Umrežavanje podataka na Web-u

Jelena Jovanovic

Email: jeljov@gmail.com

Web: http://jelenajovanovic.net

Kratko podsećanje:
Gigantski globalni graf i umreženi podaci na Web-u (Web podataka)

GIGANTSKI GLOBALNI GRAF (1)

Faza 1: ***I**nternational **I**nformation **I**nfrastructure (III)*

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

Faza 2: ***W**orld **W**ide **W**eb (WWW)*

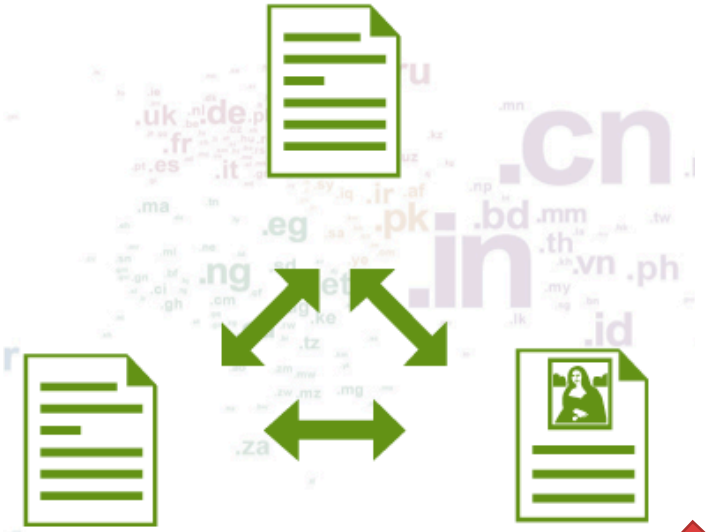
- graf/mreža dokumenata poznata kao *Web*
- "*It isn't the computers, but the documents which are interesting*"

Faza 3: ***G**igantic **G**lobal **G**raph (GGG)*

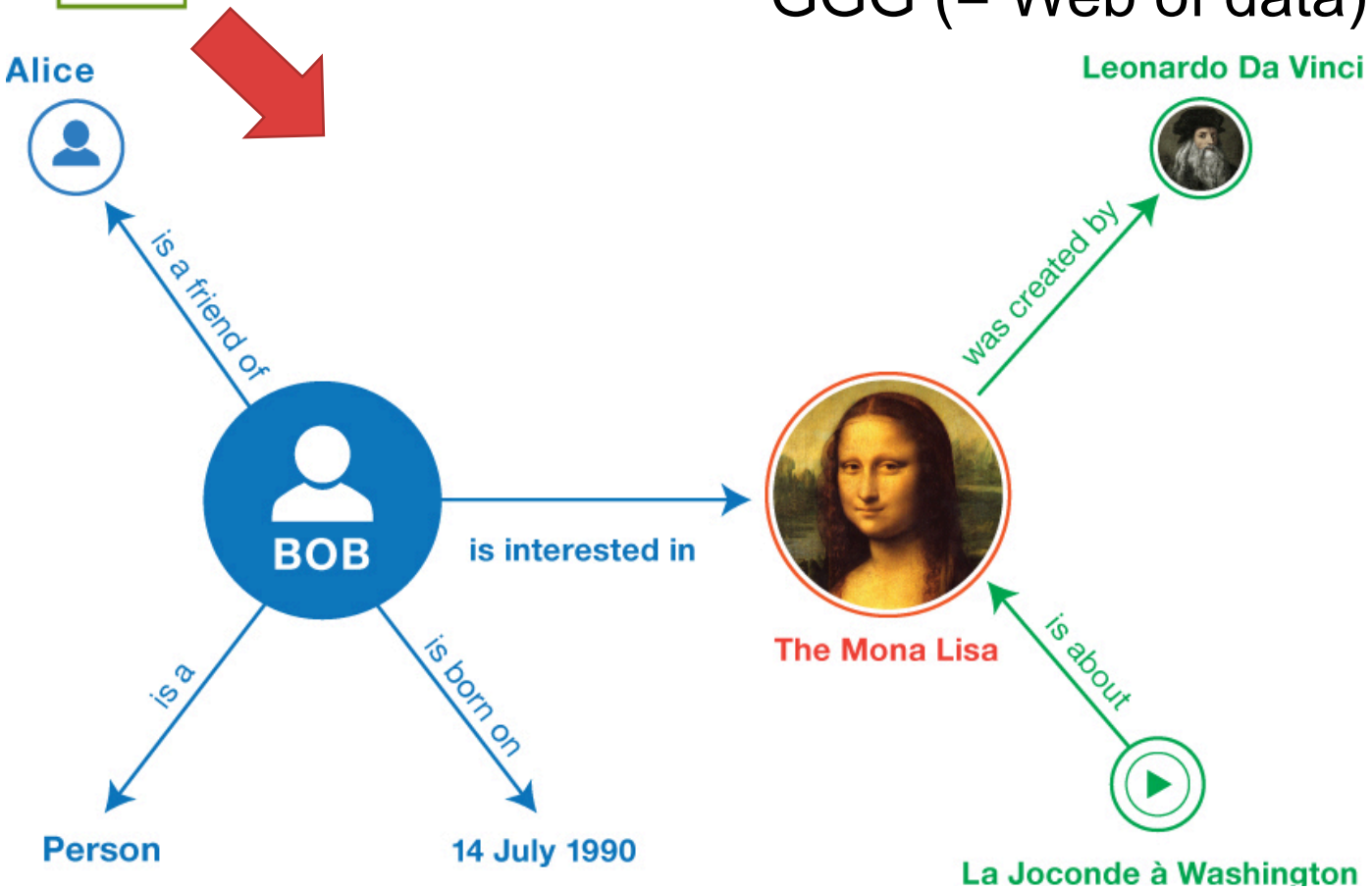
- graf/mreža entiteta (resursa) i podataka koji ih opisuju
- "*It's not the documents, it is the things they are about which are important*"

