

# MODELOVANJE ZNANJA U FORMI SEMANTIČKE MREŽE/GRAFA

**JELENA JOVANOVIĆ**

Email: [jeljov@gmail.com](mailto:jeljov@gmail.com)

Web: <http://jelenajovanovic.net>

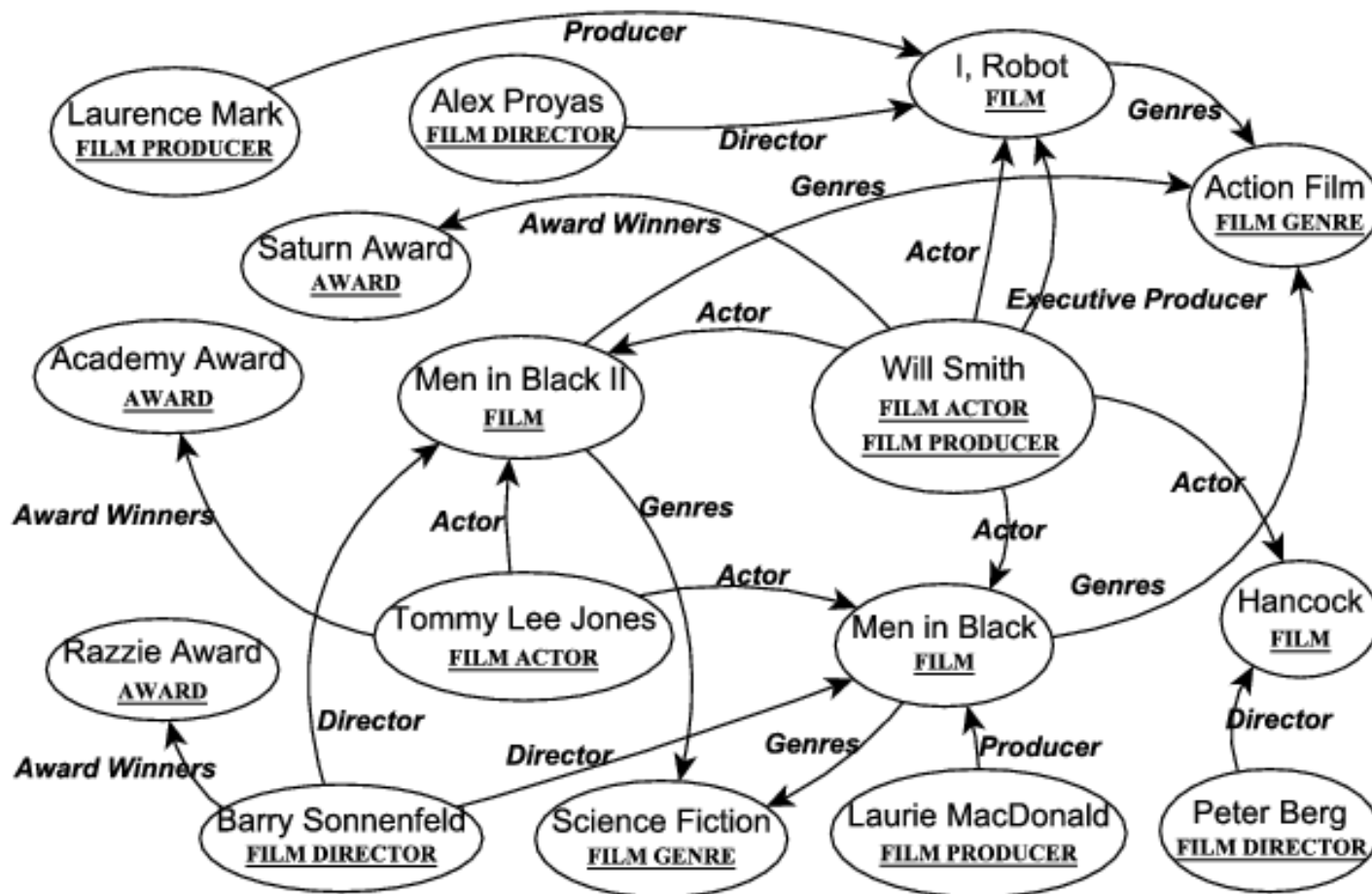


# PREGLED PREDAVANJA

- (Semantički) grafovi/mreže za modelovanje znanja
- Grafovi podataka i znanja u komercijalnoj sferi
- Otvoreni, javno dostupni grafovi podataka i znanja
- Gigantski globalni graf
  - Vizija Web-a kao gigantske globalne baze podataka i znanja
  - Kreiranje gigantskih baza znanja kroz automatsko prikupljanje podataka i činjenica sa Web-a

