Otvoreno računarstvo

4c. Povezani otvoreni podaci

Creative Commons

Otvoreno računarstvo

4c. Povezani otvoreni podaci

- **Uvod**
- RDF i primjeri

Podsjetnik na razno-razne principe otvorenih podataka

- Potpuni
- Primarni
- Pravovremeni
- Pristupačni
- Strojno čitljivi
- …
- Trajni
- Provjereni
- Dokumentirani
- …
- Sveobuhvatni
- Iskoristivi
- Interoperabilni
- …

Povezani podaci – Linked Data

- Human-readable -> Machine readable
- Iako je naglasak na strojno čitljivim podacima,
  uvijek bi trebala postojati reprezentacija podataka čitljiva i ljudima

- Web of documents -> **Web of linked data**
- Semantički web
- Strukturirani, relacijama međusobno povezani podaci označeni globalnim identifikatorima, nad kojima je moguće provoditi semantičke upite

- ***Linked Open Data*** *(LOD) is Linked Data which is released under* ***an open license****, which does not impede its reuse for free.*

Tim Berners-Lee

Četiri principa za dizajn povezanih podataka

- Tim Berners Lee, 2006.

1. uporaba URI-ja za imenovanje stvari
2. uporaba HTTP URI-ja da bi osobe i korisnički agenti mogli upućivati na stvari te ih *pretraživati/razriješiti* (*dereference*)
3. pružanje korisnih informacija prilikom pristupanja (*dereference)* URI-ju, pomoću otvorenih mrežnih normi kao što su RDF ili SPARQL
4. prilikom objavljivanja na Internetu, uključivanje poveznica na druge povezane stvari uporabom njihovih URI-ja

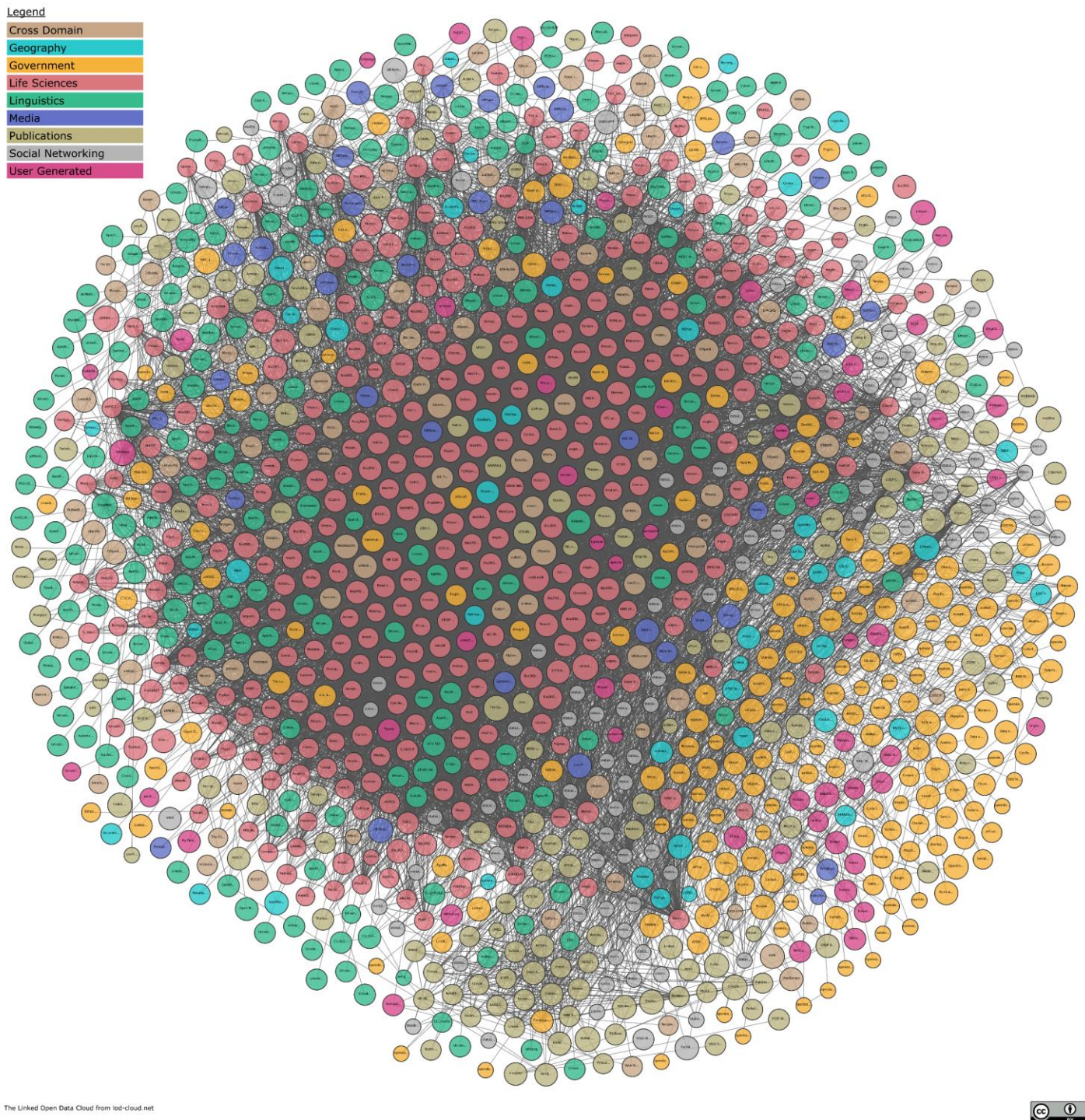
- Povezani podaci bi trebali biti visoke kvalitete
  - Potrebno je temeljito pročišćavanje podataka prije pretvorbe u RDF

