Povezani oblak otvorenih podataka

Pintek, Lucija

Sveučilište J.J. Strossmayera u Osijeku

Filozofski fakultet

Preddiplomski studij informatologije

Lucija Pintek

Povezani oblak otvorenih podataka

Završni rad

Mentor: izv. prof. dr. sc. Boris Bosančić

Osijek, 2021.

Sveučilište J.J. Strossmayera u Osijeku

Filozofski fakultet Osijek

Odsjek za informacijske znanosti

Preddiplomski studij informatologije

Lucija Pintek

Povezani oblak otvorenih podataka

Završni rad

Društvene znanosti, informacijske i komunikacijske znanosti, informacijski sustavi i informatologija

Mentor: izv. prof. dr. sc. Boris Bosančić

Osijek, 2021.

Prilog: Izjava o akademskoj čestitosti i o suglasnosti za javno objavljivanje

Obveza je studenta da donju Izjavu vlastoručno potpiše i umetne kao treću stranicu završnog odnosno diplomskog rada.

IZJAVA

Izjavljujem s punom materijalnom i moralnom odgovornošću da sam ovaj rad samostalno napravio te da u njemu nema kopiranih ili prepisanih dijelova teksta tuđih radova, a da nisu označeni kao citati s napisanim izvorom odakle su preneseni.
Svojim vlastoručnim potpisom potvrđujem da sam suglasan da Filozofski fakultet Osijek trajno pohrani i javno objavi ovaj moj rad u internetskoj bazi završnih i diplomskih radova knjižnice Filozofskog fakulteta Osijek, knjižnice Sveučilišta Josipa Jurja Strossmayera u Osijeku i Nacionalne i sveučilišne knjižnice u Zagrebu.

U Osijeku, datum 31.8.2021.

Lucija Ružić, 0122228777
ime i prezime studenta, JMBAG

Sažetak

Svrha ovog rada je pobliže objasniti i prikazati Oblak povezanih otvorenih podataka (LOD oblak). U skladu sa svrhom rada, ciljevi su objasniti koncept povezanih otvorenih podataka, pojasniti strukturu LOD oblaka, kao i pripadajuće setove podataka. Također, cilj rada je navesti i na koje poteškoće LOD oblak nailazi u svom razvoju te što se može očekivati u budućnosti. U prvom dijelu rada govori se o povezanim otvorenim podacima koji čine temelj LOD oblaka. U nastavku rada objašnjava se nastanak LOD oblaka i projekata čiji setovi podataka su prvi činili LOD oblak dijagram. Poseban naglasak stavljen je na strukturu LOD oblaka, točnije setove podataka i triplete koji su objašnjeni na primjerima. Nakon toga, opisani su nedostaci i poteškoće na koje LOD oblak nailazi u svom razvoju. Na kraju rada, izloženi su prijedlozi rješenja navedenih problema te predstavljene neke od usluga koje bi mogle biti dijelom budućnosti LOD oblaka.

Ključne riječi : oblak povezanih otvorenih podataka, setovi podataka, LOD oblak dijagram, povezani otvoreni podaci, RDF

Sadržaj

1. Uvod

Povezani otvoreni podaci ideja su Tim Berners-Lee-ja, začetnika World Wide Web-a, bez kojih ni Oblak povezanih otvorenih podataka (*Linked Open Data Cloud* - LOD *cloud*) ne bi imao smisla. Oblak povezanih otvorenih podataka (LOD oblak) prije četrnaest godina bio je mali dijagram od 12 setova podataka koji se razvio iz projekta DBpedija uz pomoć tehnologije semantičkog weba. Danas dijagram čini preko 1300 setova podataka, a taj broj se, iako u znatno manjoj mjeri, i dalje povećava.

Svrha ovog rada je prikazati Oblak povezanih otvorenih podataka ili LOD oblak. U skladu sa svrhom rada, jedan od ciljeva rada je objasniti koncept povezanih otvorenih podataka. Također, cilj rada je i objasniti i prikazati strukturu LOD oblaka, njegove setove podataka i triplete te navesti kroz koje poteškoće nailazi u svom razvoju, kao i njegove nedostatke. Samim time, izložit će se i neka predložena rješenja tih problema.

Nakon uvoda, u drugom poglavlju rada objasnit će se koncept povezanih otvorenih podataka koji je utemeljen na načelima Tim Berners-Lee-ja. Ujedno će se razjasniti na koji se način povezani podaci mogu učiniti kvalitetnijima primjenom implementacijske sheme Pet zvjezdica otvorenih podataka (*5-star Open Data*), te kao takvi, postati i dijelom LOD oblaka povezanih otvorenih podataka.