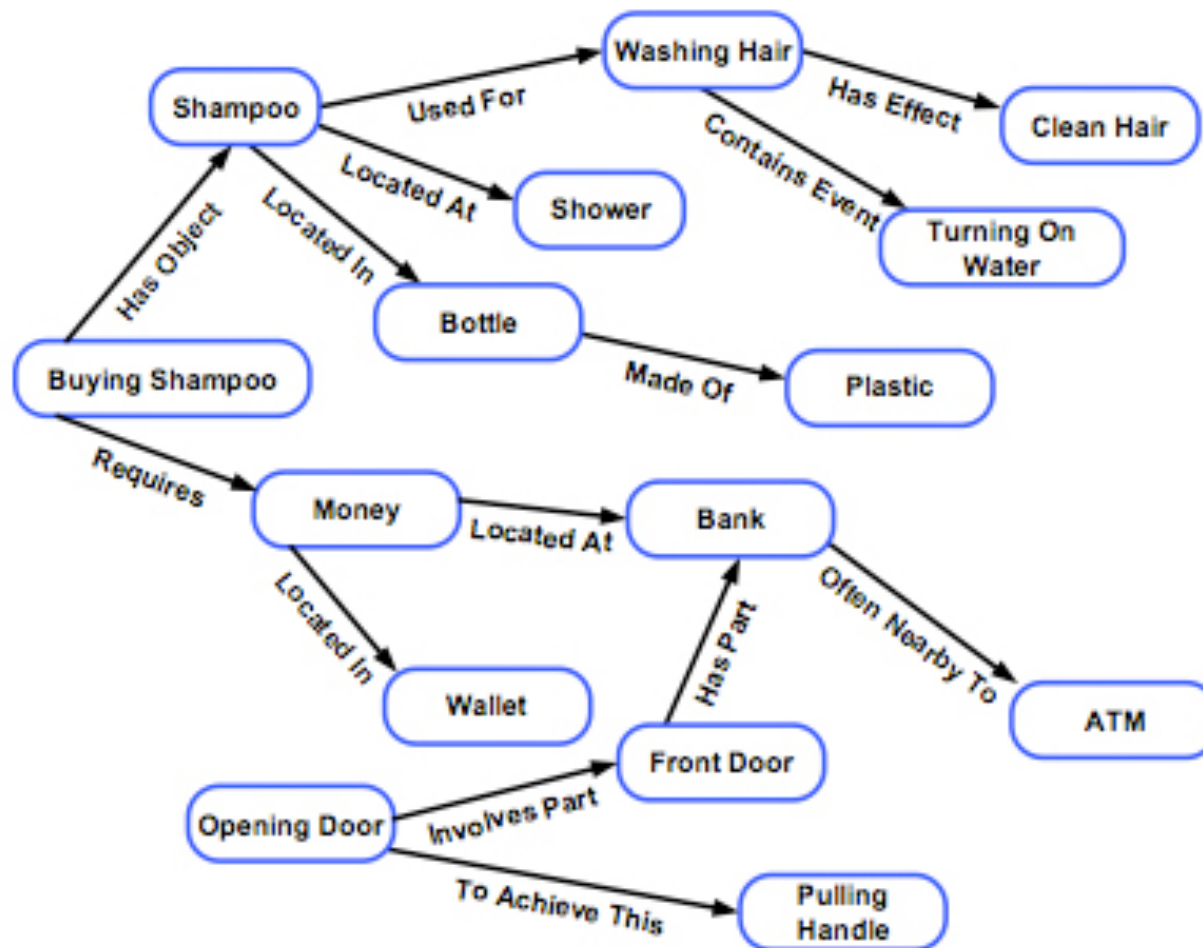
# (SEMANTIČKI) GRAFOVI/MREŽE ZA PREDSTAVLJANJE ZNANJA

# PRIMER GRAFA ENTITETA I NJIHOVIH RELACIJA



Primer ilustruje mali segment [Freebase](http://freebase.org) baze znanja

# PRIMER GRAFA ZDRAVORAZUMSKOG ZNANJA (COMMONSENSE KNOWLEDGE)

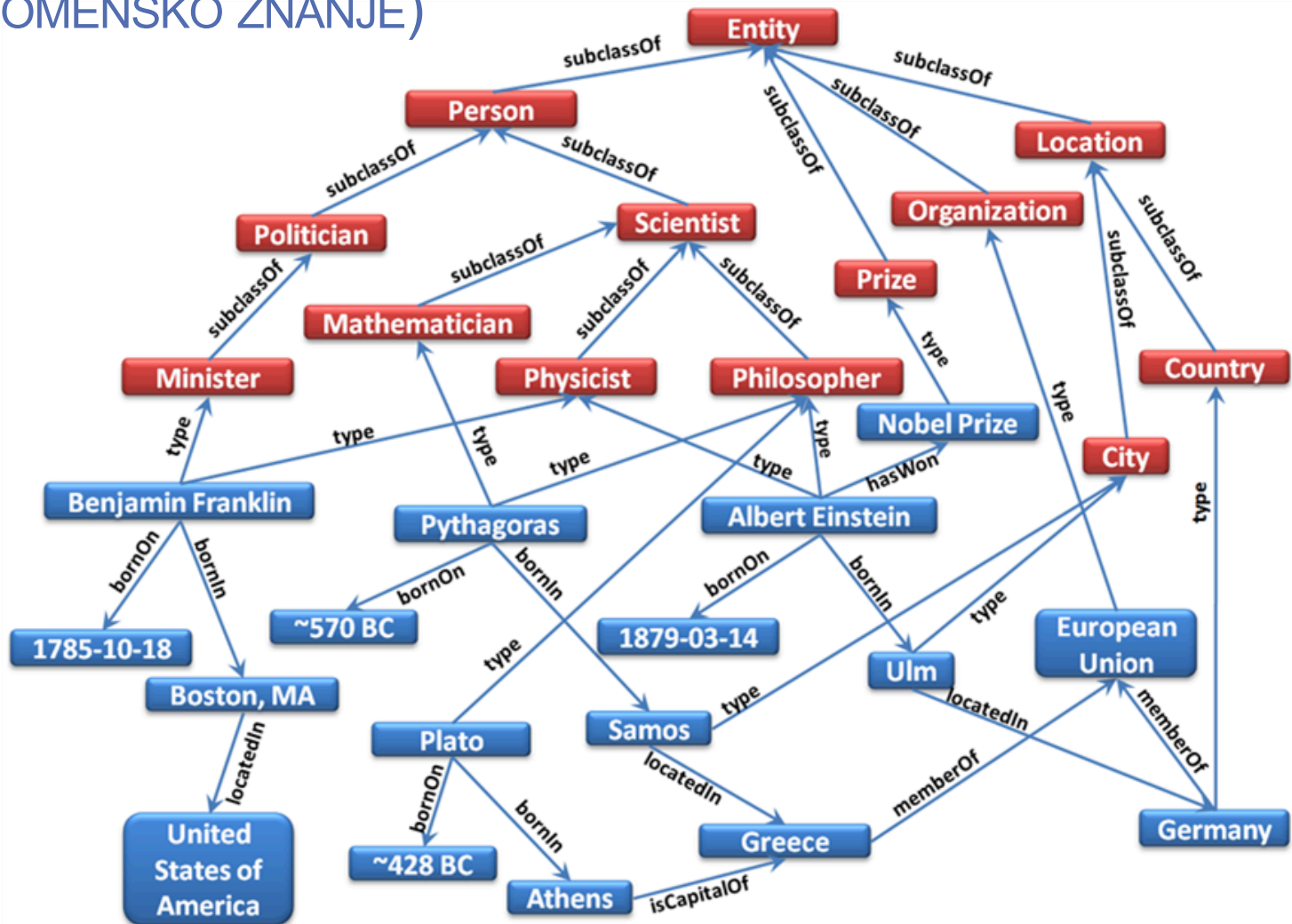


Primer ilustruje  
mali segment [ConceptNet](http://www.opasquet.fr/omcsnet/) baze znanja

# PRIMER GRAFA KOJI SADRŽI 2 NIVOVA ZNANJA:

1) KLASE/KONCEPTE (META-ZNANJE)

2) ENTITETE (DOMENSKO ZNANJE)



# GRAFOVI PODATAKA I ZNANJA U KOMERCIJALNOJ SFERI

# GOOGLE'S KNOWLEDGE GRAPH

Boyhood Jelena


Web Images Videos News Maps More Search tools


About 20,300,000 results (0.52 seconds)

**Boyhood (2014) - IMDb**  
[www.imdb.com/title/tt1065073/](http://www.imdb.com/title/tt1065073/)  
 ★★★★★ Rating: 8/10 - 217,070 votes  
 Videos. **Boyhood** -- Clip: Talk To Me · **Boyhood** -- Featurette: Behind the Scenes.  
 Photos. Still of Ellar Coltrane in **Boyhood** (2014) Patricia Arquette and Rosanna ...  
[Full Cast & Crew](#) - [Ellar Coltrane](#) - [Lorelei Linklater](#) - [Patricia Arquette](#)

**Boyhood (film) - Wikipedia, the free encyclopedia**  
[https://en.wikipedia.org/wiki/Boyhood\\_\(film\)](https://en.wikipedia.org/wiki/Boyhood_(film))  
**Boyhood** is a 2014 American independent coming-of-age drama film, written and directed by Richard Linklater, and starring Patricia Arquette, Ellar Coltrane, ...  
[Ellar Coltrane](#) - [Lorelei Linklater](#) - [Patricia Arquette](#) - [Richard Linklater](#)

**Boyhood (2014) - Rotten Tomatoes**  
[www.rottentomatoes.com/m/boyhood/](http://www.rottentomatoes.com/m/boyhood/)  
 ★★★★★ Rating: 98% - 267 votes  
 Critics Consensus: Epic in technical scale but breathlessly intimate in narrative scope, **Boyhood** is a sprawling investigation of the human condition.