TBL's blog post on GGG: http://dig.csail.mit.edu/breadcrumbs/node/215

WWW (= Web of documents)



GGG (= Web of data)



Izvor slike: http://www.w3.org/TR/2014/NOTE-rdf11-primer-20140225/

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)

Dva osnovna pojavna oblika ovakvih podataka:

- podaci umetnuti u Web stranice
- podaci sadržani u bazama podataka i dostupni na Web-u kroz različite oblike programskih interfejsa (RESTful APIs, query endpoints)

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

Entitetima je definisan tip i opis

```
<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>
```

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
- dešavanja na Ticketmaster.com
- proizvodi na BestBuy.com
- recepti na AllRecipes.com

Ekstrakcija podataka iz Web stranica

Structured data testing tool

- https://developers.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 informacija 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 osnovih 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 grupe – na temu strukturiranih podataka na Web-u, Schema.org, kao i razvoja i značaja otvorenih tehnologija za Web podataka: http://videolectures.net/iswc2013_guha_tunnel/
- Takođe, interesantan i koristan može biti i nedavno objavljen članak (Dec 2015) “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

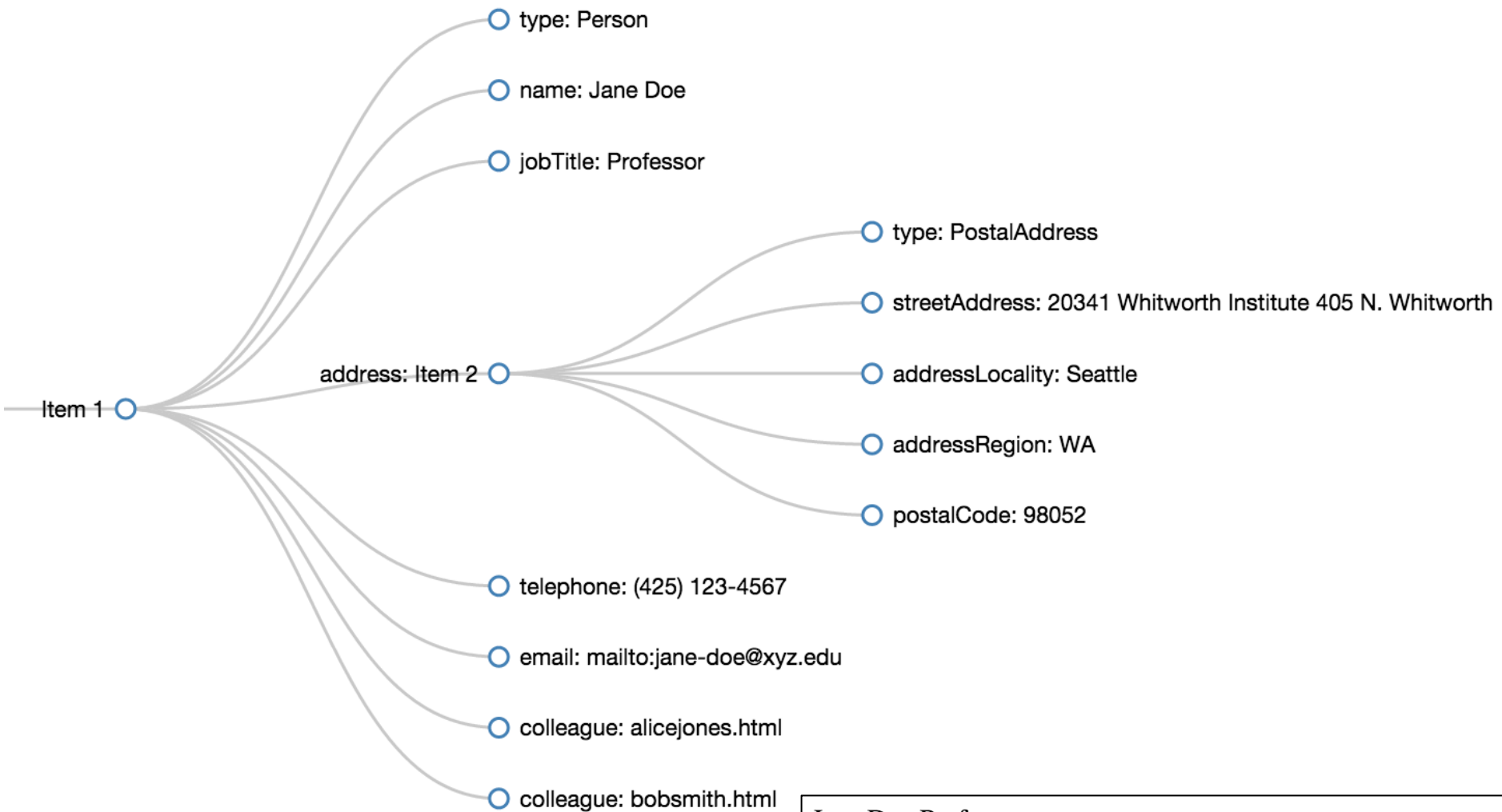
Prikaz u browser-u

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="colleague
    Bob Smith</a>
</div>
```

Primer kreiran korišćenjem http://rdfa.info/play/ editora

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;
u suštini, 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 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)
  - dokument koji opisuje Schema Actions i obezbeđuje instrukcije za njihovo korišćenje (https://goo.gl/D7oxrw)

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 stanica 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/1Ywtfg)

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>
```