Linked data = Open data? = Linked open data?

- Povezani podaci ne moraju biti otvoreni
- Otvoreni podaci ne moraju biti povezani

- Zanima nas presjek: **Povezani otvoreni podaci – Linked Open Data (LOD)**

- Linked Open Data cloud
  - https://lod-cloud.net/
  - Svibanj 2020:
    - 1260 skupova podataka
    - 16187 veza



Ciljevi

- U svijetu otvorenih državnih podataka
  - Veća transparentnost
  - Bolja suradnja među organizacijama, vladama, ...
  - Razvoj novih inovativnih usluga
  - Povećanje kvalitete donošenja odluka i politika

- Primjene općenito
  - Strukturirano pretraživanje podataka
  - Poslovna inteligencija i analitika
  - NLP
  - Strojno učenje, umjetna inteligencija
  - ...

Model „5 zvjezdica otvorenih podataka” – 5-star model

- Izvorni tekst - autor: Tim Berners-Lee
  - začetnik weba i ideje povezanih podataka
- 1: omogućite dostupnost podataka na webu u bilo kojem obliku pod otvorenom licencijom
- 2: omogućite dostupnost računalno čitljivih strukturiranih podataka (npr. Excel umjesto skenirane slike tablice)
- 3: koristite otvorene formate zapisa (npr. CSV umjesto XLS-a)
- **4: koristite otvorene norme W3C-a za označavanje pojmova tako da drugi mogu izravno pristupiti vašim podacima**
- **5: povežite svoje podatke s drugim podacima za pružanje konteksta**

4 ili 5 zvjezdica?

- Je li dovoljno označiti podatke globalnim identifikatorima?
- Je li dovoljno dodati i relacije, tj. odnose među podacima?
- „**RDF silos**” = ogromne „nakupine” podataka u RDF-u, što ćemo s njima?

- 5 zvjezdica: automatizirano **povezivanje različitih skupova podataka,** uporabom globalnih rječnika za podatke, za odnose, za vrijednosti ...

Dosta vam je kratica?

- Semantički web – standardi
- Brzinski rječnik pojmova: Linked Data Glossary

Syntax and supporting technologies	HTTP · IRI (URI) · RDF (triples · RDF/XML · JSON-LD · Turtle · TriG · Notation3 · N-Triples · TriX (no W3C standard)) · RRID · SPARQL · XML · Semantic HTML
Schemas, ontologies and rules	Common Logic · OWL · RDFS · Rule Interchange Format · Semantic Web Rule Language · ALPS · SHACL
Semantic annotation	eRDF · GRDDL · Microdata · Microformats · RDFa · SAWSDL · Facebook Platform
Common vocabularies	DOAP · Dublin Core · FOAF · Schema.org · SIOC · SKOS
Microformat vocabularies	hAtom · hCalendar · hCard · hProduct · hRecipe · hResume · hReview