**Boyhood - International Trailer (Universal Pictures) HD ...**  
 [www.youtube.com/watch?v=Ys-mbHXyWX4](http://www.youtube.com/watch?v=Ys-mbHXyWX4)  
 Apr 25, 2014 - Uploaded by Universal Pictures UK  
[www.facebook.com/BoyhoodMovieUK](http://www.facebook.com/BoyhoodMovieUK). Richard Linklater's **BOYHOOD** -- a fictional drama made with the same ...

In the news  
 [Gordon honored at boyhood home before](#)

## Boyhood

2014 film





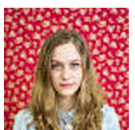
★★★★★ 8/10 · IMDb  
 ★★★★★ 98% · Rotten Tomatoes  
 ★★★★★ 100% · Metacritic



The joys and pitfalls of growing up are seen through the eyes of a child named Mason (Ellar Coltrane), his parents (Patricia Arquette, Ethan Hawke) and his sister (Lorelei Linklater). Vignettes, filmed with the same cast over the course of 12 years, capture family meals, road trips, birthday parties... [More](#)

**Initial release:** July 11, 2014 (USA)  
**Director:** [Richard Linklater](#)  
**Running time:** 2h 46m  
**Screenplay:** [Richard Linklater](#)  
**Awards:** [Academy Award for Best Actress in a Supporting Role](#), more

**Cast** View 5+ more

 <a href="#">Ellar Coltrane</a>	 <a href="#">Patricia Arquette</a>	 <a href="#">Ethan Hawke</a>	 <a href="#">Lorelei Linklater</a>	 <a href="#">Zoe Graham</a>
---	--	--	--	---



# GOOGLE'S KNOWLEDGE GRAPH

*“...Google users will be able to browse through the company’s ‘**knowledge graph**,’ or its ever-expanding **database** of information about ‘**entities**’ – people, places and things – **the ‘attributes’ of those entities and how different entities are connected to one another.**”*

[What Google's Search Changes Might Mean for You](#)

Wall Street Journal, March 14, 2012

# GOOGLE'S KNOWLEDGE GRAPH

*“...Every piece of information that we crawl, index, or search is analyzed in the context of Knowledge Graph.”*

*“...Almost all the structured data from all of our products like Maps and Finance and Movies and Music are all in the Knowledge Graph, so we can reasonably say that everything we know about is in this canonical form.”*

[How a Database of the World's Knowledge Shapes Google's Future](#)

MIT Technology Review, January 27, 2014

# GOOGLE'S KNOWLEDGE GRAPH

*“[Google Now] works by using machine learning algorithms to determine what you’re doing, then matches this understanding with information stored in what the company calls the Google Knowledge Graph—a database of semantic data describing more than 1 billion people, places, and things. ‘To be able assist you,’ says Aparna Chennapragada, who oversees Google Now, ‘we have to understand the world.’”*

[Startup Unleashes Its Clone of Google's 'Knowledge Graph'](#)

Wired, April 6, 2014

# FACEBOOK'S ENTITY GRAPH

*“Facebook is building a rich stock of knowledge that could make its software smarter and boost the usefulness of its search engine...*

*...Entities such as colleges and employers are **learned from data typed in profile pages**; businesses, movies, fictional characters, and other concepts are **learned from fan pages** created by Facebook users. ... **analyzing many employment histories** on the site allows Facebook's search engine to know that a search for “software engineers” should also return people who say they are “coders.”*

[Facebook Nudges Users to Catalog the Real World](#)

MIT Technology Review, February 27, 2013

# MICROSOFT'S SATORI

*“At the core of Microsoft's work to create a state-of-the-art Bing digital assistant is **Satori**, a **knowledge repository** of more than a billion objects digested in the past 3.5 years...*

*...Satori catalogs **entities** and the **associated data and relationships** among them...*

*...Satori is a **self-learning system** that is running every day and learning more, adding 28,000 DVDs of content every day...*

*...Bing search and Windows already are using Satori's knowledge repository...”*

[Microsoft's Bing seeks enlightenment with Satori](#)

CNET News, July 30, 2013

# BING'S KNOWLEDGE AND ACTION GRAPH

*“Bing has over a billion **entities** (people, places, and things) and the number is growing every day. For those entities, we have over 21 billion associated **facts**, 18 billion links to **key actions** and over 5 billion **relationships between entities**.*

*Millions of Bing users around the globe use this rich information every day, in bing.com, Cortana, Xbox, Office and more”*

*“knowledge and action graph will be available to developers via a new API”*

[Bing announces availability of the knowledge and action graph API](#)

Bing Blogs, 20 August 2015

# YAHOO'S KNOWLEDGE GRAPH

*“Spark [a semantic search assistance tool] takes a large **entity graph** as input, ... consisting of the most important entities, their most important related entities, and their respective types. This entity graph is drawn from a larger Yahoo! Knowledge Graph, a unified **knowledge base** that provides **key information** about all the **entities** we care about, and **how they relate** to each other.”*

[Entity Recommendations in Web Search](#)

The 12th International Semantic Web Conference, Oct 2013

LINKEDIN:

## PROFESSIONAL GRAPH => ECONOMIC GRAPH

*“[Economic graph is] a digital mapping of the global economy, comprised of a profile for every professional, company, job opportunity, the skills required to obtain those opportunities, every higher education organization, and all the professionally relevant knowledge associated with each of these entities”*

*“With these elements in place, we can connect talent with opportunity at massive scale”*

[Announcing The LinkedIn Economic Graph Challenge](#)