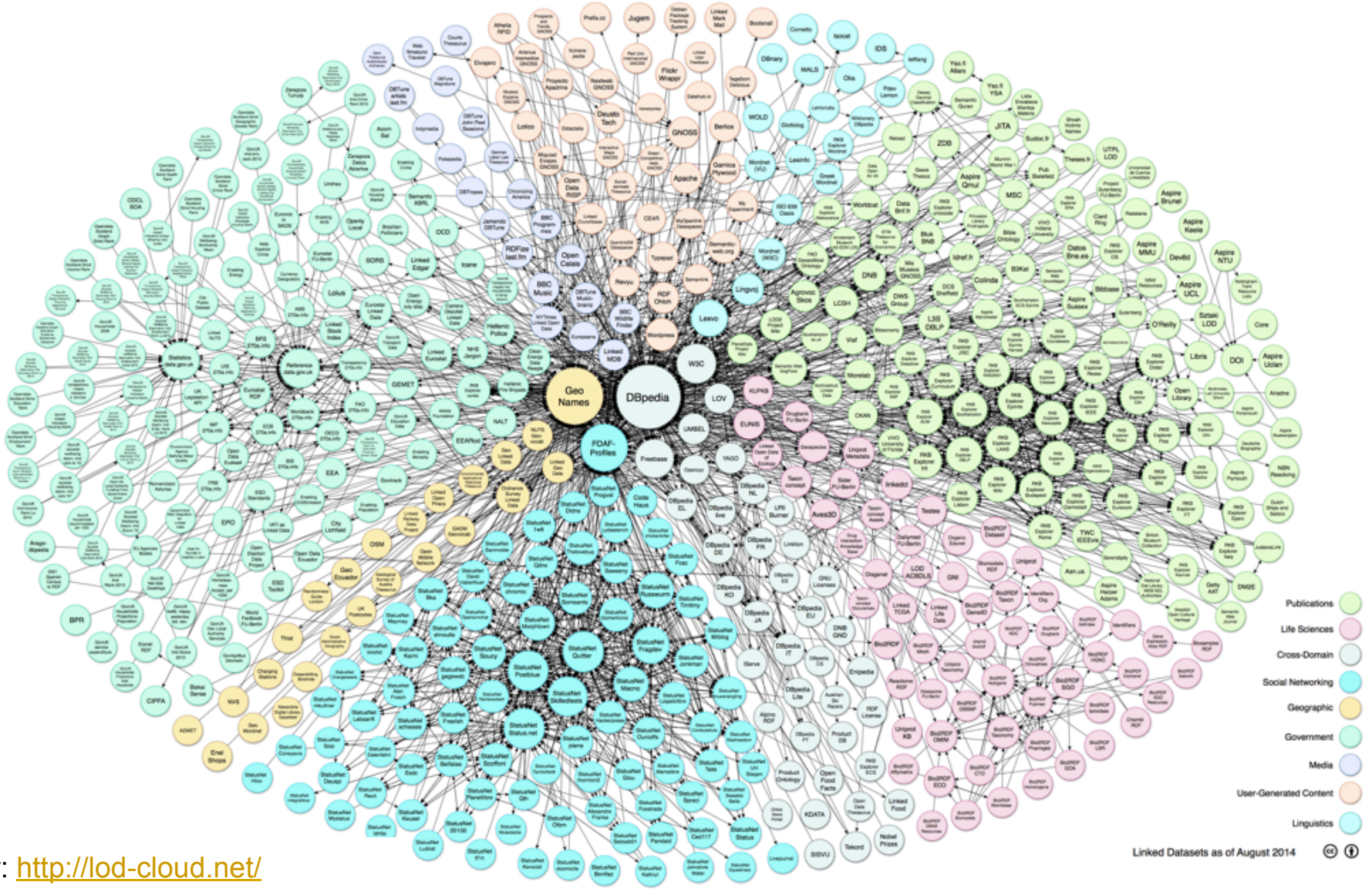
3) Opisati entitete korisnim podacima primenom RDF modela; u te svrhe, preporučuje se korišćenje postojećih RDF vokabulara

```
<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/>
```