Otvoreno računarstvo

4c. Povezani otvoreni podaci

- Uvod
- **RDF i primjeri**

RDF

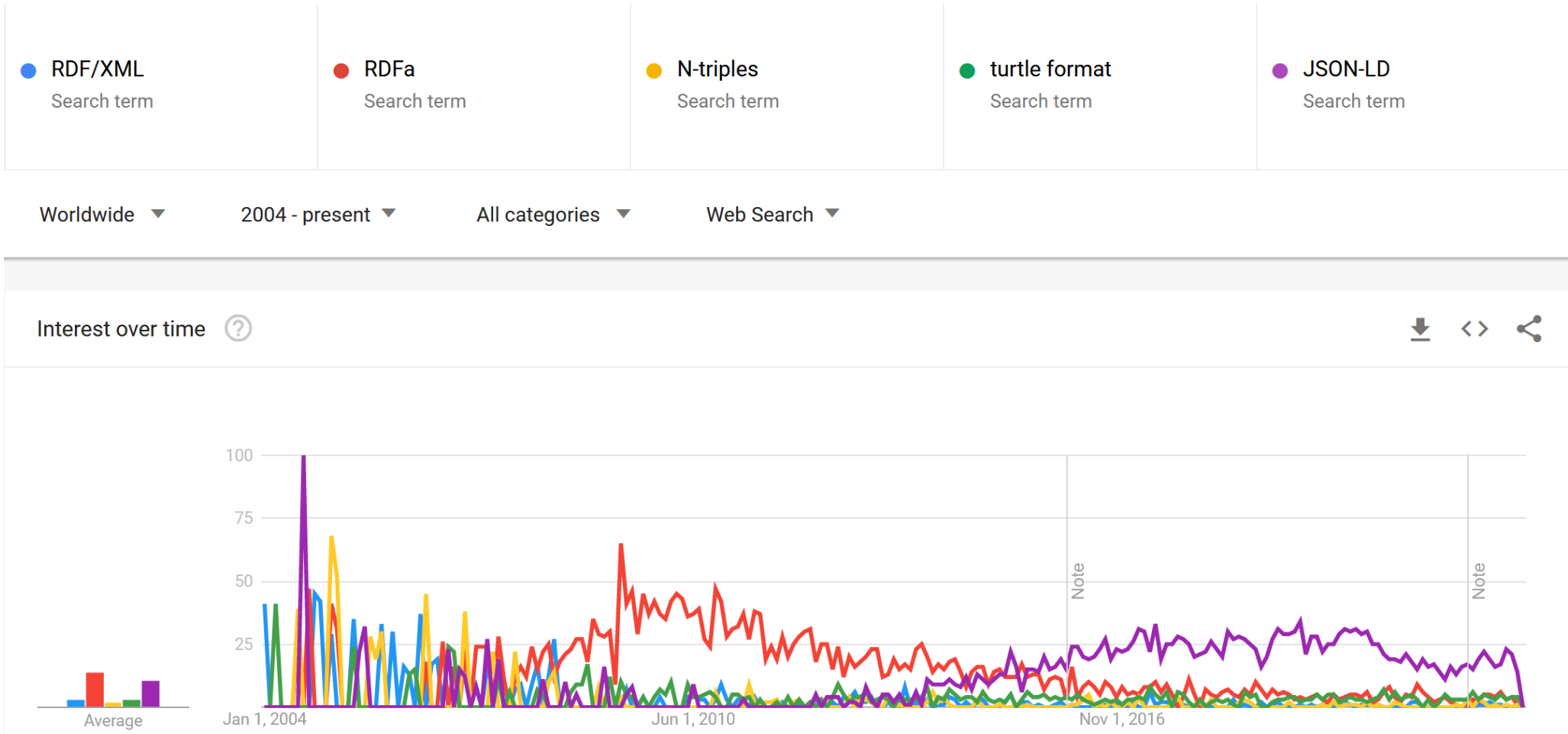
- **R**esource **D**escription **F**ramework
- Resursi?
  - „Bilo što” – jedinstveno identificirano URI-jem
- Description?
  - Formalni opis resursa (svojstva i veze), u obliku **grafa**
- Framework?
  - Cijeli okvir temeljen na modelu:
  - HTTP-protokol, URI, formati za serijalizaciju
- **Trojke** - Subjekt, predikat, objekt
  - **Subjekt**: URI
  - **Predikat**: URI
    - Jednosmjerna relacija, svojstvo, odnos subjekta
  - **Objekt**: URI ili literal – podatkovna vrijednost, s kojom se uspostavlja odnos iz predikata
    - U slučaju vrijednosti, može imati povezan drugi URI za tip podataka