Oct 14, 2014



OTVORENI, JAVNO DOSTUPNI  
GRAFOVI PODATAKA I ZNANJA

# DBPEDIA

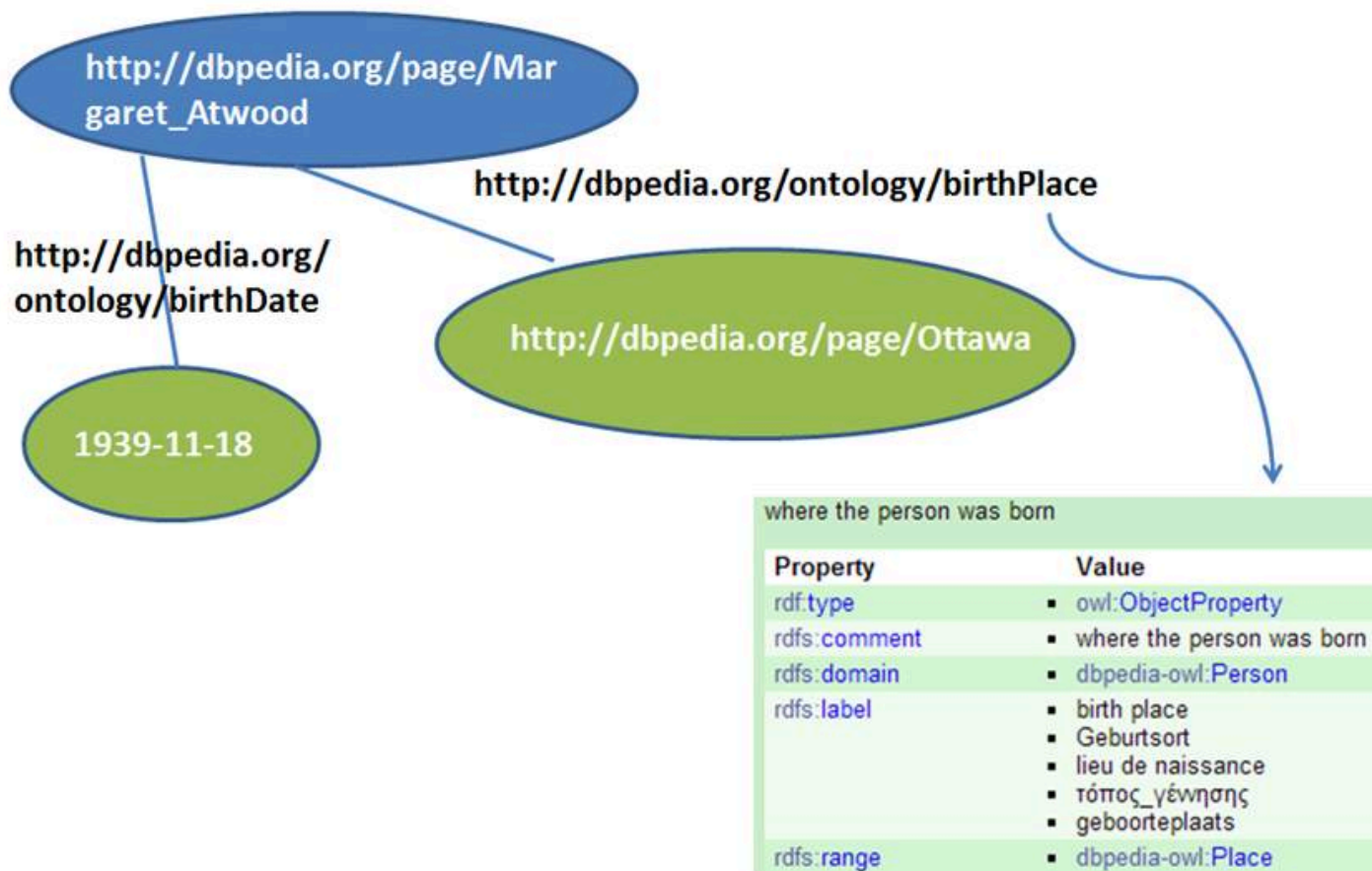
Mašinski čitljiva verzija Wikipedia-e

Podaci preuzeti iz Wikipedia-e su:

- Strukturirani: predstavljeni u formi {subjekat-predikat-objekat} tripleta pogodnih za procesiranje
- Semantički opisani: semantika svakog elementa tripleta je eksplicitno definisana => može se direktno interpretirati od strane računara

# DBPEDIA

## Predstavljanje podataka i znanja u Dbpedia-i



# DBPEDIA

[http://en.wikipedia.org/wiki/San\\_Francisco](http://en.wikipedia.org/wiki/San_Francisco)

<b>Country</b>	 United States
<b>State</b>	 California
<b>Founded</b>	June 29, 1776
<b>Incorporated</b>	April 15, 1850 <sup>[9]</sup>
<b>Founded by</b>	José Joaquín Moraga Francisco Palóu
<b>Named for</b>	St. Francis of Assisi
<b>Government</b>	
• <b>Type</b>	Mayor-council
• <b>Body</b>	Board of Supervisors
• <b>Mayor of San Francisco</b>	Ed Lee
• <b>Board of Supervisors</b>	<b>Supervisors</b> <span>[show]</span>
• <b>California State Assembly</b> <sup>[10][11]</sup>	Tom Ammiano (D) Phil Ting (D)
• <b>CA State Senate</b> <sup>[12]</sup>	Mark Leno (D)
• <b>United States House of Representatives</b> <sup>[13][14]</sup>	Nancy Pelosi (D) Barbara Lee (D) Jackie Speier (D)
<b>Area</b> <sup>[15]</sup>	
• <b>Consolidated city-county</b>	231.89 sq mi (600.6 km <sup>2</sup> )
• <b>Land</b>	46.87 sq mi (121.4 km <sup>2</sup> )

```
<http://dbpedia.org/resource/San_Francisco>
db:country dbpedia:United_States ;
...
db:foundingDate "1776-6-29"^^xsd:date ;
dbpprop:namedFor
    dbpedia:Francis_of_Assisi ;
db:governmentType
    dbpedia:Mayor-council_government ;
...
```

# WIKIDATA

*“**Wikipedia’s data** is buried in 30 million Wikipedia articles in 287 languages from which **extraction is inherently very difficult.**”*

*“Population numbers for Rome, for example, can be found in English and Italian articles about Rome but also in the English article “Cities in Italy.” The **numbers are all different.**”*



Osnovni ciljevi WikiData projekta:

- učiniti Wikipedia podatke pristupačnim aplikacijama, tj spremnim za direktno procesiranje;
- obezbediti kontinuiranu tačnost i ažurnost podataka

