokabular koji je uveo Facebook kako bi omogućio prikupljanje dodatnih informacija o sadržajima koje korisnici Like-uju na Web-u
 - OGP vokabular u kombinaciji sa RDFa standardom za proširenje HTML-a, obezbeđuje eksplicitnu semantiku Like-ovanih sadržaja
 - Tako prikupljene informacije Facebook koristi kao input za dalji razvoj svog Entity Graph-a
- OGP omogućuje opis različitih tipova sadržaja popularnih među korisnicima Web-a, kao što su muzika, knjige, video zapisi, profili korisnika i sl

ALATI ZA RAD SA PODACIMA NA WEB-U

Google je razvio više alata namenjenih

- dodavanju strukturiranih podataka u Web stranice
- praćenju korišćenja stranica sa umetnutim podacima,
- detektovanju grešaka u podacima

Ti alati su:

- Structured Data Dashboard (<https://goo.gl/V8NZ8L>)
- Data Highlighter (<https://goo.gl/P5SZOc>)
- Structured Data Markup Helper (<https://goo.gl/1Ywtfq>)

Video sa Google IO 2013 konferencije opisuje ove alate i objašnjava njihovu namenu i korišćenje:

<https://developers.google.com/events/io/sessions/351340935>

ALATI ZA RAD SA PODACIMA NA WEB-U

Popularne Web platforme koje podržavaju RDFa/Microdata

- Drupal
 - podrška za RDFa je deo Drupal-ovog core modula (od v.7);
- Webnodes
 - obezbeđuju punu podršku za rad sa Microdata i Schema.org (pogledati [ovaj članak](#))
- WordPress
 - Obezbeđuje više proširenja za rad sa RDFa, Microdata, Schema.org (pogledati, npr., [ovu listu](#))

PRINCIPI UMREŽAVANJA (LINKOVANJA) PODATAKA NA WEB-U

5 STAR LINKED OPEN DATA



Linked Open Data star scheme by example:

<http://5stardata.info/>

OSNOVNI PRINCIPI LINKOVANJA PODATAKA NA WEB-U

1) Koristiti URI za jedinstvenu identifikaciju entiteta/objekata/pojava/...

ISBN: 9781775411840

2) Koristiti HTTP URI tako da se informacije o entitetima učine dostupnim posredstvom Web-a

```
<http://www.worldcat.org/title/north-and-south/oclc/606818482>
```

3) Opisati entitete korisnim podacima primenom RDF modela; u te svrhe, preporučuje se korišćenje postojećih RDF vokabulara

```
<http://www.worldcat.org/title/north-and-south/oclc/606818482>  
rdf:type schema:Book
```