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

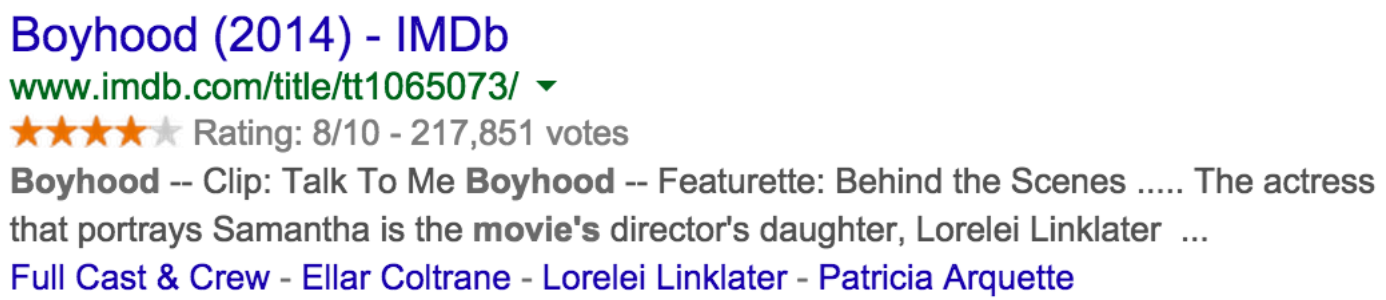


Izvor: http://lod-cloud.net/

PRIMERI PRIMENE

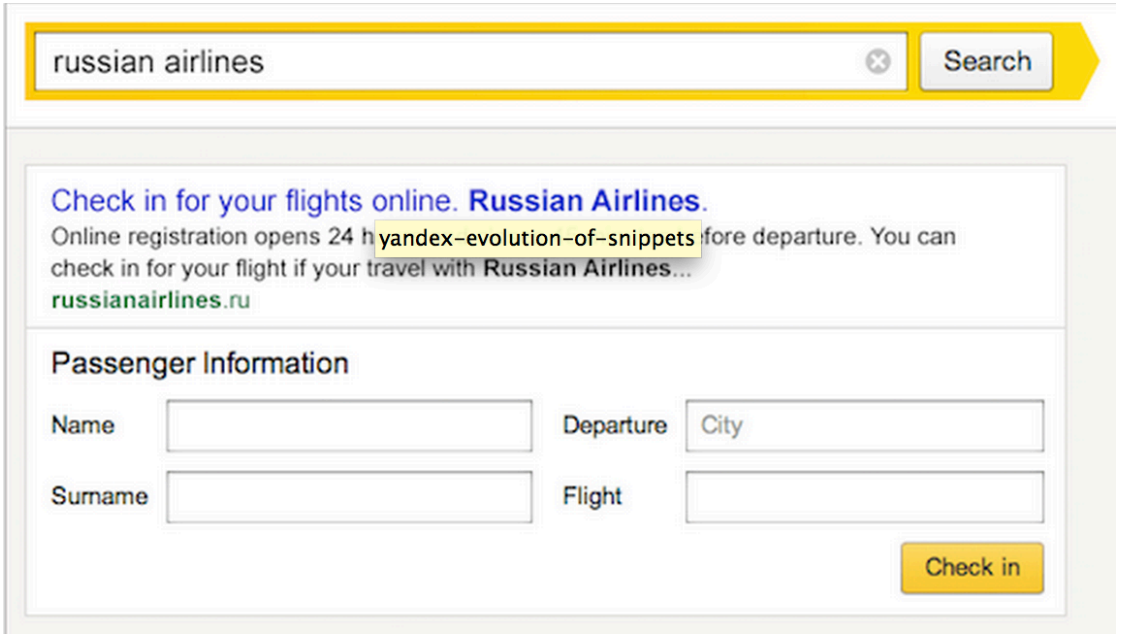
Rich Snippets

- 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



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



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

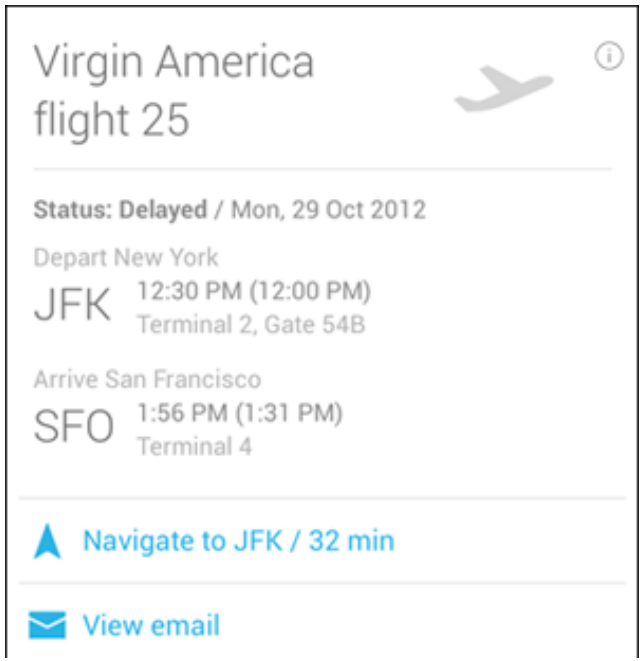
Neki poznatiji primeri:

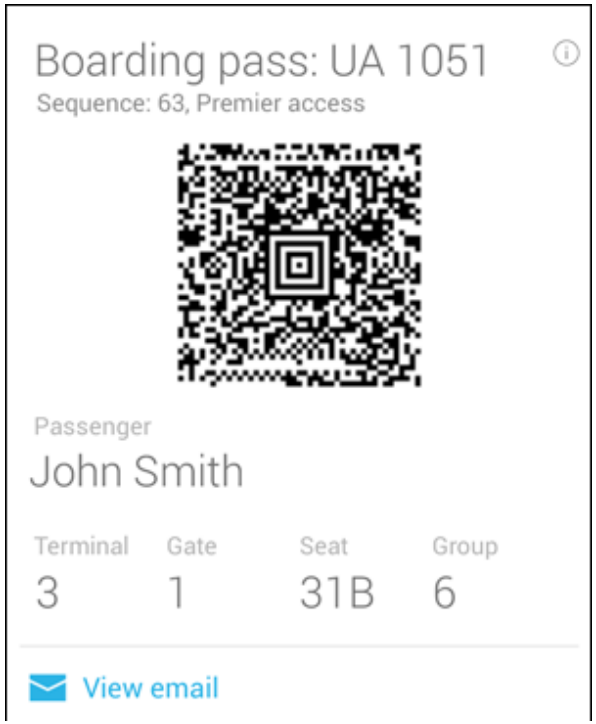
- Siri (http://www.apple.com/ios/siri/)
- Google Now (http://www.google.com/landing/now/)
- Cortana (http://www.microsoft.com/en/mobile/experiences/cortana/)
- Evi (http://www.evi.com/)
- Skyvi (http://www.skyviapp.com/)
  - kupljen od strane Google-a; od Sept 2015 prestaje da postoji kao posebna apikacija; tehnologija integraisana u Google Now

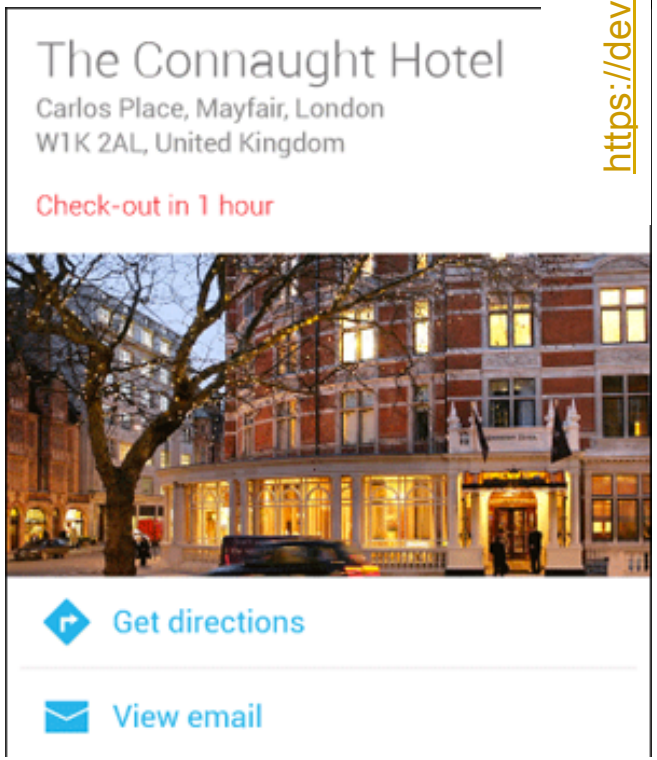
GOOGLE NOW

"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"