# WIKIDATA

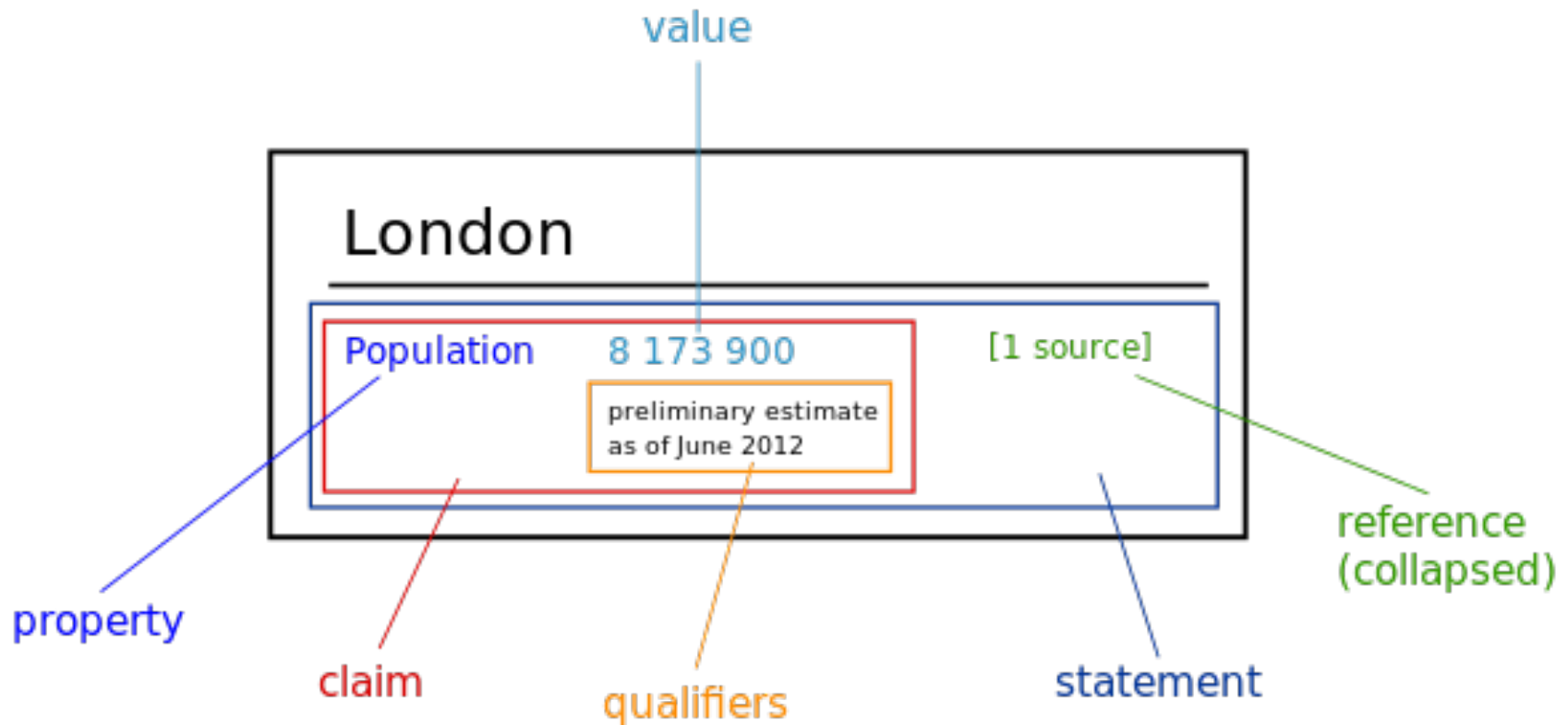
*“Wikidata is a project of the Wikimedia Foundation: a **free, collaborative, multilingual, secondary database**, collecting **structured data** to provide support for Wikipedia, Wikimedia Commons, the other Wikimedia projects, and well beyond that”*

*“A secondary database: Wikidata can record not just statements, but also their sources, thus reflecting the **diversity of knowledge** available and supporting the notion of **verifiability**”*

*“Collecting structured data: [to] allow easy **reuse of that data** by Wikimedia projects and third parties, and enable computers to easily process and “understand” it.”*

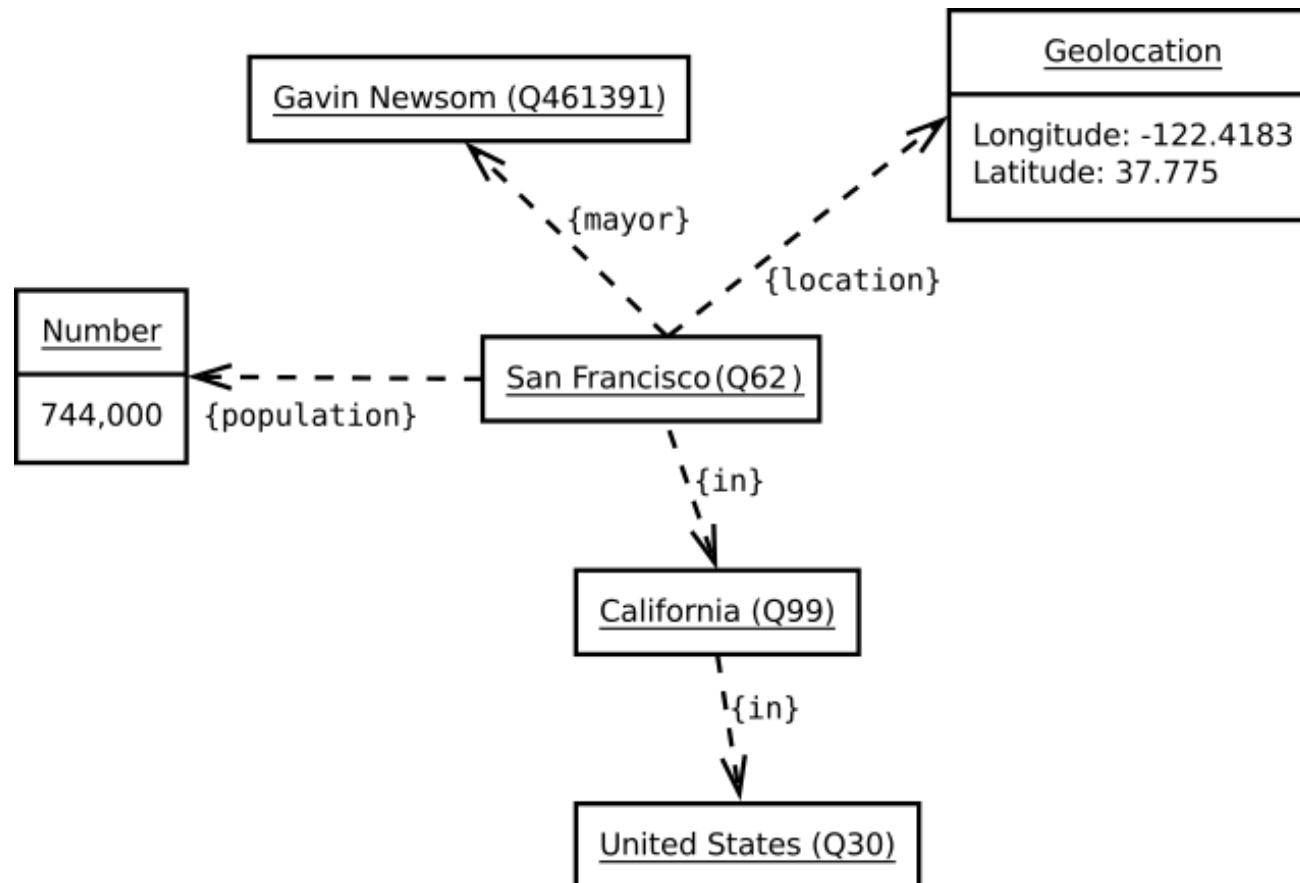
# WIKIDATA

## Predstavljanje podataka i znanja u WikiData bazi



# WIKIDATA

## Predstavljanje podataka i znanja u WikiData bazi





# GIGANTSKI GLOBALNI GRAF

GIGANTSKI GLOBALNI GRAF (1):  
VIZIJA WEB-A KAO  
GIGANTSKE GLOBALNE BAZE (GRAFA)  
PODAKA I ZNANJA  
BY SIR TIM BERNERS-LEE