U trećem poglavlju najprije se, definira što je to LOD oblak, od čega se sastoji i kako se odvijao njegov razvoj. U izlaganju povijesnog razvoja LOD oblaka uključeni su i projekti - DBpedija i Bio2RDF, u okviru kojih je isti i utemeljen. U nastavku poglavlja donosi se opis seta podataka i tripleta koji čine LOD oblak. Također, prikazuje se način na koji se isti mogu uključivati u LOD oblak putem službene stranice u skladu sa zahtjevima koji se postavljeni. Na primjeru VIAF seta podataka iz knjižničarske struke, pobliže je prikazan jedan tipični set podataka u LOD oblaku, njegove međusobne poveznice s drugim setovima te osnovne informacije o njemu. Tripleti koji čine set podataka također su prikazani na jednostavnom primjeru, u RDF/XML notaciji te u grafičkom obliku. Nadalje, navodi se s kojim se poteškoćama susreću korisnici, ali i sama LOD zajednica. Ujedno je i naveden popis nedostatka vezanih uz kvalitetu i otvorenost podataka opisanih kroz primjere setova podataka. Na kraju trećeg poglavlja opisuje se kakva je budućnost LOD oblaka uz analizu nekih predloženih rješenja za spomenute poteškoće i nedostatke. Poseban naglasak stavljen je na moguće nove usluge LOD oblaka i mogućnost u što se može razviti ako bude koristio postojeće i potencijalno buduće tehnologije semantičkog weba.

2. Povezani otvoreni podaci

„Postoje podaci u svakom aspektu naših života, svakom aspektu posla i zadovoljstva i nije riječ samo o broju mjesta odakle podaci dolaze nego i o njihovom povezivanju. … Dakle, zovu se povezani podaci. Želim da ih stvarate. Želim da to zahtijevate. I mislim da je to ideja vrijedna širenja.“[1]

To je rekao 2009. godine Tim Berners-Lee na TED konferenciji pred mnoštvom gledatelja. Tim govorom potaknuo je objavljivanje povezanih podataka. Povezani otvoreni podaci (*Linked Open Data - LOD)* strojno su čitljivi podaci koji se prikazuju u RDF (*Resource Description Framework*) formatu.[2] Tim Berners-Lee, osnivač World Wide Web-a, prvi je u osobnim zapisima 2006. godine zabilježio prema koja četiri načela povezani podaci trebaju biti oblikovani.[3] U prvom načelu navodi se korištenje URI/IRI-ja ne samo za identificiranje nego i nazive 'stvari' (engl. *thing*), što je od velike važnosti za semantički web i LOD oblak. Drugim načelom insistira se na korištenju HTTP URI-ja za nazive 'stvari', npr. http://www.w3.org/. Nadalje, treće načelo daje preporuke da se kod pregledavanja URI-ja pruže korisne informacije u RDF/XML formatu. Posljednje načelo oblikovanja povezanih podataka vezano je uz zahtjev da poveznice trebaju voditi na druge URI/IRI-je kako bi korisnik otkrio i druge 'stvari' prilikom pretraživanja. Dakle, bez povezivanja podataka jedne stranice s drugom, povezani podaci ne bi imali smisla.[4] Kako bi povezani podaci postali dijelom LOD oblaka, potrebno ih je učiniti otvorenima. U skladu s tim, Berners-Lee je 2010. godine osmislio i implementacijsku shemu Pet zvjezdica otvorenih podataka putem koje je pokazao na koji način otvoreni povezani podaci mogu postati kvalitetniji i otvoreniji, dobivanjem po jedne „zvjezdice“ na svakoj od pet razina poboljšanja. Otvoreni podaci s jednom zvjezdicom su podaci koji su dostupni na internetu u bilo kojem formatu, te imaju otvorenu licencu. Korisnici takve podatke mogu pretraživati, pohranjivati, mijenjati te ih dijeliti s kim god žele.[5] Otvoreni podaci s dvije zvjezdice su dostupni kao strojno čitljivi strukturirani podaci, primjerice, u Excel softverskom programu. Budući da su podaci podložni izmjenama, korisnici ih mogu slobodno mijenjati i objavljivati u drugom formatu, ali i kao otvoreni podaci s jednom zvjezdicom, mogu se pretraživati, pohranjivati i sl. Međutim, takvi podaci se i dalje smatraju zatvorenima, jer korisnici ovise o vlasničkom softveru u kojem jedino mogu otvoriti i mijenjati