Izvor: https://developers.google.com/schemas/now/cards

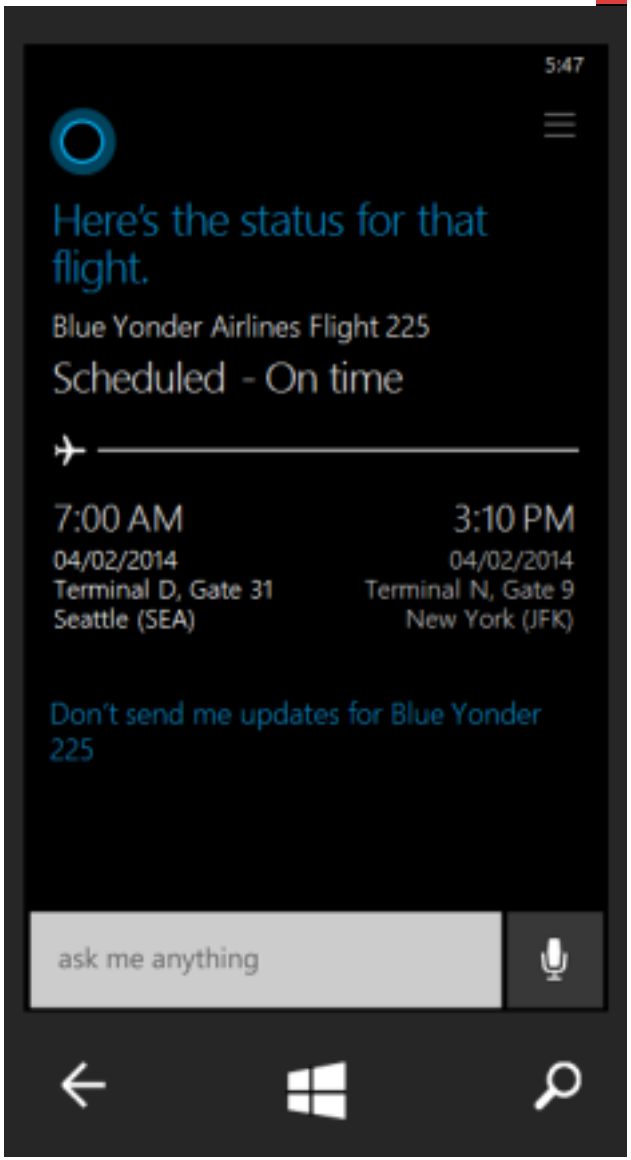






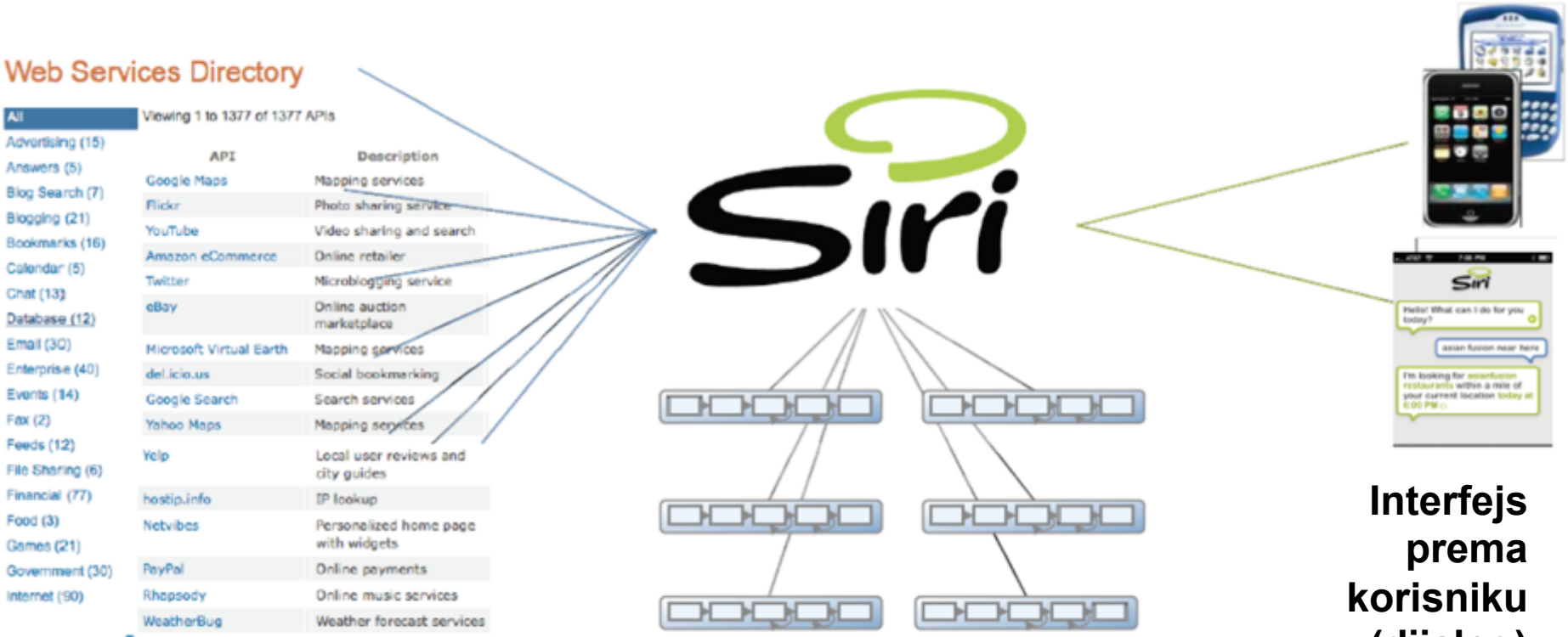
CORTANA

"*Airline flight providers can add schema.org markup to their outgoing mails to enable flight tracking for customers that use Microsoft Cortana*"



Izvor: https://msdn.microsoft.com/en-us/library/dn632190.aspx

WEB (RESTFUL) APIS ZA PRISTUP PODACIMA