# GIGANTSKI GLOBALNI GRAF (1)

## Faza 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"*

## Faza 2: **World Wide Web (WWW)**

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

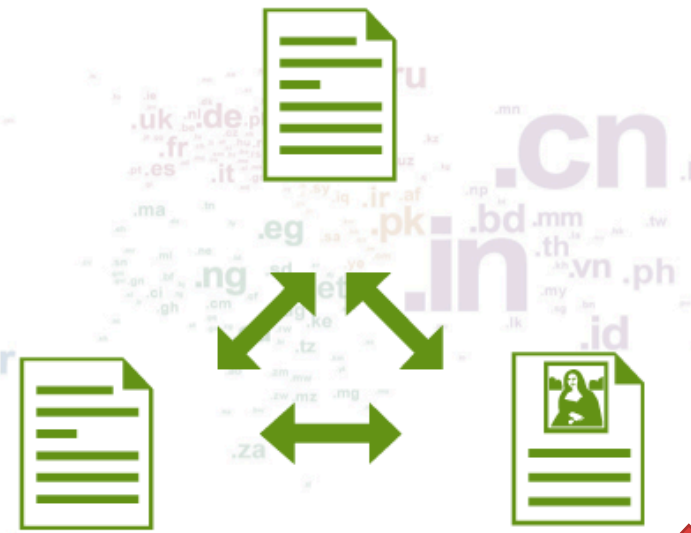
## Faza 3: **Gigantic Global Graph (GGG)**

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

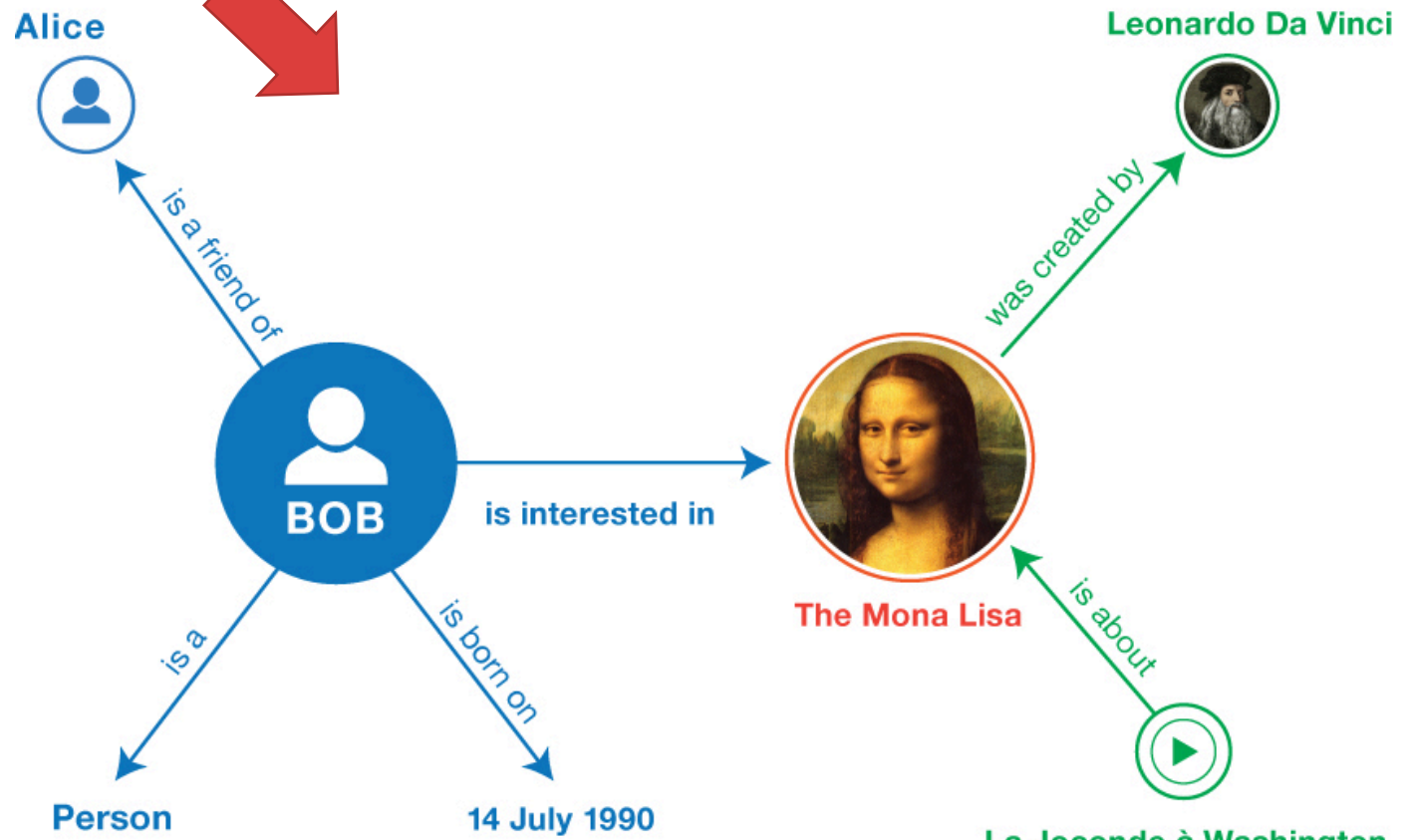
# GIGANTSKI GLOBALNI GRAF (1)

*“...when I book a flight it is the flight that interests me. Not the flight page on the travel site, or the flight page on the airline site, but the URI (issued by the airlines) of the flight itself. ... And whichever device I use ... it will access a situation-appropriate view of an integration of everything I know about that flight from different sources. The task of booking and taking the flight will ... be primary things in my awareness, the websites involved will be secondary things, and the network and the devices tertiary.”*