[1] Berners-Lee, Tim. The next web. TED Talks. Edinburgh, 2009. [Predavanje] URL: https://www.ted.com/talks/tim_berners_lee_the_next_web/reading-list (2021-08-23)
[2] Usp. Siebes, Ronald…[et al.] Top 10 fair data & software things : Linked open data. URL: https://librarycarpentry.org/Top-10-FAIR/2019/09/05/linked-open-data/ (2021-05-20)
[3] Usp. World Wide Web Consortium. Linked data. URL: https://www.w3.org/wiki/LinkedData (2021-05-20)
[4] Usp. Berners-Lee, Tim. Linked data, 2009. URL: https://www.w3.org/DesignIssues/LinkedData.html (2021-05-20)
[5] Usp. What is five-star linked open data? URL: https://www.ontotext.com/knowledgehub/fundamentals/five-star-linked-open-data/ (2021-05-20)

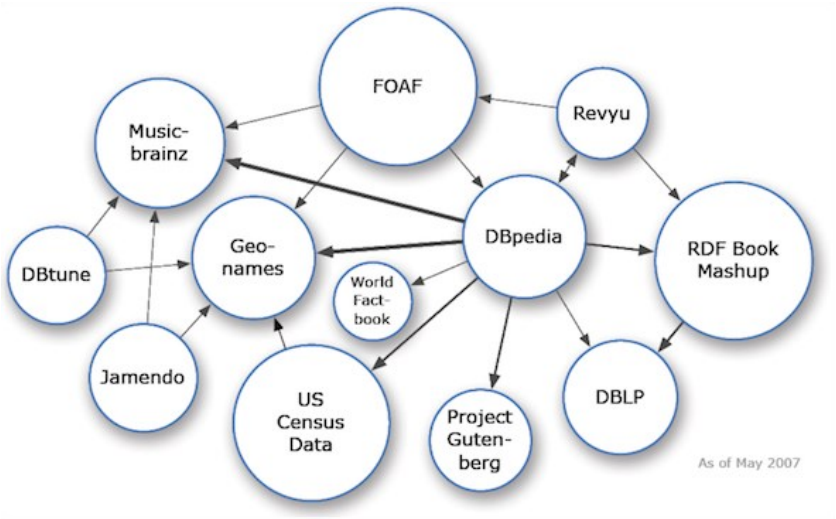
podatke.[6] Otvoreni podaci pohranjeni u nevlasničkim formatima, za koje korisnici ne trebaju vlasničke softvere, imaju tri zvjezdice. Otvoreni podaci s tri zvjezdice nisu pohranjeni u Excel programu, nego u otvorenom programu poput *Comma-separated values (CSV)*[7]. U takvom programu tekstualne datoteke, za odvajanje vrijednosti, koriste zareze.[8] Otvoreni podaci s četiri zvjezdice su podaci koji koriste otvorene standarde poput W3C-a, RDF-a i SPARQL-a za identificiranje stvari.[9] RDF je standardni format koji se koristi u semantičkim grafičkim bazama podataka koje pak svoje podatke prikazuju u obliku čvorova (engl. *nodes*) i rubova (engl. *edges*), pritom čvorovi predstavljaju zapise, a rubovi odnose među zapisima.[10] Takve se baze podataka nazivaju i RDF triplet spremištima (engl. *RDF triplestore*). Za razliku od relacijskih baza podataka, RDF triplet spremišta su u mogućnosti putem mreže tripleta prikazati različite odnose među entitetima u RDF bazi podataka. Temelj svakog RDF tripleta je jedinstveni identifikator resursa (URI ili IRI)[11] i podaci prikazani u grafovima zajedno s priloženim URI/IRI-jima na koje se korisnik može povezati s bilo kojeg mjesta ili ponovno upotrijebiti neke dijelove povezanih podataka. S druge strane, SPARQL je W3C standardizirani upitni jezik za RDF baze podataka.[12] Kao preduvjet za dobivanje pete zvjezdice, korisnici prilikom objavljivanja svojih podataka dužni su povezati te podatke s podacima iz drugih setova podataka kako bi stvorili kontekst. Semantičke grafičke baze podataka sposobne su povezati različite setove podataka s onima iz otvorenih izvora poput DBpedije. Prema tome, takvi povezani otvoreni podaci su dostupni na internetu i cijela mreža povezanih otvorenih podataka postaje korisna i konzumentima i objavljivačima podataka.[13]

[6] Isto.
[7] Usp. Fileformat. URL: https://docs.fileformat.com/spreadsheet/csv/ (2021-05-20)
[8] Usp. What is five-star linked open data? URL: https://www.ontotext.com/knowledgehub/fundamentals/five-star-linked-open-data/ (2021-05-20)
[9] Isto.
[10] Usp. Best NoSQL Databases Software. URL: https://www.g2.com/categories/nosql-databases (2021-05-20)
[11] URI (*Uniform Resource Identifier*) počiva na ASCII kodnoj stranici, dok IRI (*International Resource Identifier*) počiva na UNICODE kodnoj stranici.
[12] Usp. What is five-star linked open data. URL: https://www.ontotext.com/knowledgehub/fundamentals/five-star-linked-open-data/ (2021-05-20)
[13] Isto.