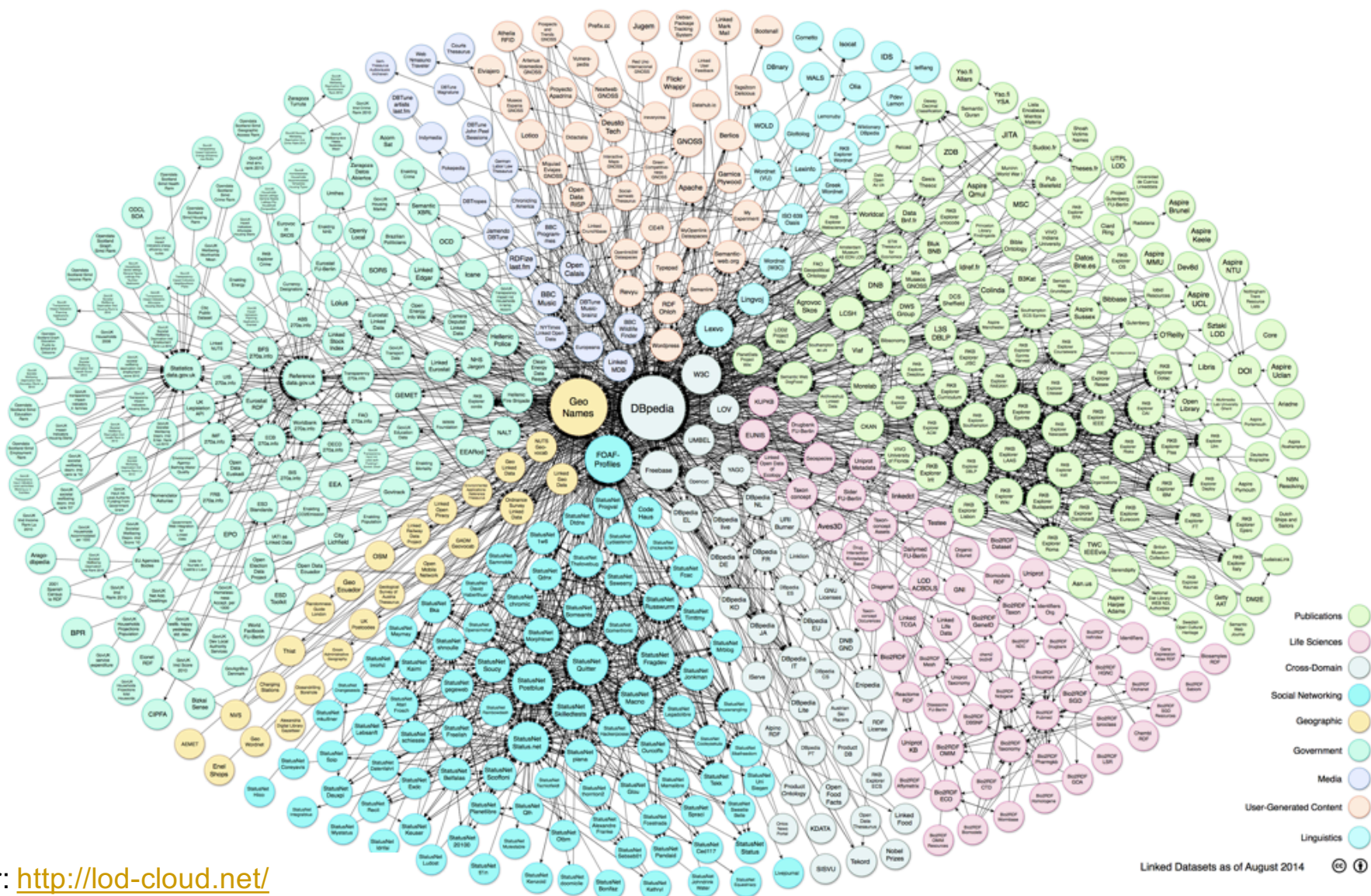
4) Uspostaviti imenovane linkove ka drugim entitetima/objektima/pojavama...

```
<http://www.worldcat.org/title/north-and-south/oclc/606818482>  
schema:author <http://viaf.org/viaf/39377536/>
```



URI jedinstveno identifikuje spisateljicu Elizabeth Gaskell

PODACI PUBLIKOVANI U SKLADU SA PRINCIPIMA LINKOVANJA PODATAKA NA WEB-U



PRIMERI PRIMENE

RICH SNIPPETS

- Bogatiji prikaz rezultata pretrage na Google-u za stranice koje sadrže strukturirane i semantički opisane podatke
- Npr., potražite na Google.com bilo koji film, koncert, aplikaciju za mobilni uređaj, projekat sa Sourceforge-a, ...
- Detaljan prikaz Rich Snippets-a je rasoloživ na <https://goo.gl/6JBY9k>