# WWW (= Web of documents)



# GGG (= Web of data)



# WEB DOKUMENATA

Dizajniran za: **direktno korišćenje od strane ljudi**

Primarni objekti: **dokumenti i multi-medija**

Stepen strukturiranosti objekata: **prilično nizak**

Linkovi: **između dokumenata (ili njihovih delova)**

Semantika sadržaja i linkova: **implicitna**

Analogija: **globalni fajl sistem**

# WEB PODATAKA

Dizajniran za: ljude koje 'opslužuju' programi

Primarni objekti: resursi\* i opisi resursa

Stepen strukturiranosti objekata: visok

Linkovi: između dokumenata i između resursa

Semantika sadržaja i linkova: eksplicitna

Analogija: globalna baza podataka

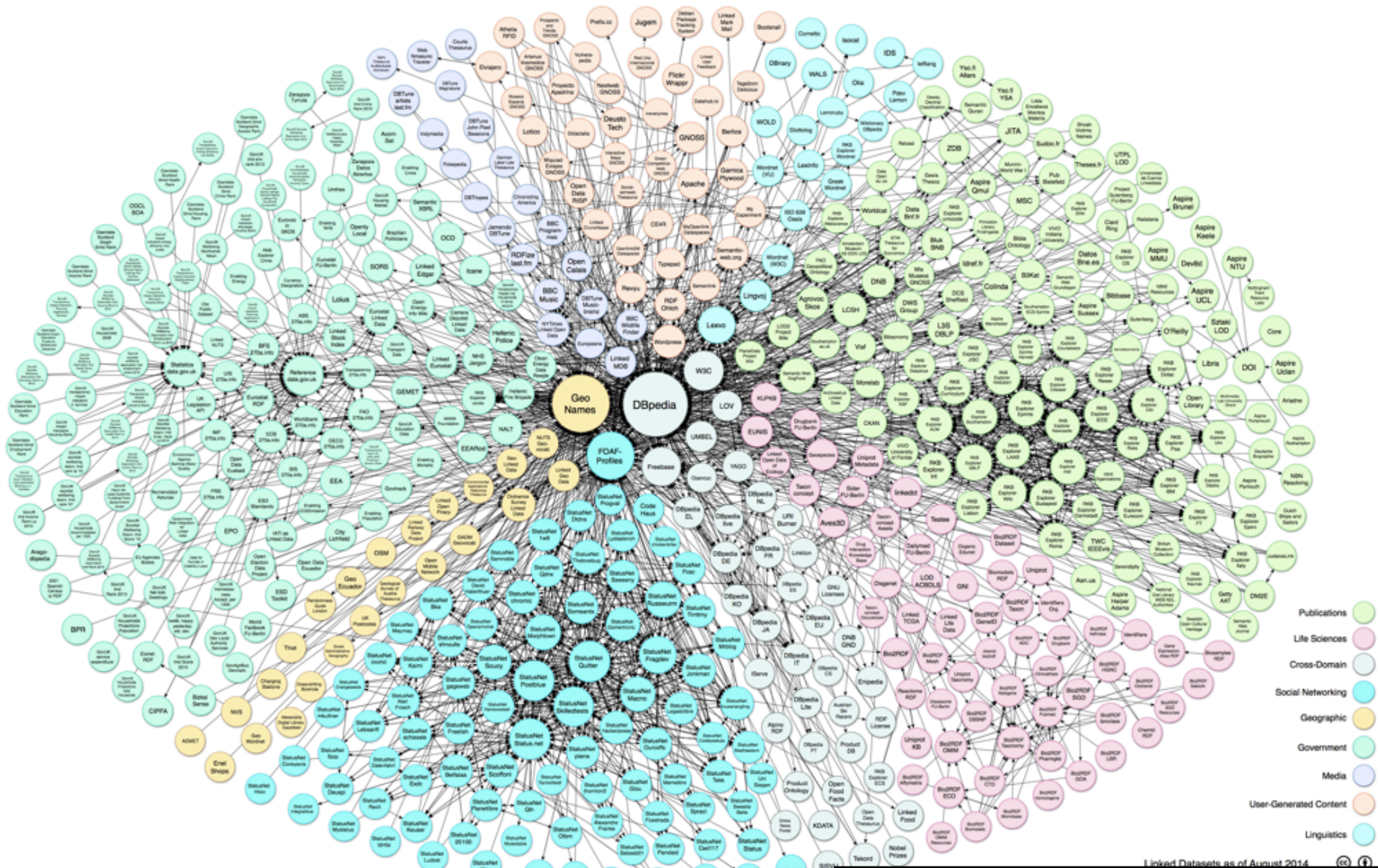
\* Resurs je bilo šta što se može jedinstveno identifikovati tj. ima svoj URI. Npr., Beograd je resurs, čiji je URI: <http://dbpedia.org/resource/Belgrade>

# WEB DOKUMENATA VS. WEB PODATAKA

	<b>Web dokumenata</b>	<b>Web podataka</b>
Dizajniran za	Direktno korišćenje od strane ljudi	Ljude koje 'opslužuju' programi
Primarni objekti	Dokumenti (uključujući multimediju)	Resursi i opisi resursa
Linkovi između	Dokumenata	I dokumenata i resursa
Stepen strukturiranosti	Prilično nizak	Visok
Semantika sadržaja i linkova	Implicitna	Eksplicitna
Analogija	Globalni fajl sistem	Globalna baza podataka



# WEB (OTVORENIH) PODATAKA



Animacija koja ilustruje razvoj LOD-a:

<http://goo.gl/49p9Eh>

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

# GIGANTSKI GLOBALNI GRAF (2)

Gigantske graf baze znanja koje

- sadrže strukturirane podatke ekstrahirane iz Web stranica
- kontinuirano rastu i evoluiraju kako bi njihov sadržaj uvek odražavao podatke i znanje Web-a

Karakteristike

- automatizovani sistemi
- kombinuju različite metode m. učenja radi kontinuiranog unapređenja performansi ekstrakcije podataka/činjenica
- predmet intenzivnih istraživanja (u kompanijama i akademiji) sa ciljem unapređenja ekstrakcije podataka i znanja iz teksta

# READ THE WEB

Istraživački projekat pri Carnegie Mellon University

- <http://rtw.ml.cmu.edu/rtw/>

Ciljevi:

- Razvoj sistema m. učenja koji će kontinuirano vršiti ekstrakciju strukturiranih informacija iz nestrukturiranih Web sadržaja
- Razvoj najveće baze strukturiranog znanja koja
  - u svakom trenutku reflektuje činjenični sadržaj Web-a (tj poseduje svo činjenično znanje koje Web u tom trenutku sadrži),
  - kontinuirano raste kako na nivou instanci, tako i koncepata i relacija
  - može poslužiti kao osnova za brojne AI sisteme

# NEVER ENDING LANGUAGE LEARNER (NELL)

NELL je sistem koji se razvija u okviru Read the Web projekta




















Kontinuirano (24/7) izvršava 2 zadatka:

- *Reading task*: identifikuje u Web stranicama nove instance klasa i relacija iz svoje baze znanja, i tako proširuje bazu
- *Learning task*: uči da “bolje čita”, odnosno da vrši precizniju ekstrakciju informacija iz Web stranica
  - modeli mašinskog učenja na kojima je NELL zasnovan kontinuirano se iznova obučavaju koristeći stalno sve veću bazu znanja kao izvor podataka za trening