3. Povijest LOD oblaka

3.1. Uvodna razmatranja

LOD oblak predstavlja mrežu povezanih dokumenata pisanih RDF standardom koji su oblikovani prema spomenutim načelima povezanih podataka.[14] LOD oblak prikazan je kao graf znanja, a predstavlja semantički web povezanih podataka. Osim RDF-a, temelji se na standardima poput URI/IRI-ja, HTTP-a i sl., te ga pokreće moderni sustav za upravljanje bazama podataka kao što je *Virtuoso from OpenLink Software*.[15] Tako su 2007. godine Sören Auer i Jens Lehmann sa Sveučilišta u Leipzigu te Christian Bizer sa Sveučilišta u Mannheimu započeli projekt DBpedije. DBpedija je projekt koji je imao za cilj izvući strukturirani sadržaj iz informacija stvorenih kroz različite projekte Wikimedije. Nakon toga su dobivene setove podataka uključili u LOD oblak, i tako je strukturirani sadržaj prikazan u obliku grafa znanja.[16] Prema tome, prvi setovi podataka u LOD oblaku bili su oni iz projekta DBpedije, a prvi skupni LOD oblak dijagram objavljen je u svibnju 2007. godine (Slika 1).



Slika 1 - Prvi LOD oblak iz 2007. godine.

Uz DBpediju, još jedan projekt utemeljio je svoj vlastiti LOD oblak, a to je Bio2RDF. Za razliku od DBpedije koja je prikupljala podatke iz Wikipedije, Bio2RDF je projekt koji je imao za cilj

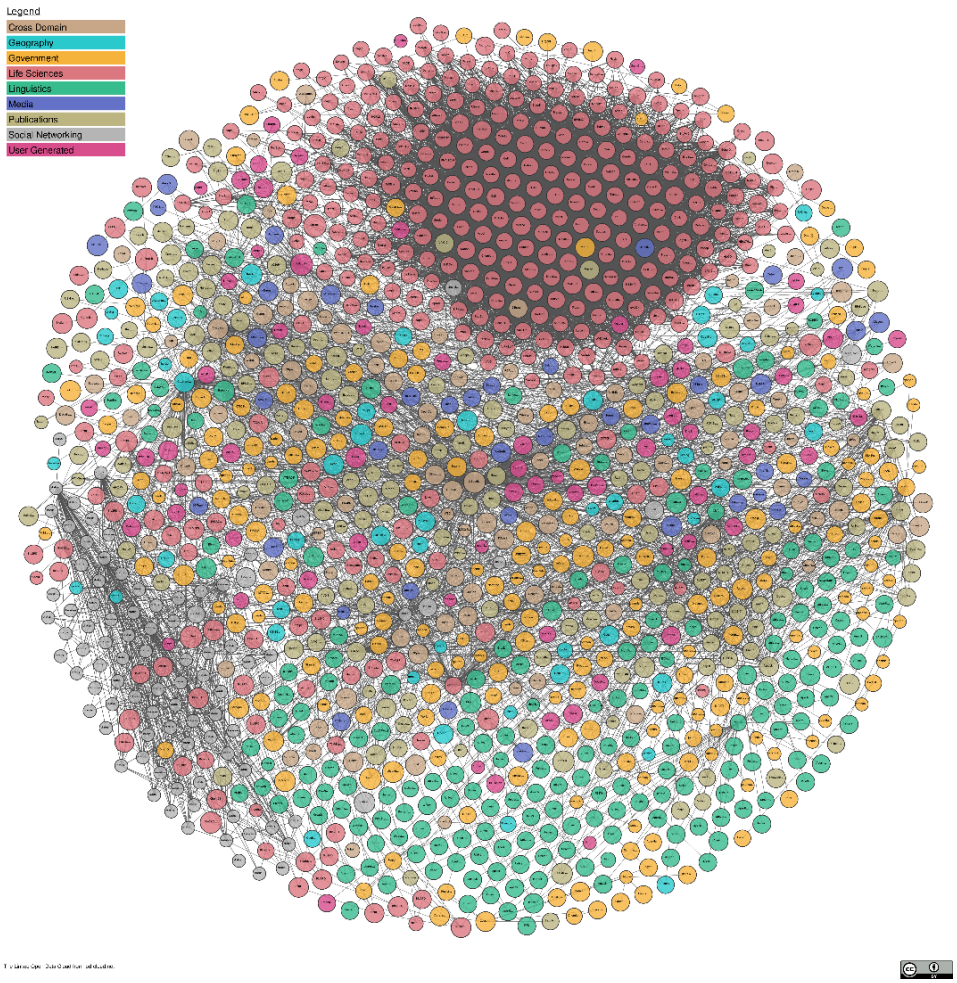
[14] Usp. Open link software. Linked open data (LOD) cloud. URL: https://www.openlinksw.com/describe/?url=http%3A%2F%2Fdata.openlinksw.com%2Foplweb%2Fglossary-term%2FLODCloud%23this&graph=urn%3Adata%3Aopenlink%3Aglossary (2021-08-03)
[15] Usp. Idehen, Kingsley Uyi. What is the linked open data cloud, and why is it important, 2019. URL: https://medium.com/virtuoso-blog/what-is-the-linked-open-data-cloud-and-why-is-it-important-1901a7cb7b1f (2021-08-03)
[16] Usp. DBpedia. URL: https://www.dbpedia.org/about/ (2021-08-03)