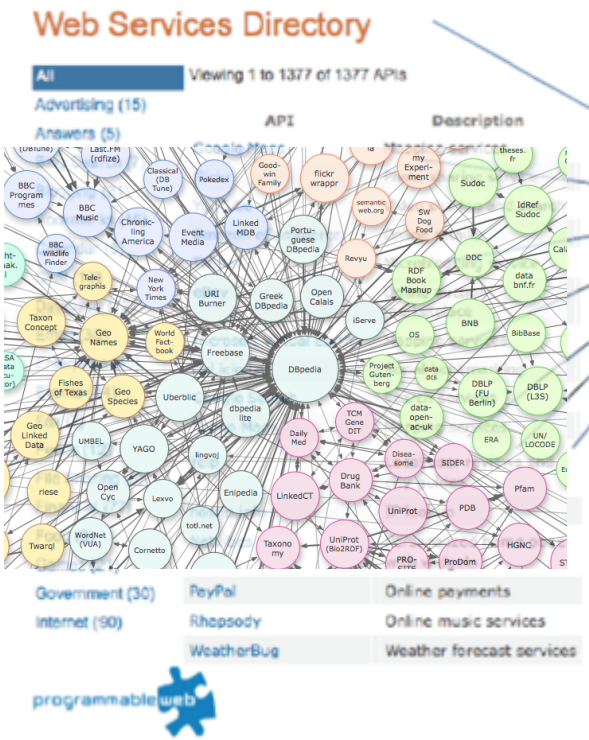
Tom Gruber. *Siri: A Virtual Personal Assistant.* Keynote presentation at Semantic Technologies conference, June 2009. http://tomgruber.org/writing/semtech09.htm

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



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

Siri

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

Interfejs
prema
korisniku
(dijalog)