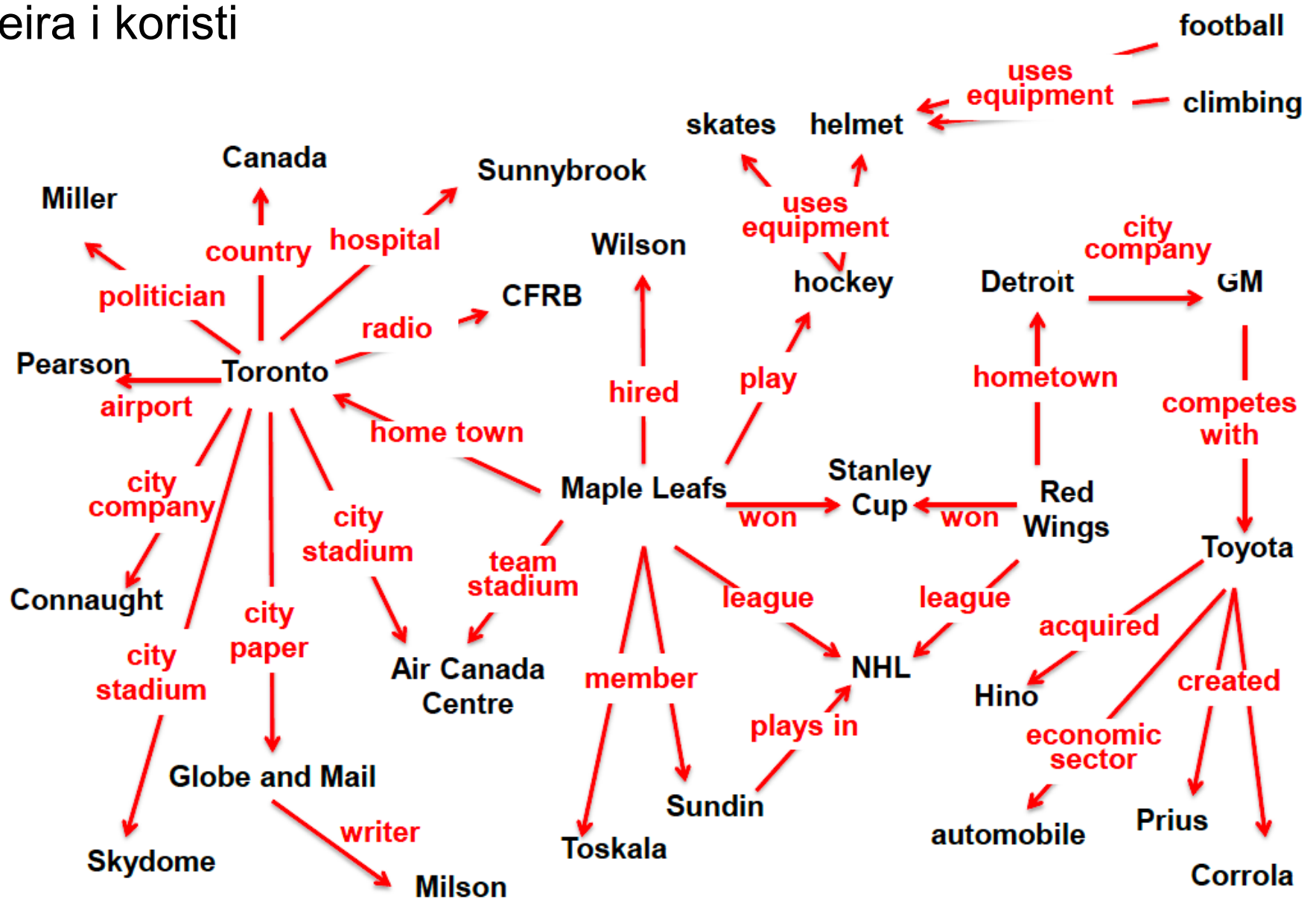
Možete pratiti NELL dok 'čita',  
i pomoći mu da 'nauči' da bolje 'čita'

## Recently-Learned Facts

Refresh

instance	iteration	date learned	confidence		
<u>estonians</u> is an <u>ethnic group</u>	956	23-oct-2015	97.9		
<u>jane krakowski</u> is an <u>architect</u>	955	20-oct-2015	100.0		
<u>cups</u> <u>spoons</u> is an <u>item found in the kitchen</u>	956	23-oct-2015	99.6		
<u>voight park</u> is a <u>zoo</u>	959	07-nov-2015	90.5		
<u>michael mayer</u> is <u>american</u>	955	20-oct-2015	100.0		
<u>bob</u> is a U.S. politician who <u>holds the office of president</u>	959	07-nov-2015	99.2		
<u>karl rove</u> <u>works for fox</u>	956	23-oct-2015	93.8		
<u>david toseland</u> <u>died in</u> the country <u>england</u>	960	23-nov-2015	100.0		
<u>cavaliers</u> is a sports team that <u>won</u> the <u>nba finals</u>	955	20-oct-2015	98.4		
<u>logan</u> <u>was born in</u> <u>chicago south</u>	960	23-nov-2015	99.6		

# Fragment baze (grafa) znanja koji NELL kreira i koristi



Source: [ACM Webinar “Never-Ending Learning to Read the Web” by Tom Mitchell](#), April 2013

# GOOGLE'S KNOWLEDGE VAULT (KV)

Probabilistička baza znanja koja bi trebalo da sadrži svo činjenično znanje koje Web poseduje, kao i da se uvećava i menja kako se Web razvija i menja

Predstavljanje znanja:

- koristi {subjekat-predikat-objekat} triplete (poput DBpedia-e),
- svaki triplet ima pridruženi *confidence score* koji predstavlja procenjenu verovatnoću da je triplet tačan

Kombinuje:

- automatizovanu ekstrakciju činjenica iz Web stranica (neizvesno, neprovereno znanje)
- znanje preuzeto iz postojećih baza znanja (provereno, validirano znanje)

# GOOGLE'S KNOWLEDGE VAULT (KV)

## Poređenje KV-a sa drugim, sličnim bazama znanja

Name	# Entity types	# Entity instances	# Relation types	# Confident facts (relation instances)
<i>Knowledge Vault (KV)</i>	1100	45M	4469	271M
DeepDive [32]	4	2.7M	34	7M <sup>a</sup>
NELL [8]	271	5.19M	306	0.435M <sup>b</sup>
PROSPERA [30]	11	N/A	14	0.1M
YAGO2 [19]	350,000	9.8M	100	4M <sup>c</sup>
Freebase [4]	1,500	40M	35,000	637M <sup>d</sup>
Knowledge Graph (KG)	1,500	570M	35,000	18,000M <sup>e</sup>

**Table 1: Comparison of knowledge bases. KV, DeepDive, NELL, and PROSPERA rely solely on extraction, Freebase and KG rely on human curation and structured sources, and YAGO2 uses both strategies. Confident facts means with a probability of being true at or above 0.9.**

NELL (Never ending Language Learner) je sistem koji se razvija u okviru pomenutog Read the Web projekta



# DIFFBOT'S GLOBAL INDEX

*“...in recent months Diffbot has been analyzing websites to build its index at a rate of up to 15 million pages a day.*

*Its Global Index now contains more than 600 million objects (this can be anything from a celebrity to an Ikea chair model) and 19 billion facts.*

*‘Our approach is fairly radical in that there’s no human behind the curtain’*

*Diffbot ... [is] enhancing other search engines including Microsoft’s Bing and DuckDuckGo, and powering apps for companies such as Cisco and AOL”*

[Diffbot Challenges Google Supremacy With Rival Knowledge Graph](#)

Xconomy, June 4, 2015

# PREPORUKE

- [article] Diffbot Bests Google's Knowledge Graph To Feed The Need For Structured Data ([link](#))
- [video] Mike Tung, DiffBot CEO, Turning the Web into a Structured Database ([link](#))
- [video] Knowledge Vault: A Web-Scale Approach to Probabilistic Knowledge Fusion ([link](#))
  - predavanje koje objašnjava ideju i značaj KV-a, kao i principe na kojima se prikupljanje znanja zasniva
- [video] Tom Mitchell, Never-Ending Learning to Read the Web ([link](#))
- [video] From Structured Data to Knowledge Graph, Google I/O 2013 ([link](#))