Boyhood (2014) - IMDb

www.imdb.com/title/tt1065073/ ▼

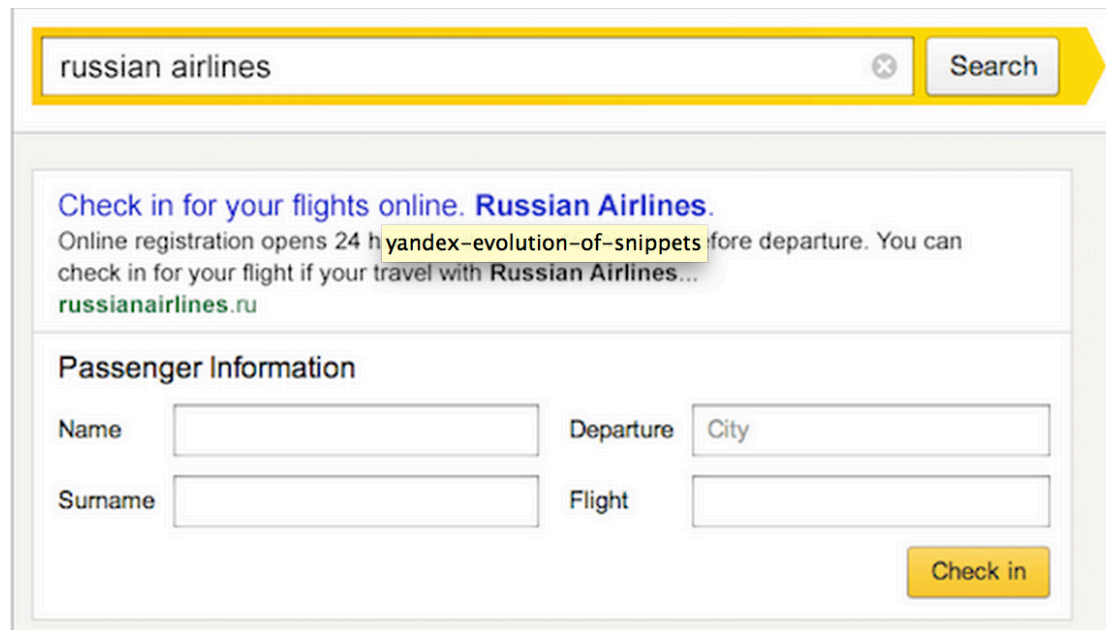
★★★★★ Rating: 8/10 - 217,851 votes

Boyhood -- Clip: Talk To Me **Boyhood** -- Featurette: Behind the Scenes The actress that portrays Samantha is the **movie's** director's daughter, Lorelei Linklater ...

[Full Cast & Crew](#) - [Ellar Coltrane](#) - [Lorelei Linklater](#) - [Patricia Arquette](#)

INTERACTIVE SNIPPETS

- Funkcionalniji rezultati pretrage Yandex pretraživača, zasnovani na strukturiranim i semantički opisanim podacima
 - Npr., moguće je odmah uraditi check-in na avio letu
- Za više informacija pogledati: <http://goo.gl/uxjb6g>



The image shows a search interface for 'russian airlines'. The search bar contains the text 'russian airlines' and a 'Search' button. Below the search bar, there is a snippet for 'Check in for your flights online. Russian Airlines.' The snippet text includes 'Online registration opens 24 h yandex-evolution-of-snippets before departure. You can check in for your flight if your travel with Russian Airlines...' and the URL 'russianairlines.ru'. Below the snippet, there is a form titled 'Passenger Information' with input fields for 'Name', 'Surname', 'Departure City', and 'Flight'. A 'Check in' button is located at the bottom right of the form.

Search bar: russian airlines [x] Search

Check in for your flights online. **Russian Airlines.**
Online registration opens 24 h yandex-evolution-of-snippets before departure. You can check in for your flight if your travel with Russian Airlines...
russianairlines.ru

Passenger Information

Name Departure

Surname Flight

Check in

PINTEREST'S RICH PINS

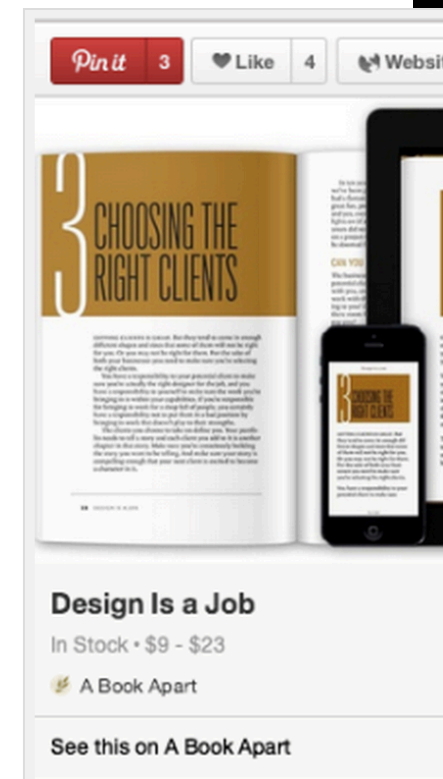
- Reč je o pinovima (Pins) sa dodatnim mogućnostima
 - Npr., za proizvode obezbeđuju informacije o trenutnoj ceni, raspoloživosti, aktuelnim popustima

- Pregled različitih tipova Rich Pins:

<https://business.pinterest.com/en/rich-pins>

- Dokumentacija namenjena developerima detaljno opisuje kako se strukturirani podaci koriste za generisanje Rich Pins:

<https://developers.pinterest.com/docs/rich-pins/>




PERSONALNI DIGITALNI ASISTENTI

Neki poznatiji primeri:

- Siri (<http://www.apple.com/ios/siri/>)
- Google Now (<https://www.google.com/search/about/>)
- Cortana (<http://www.microsoft.com/en/mobile/experiences/cortana/>)
- Evi (<http://www.evi.com/>)
- Skyvi (<http://www.skyviapp.com/>)
 - kupljen od strane Google-a; od Sept 2015 prestaje da postoji kao posebna aplikacija; tehnologija integrisana u Google Now

GOOGLE NOW


“Google Now provides updates to restaurant and hotel reservations or flight information received in Gmail. By marking up email notifications to your users, you can use Google Now to bring them similar updates about your services and products”


Virgin America
flight 25 


Status: Delayed / Mon, 29 Oct 2012


Depart New York
JFK 12:30 PM (12:00 PM)
Terminal 2, Gate 54B

Arrive San Francisco
SFO 1:56 PM (1:31 PM)
Terminal 4

 [Navigate to JFK / 32 min](#)


 [View email](#)

Boarding pass: UA 1051 
Sequence: 63, Premier access




Passenger
John Smith


Terminal	Gate	Seat	Group
3	1	31B	6


 [View email](#)

The Connaught Hotel
Carlos Place, Mayfair, London
W1K 2AL, United Kingdom

Check-out in 1 hour

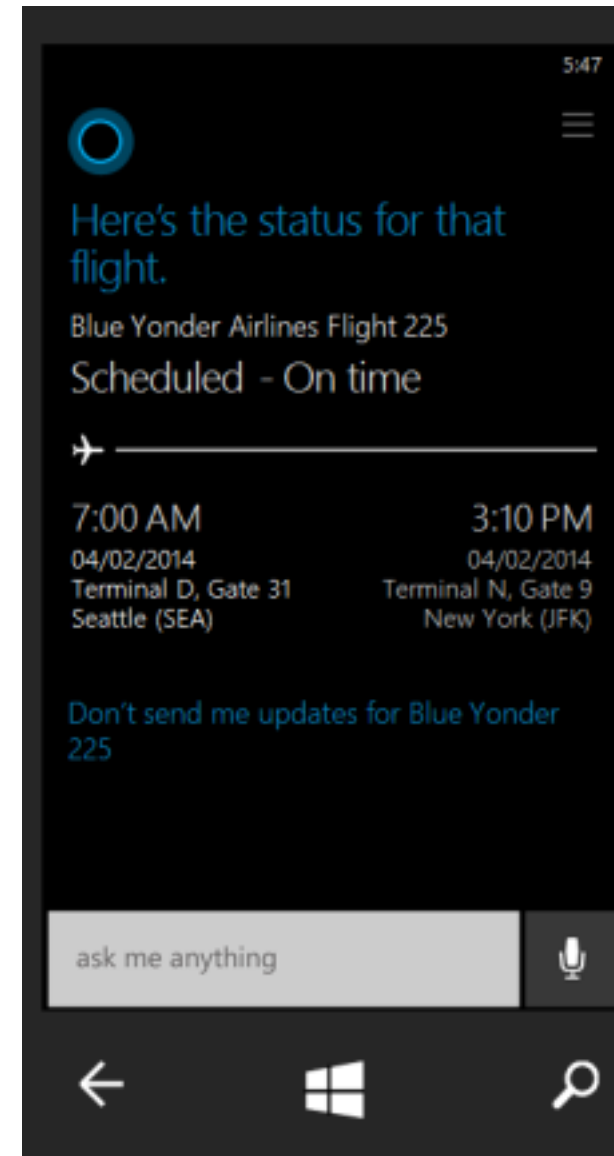


 [Get directions](#)

 [View email](#)

CORTANA

“Airline flight providers can add schema.org markup to their outgoing emails to enable flight tracking for customers that use Microsoft Cortana”