Formati za serijalizaciju RDF-a

- Graf je potrebno prikazati u nekom tekstualnom formatu – **serijalizacija**

- RDF/XML
- RDFa
- N-Triples
- Turtle
- **JSON-LD**

- Usporedba:
  - 1 – W3C
  - 2 – usputni blog :-)



Vokabulari - Rječnici

- Dublin Core
- DCAT
- Schema.org

- FOAF – Friend of a Friend
  - rječnik pojmova vezanih uz osobe:
    - društvene mreže koje opisuju suradnju, prijateljstva, veze
    - reprezentacijske mreže koje opisuju pojednostavljen prikaz svijeta
    - informacijske mreže koje prikazuju neovisne opise povezanih stvari u svijetu
    - Primjeri: foaf:Person, foaf:knows, foaf:name, foaf:birthday, ...
- SKOS – Simple Knowledge Organization System
  - Organizacija „znanja” – rječnika i slično - u RDF-u
    - Koncepti (skos:Concept), relacije (skos:broader, ..),
      mapiranja (skos:closeMatch, ...), kolekcije (skos:orderedCollection, ...)

Rječnici - Česti prefiksi

- Važno je koristiti **postojeće rječnike**, ako su prilagođeni našim potrebama
- Provjeriti
  - **Rasprostranjenost** – hoće li rječnik povećati (ili smanjiti? :-( ) povezanost skupa podataka?
  - **Održavanje** – održava li se rječnik redovito?
  - **Pokrivenost** – pokriva li opsegom dovoljno podataka?
  - **Izražajnost** – je li svojom razinom detalja dovoljan za naše potrebe i podatke?

- Zgodan link:
  - www.prefix.cc
  - Npr:
    - http://prefix.cc/foaf,skos,dcterms.**xml**
    - http://prefix.cc/foaf,skos,dcterms.**json**

Ima još...

- OWL – Web Ontology Language
  - Moćniji od RDF-a:
    - Složene konstrukcije
      - dvosmjerne veze (owl:inverseOf)
      - omogućuje veće restrikcije od RDF-a (RDF ne pazi kakve trojke se zapisuju, OWL može imati moćna pravila)
    - mnogi dodatni predikati, npr. owl:sameAs usporedba po različitim bazama, tj. rječnicima
    - Meta-metapodaci :-)
      - anotacije poput: owl:versionInfo, owl:backwardsCompatibleWith ...
- RDFS – RDF Schema
  - Skup pravila o osnovnim klasama i njihovim svojstvima, za izražavanje odnosa/predikata u RDF-u
    - Klase: rdfs:Resource, rdfs:Class, rdfs:Literal, rdfs:Property...
    - Svojstva: rdfs:label, rdfs:domain, rdfs:range, rdfs:subClassOf...
    - Primjer na Wikipediji
- Shape Expressions (ShEx) – opisivanje, validacija i transformacija RDF-a
  - http://shex.io/
  - Validator Demo

SPARQL

- Simple Protocol and RDF Query Language
  - *(jedino što na prvi pogled nije baš simple...?)*
- Jezik za
  - upite nad grafovima - dohvat podataka
    - upit je moguć nad više izvora podataka, raznih vrsta – *federated query*
      - Podaci mogu biti u RDF-u
      - Podaci mogu biti u nekoj bazi, koja će preko dodatnog alata mapirati podatke u RDF
  - transformaciju RDF podataka, tj. izradu novog grafa iz odgovora
- Upiti: SELECT, ASK, DESCRIBE, CONSTRUCT
- Elementi SELECT upita
  - PREFIX, FROM, SELECT, WHERE, ORDER BY

Primjeri

- Wikidata Query Service
  - https://query.wikidata.org/
    - Wikidata Query Service Tutorial: https://wdqs-tutorial.toolforge.org/
- Geonames
  - 11 milijuna geografskih pojmova
  - http://www.geonames.org/
    - http://www.geonames.org/8531820/gradska-cetvrt-trnje.html
    - https://www.geonames.org/datasources/
- Dbpedia <- zaslužuje svoje slideove :-)