WEB (RESTFUL) APIS ZA PRISTUP PODACIMA

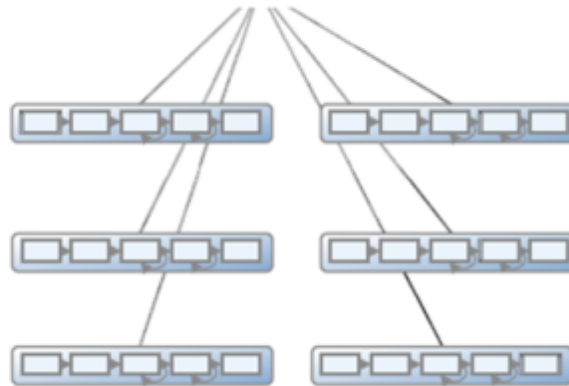
Web Services Directory

All Viewing 1 to 1377 of 1377 APIs

API	Description
Google Maps	Mapping services
Flickr	Photo sharing service
YouTube	Video sharing and search
Amazon eCommerce	Online retailer
Twitter	Microblogging service
eBay	Online auction marketplace
Microsoft Virtual Earth	Mapping services
del.icio.us	Social bookmarking
Google Search	Search services
Yahoo Maps	Mapping services
Yelp	Local user reviews and city guides
hostip.info	IP lookup
Netvibes	Personalized home page with widgets
PayPal	Online payments
Rhapsody	Online music services
WeatherBug	Weather forecast services

programmableweb **Izvori podataka:
Web APIs**

Siri



**Domenski modeli
i programska
logika**



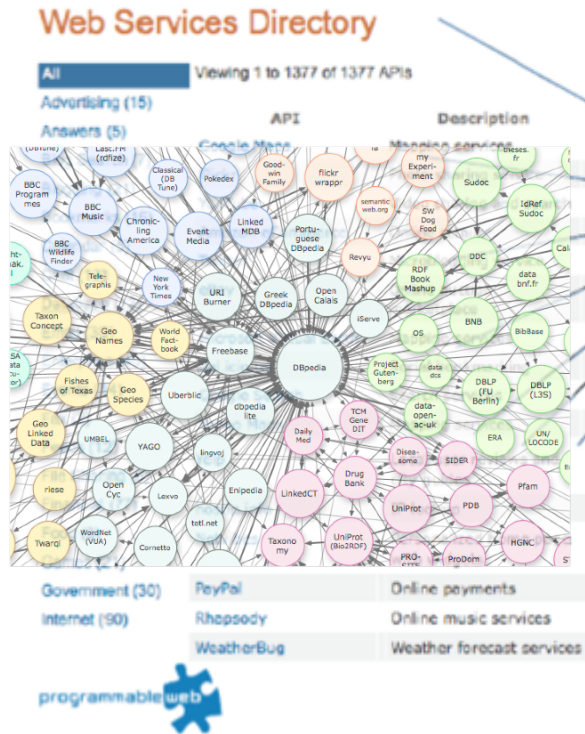
**Interfejs
prema
korisniku
(dijalog)**

WEB (RESTFUL) APIs ZA PRISTUP PODACIMA

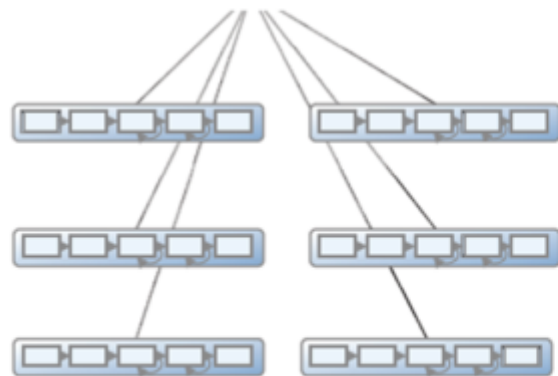
Nedostaci / poteškoće ovog pristupa:

- Potreba za upoznavanjem sa specifičnostima svakog novog API-a radi pristupa podacima koje obezbeđuje
- Potreba za usklađivanjem heterogenih formata i značenja podataka prikupljenih iz različitih (Web) izvora
- Potreba za kontinuiranim ažuriranjem koda u skladu sa izmenama Web APIs
- Promena uslova pod kojima su podaci dostupni
 - promene količine i/ili vrste podataka kojima se može pristupiti posredstvom API-a

NOVI IZVOR PODADATA: UMREŽENI (OTVORENI) PODACI NA WEB-U



Siri



Interfejs
prema
korisniku
(dijalog)

Domenski modeli
(**RDFS vokabulari**) i
programska logika

Izvori podataka:
Web APIs + **Linked Data**