Primjer: DBpedia

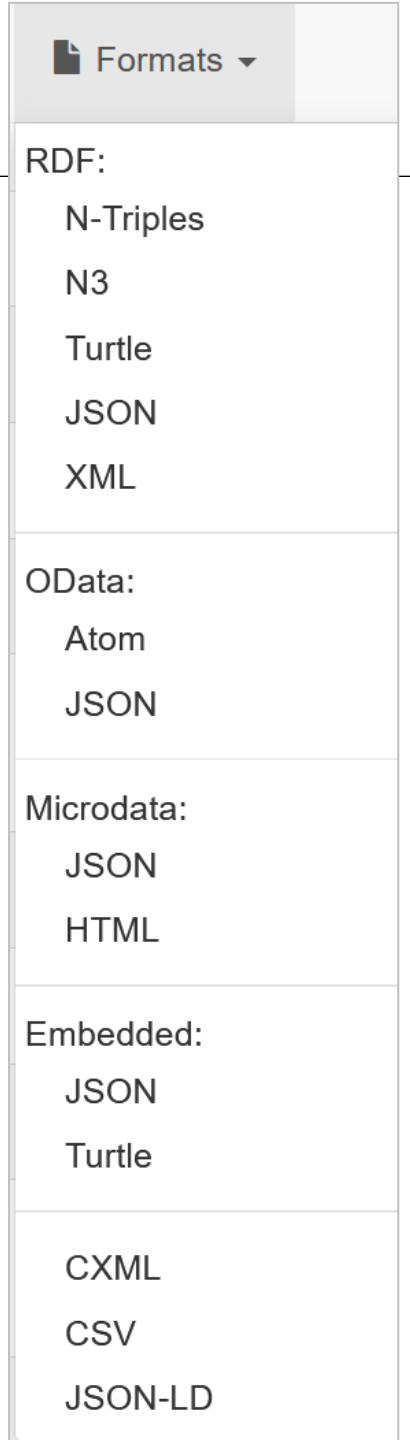
- Otvoreni graf znanja,
  baza strukturiranih, povezanih podataka
  iz 130+ WikiMedia projekata,
  uključujući sve jezike **Wikipedije**,
  WikiMedia Commons i **Wikidata**
- Striktno pridržavanje načela izrade
  povezanih otvorenih skupova podataka
  - URI-ji, HTTP, HTML, RDF, SPARQL
  - Otvorena licencija
- Brojkice – 2020.:
  - 38.3 milijuna stvari u 125 jezika
  - 25 milijuna poveznica na slike
  - 50 milijuna RDF-poveznica, **3 milijarde RDF-trojki**

- https://wiki.dbpedia.org/about
- https://databus.dbpedia.org/dbpedia/collections/latest-core :-)

- Primjer korištenja:
  https://www.dbpedia-spotlight.org/

Primjer: DBpedia

- http://dbpedia.org/page/Zagreb



Alati

- Open Refine
  - Alat otvorenog kôda za upravljanje podacima (čišćenje, pretvorbe...)
  - Proširivanje podataka
  - *Data reconciliation* – povezivanje postojećih podataka s drugim izvorima podataka
    - Defaultno Wikidata, postoje dodaci za mnoge druge izvore
    - Primjer: https://openrefine.org -> 3. Reconcile and Match Data video
- Protégé
  - alat otvorenog kôda za uređivanje ontologija
  - OWL 2
  - uvoz/izvoz u formatima: RDF/XML, Turtle, OWL/XML, ...
  - https://protege.stanford.edu/

Best Practices for Publishing Linked Data

1. Prepare stakeholders
2. Select a dataset
3. Model the data
4. Specify an appropriate license
5. Good URIs for linked data
6. Use standard vocabularies
7. Convert data
8. Provide machine access to data
9. Announce new datasets
10. Recognize the social contract

W3C Best Practices for Publishing Linked Data

Linked Data Publishing Checklist

1. Does your data links to other data sets?
2. Do you provide provenance metadata?
3. Do you provide licensing metadata?
4. Do you use terms from widely deployed vocabularies?
5. Are the URIs of proprietary vocabulary terms dereferenceable?
6. Do you map proprietary vocabulary terms to other vocabularies?
7. Do you provide dataset-level metadata?
8. Do you refer to additional access methods?

Linked Data: Evolving the Web into a Global Data Space
Poglavlje 5.5

Korišten *CreativeCommons* sadržaj

- Open Data Support: Training module 1.2 Introduction to Linked Data, CC BY
- The Linked Open Data cloud, CC BY