izgraditi mrežu povezanih podataka za biomedicinske znanosti.[17] Nakon DBpedijinih i Bio2RDF-ovih setova podataka, tijekom sljedećih četrnaest godina, setovi podataka drugih projekata i organizacija uključivali su se u LOD oblak dijagram, kao što su VIAF, data-open-ac-uk i sl.[18]

LOD oblak dijagram danas održava John P. McCrae s *Insight Centre for Data Analytics*, a tijekom godina izmijenilo se nekoliko osoba u njegovu održavanju - od Jamala Nasira, Jeremyja Debattista, Vladimira Andryushechkina, Anje Jentzsch, Richarda Cyganiaka, Paula Buitelaara do Andrejsa Abelea.[19] A kako bi grafički prikaz ili dijagram LOD oblaka bio konstantno ažuriran, zajednica prikuplja 'meta' informacije o povezanim setovima podataka isključivo na službenoj stranici LOD oblaka dok su se u prošlosti podaci prikupljali i s datahub.io mrežne stranice, koja je nekada imala ulogu registra otvorenih podataka i paketa sadržaja pod sponzorstvom Open Knowledge Foundation.[20] LOD oblak dijagram je 2007. godine činilo samo 12 setova podataka i 20 poveznica (Slika 1), dok se u posljednjem LOD oblak dijagramu iz 2021. godine nalazi 1301 set podataka te 16 283 poveznice (Slika 2).

Od 2009. godine organizacije i projekti čiji setovi podataka su uključeni u LOD oblak dijagram označene su različitim bojama s obzirom na polja i grane znanosti kojima pripadaju. Za razliku od prvog LOD oblak dijagrama i setova podataka iz Wikipedije, danas su tu geografija, lingvistika, mediji i drugi. Prema LOD oblak dijagramu iz 2021. godine, može se uočiti da su biomedicinske znanosti dosada objavile najveći broj setova podataka i da posjeduju i najveći broj poveznica između istih. (Slika 2).

[17] Usp. Dumontier, Michel. Bio2rdf, 2017. URL: https://github.com/bio2rdf/bio2rdf-scripts/wiki (2021-08-03)
[18] Usp. Idehen, Kingsley Uyi. Nav. dj.
[19] Usp. The linked open data cloud. URL: https://lod-cloud.net/#about (2021-08-06)
[20] Usp. World Wide Web Consortium. TaskForces/CommunityProjects/LinkingOpenData/DataSets/CKANmetainformation. URL: https://www.w3.org/wiki/TaskForces/CommunityProjects/LinkingOpenData/DataSets/CKANmetainformation (2021-08-08)



Slika 2 - LOD oblak iz 2021. godine.

Kao što je već navedeno, LOD oblak semantički je web povezanih podataka, gdje RDF poveznice omogućuju kretanje među srodnim povezanim podacima. No iste poveznice mogu se pratiti pomoću semantičkih web paukova (engl. *web crawler*), semantičkih web tražilica poput Swoogle-a (koji više nije aktivan), i koje pružaju jednostavnije načine pretraživanja s obzirom da rezultati upita nisu samo poveznice na HTML mrežne stranice.[21] Što je, zapravo, LOD oblak, dodatno je objašnjeno u idućem poglavlju.

[21] Usp. World Wide Web Consortium. SweoIG/TaskForces/CommunityProjects/LinkingOpenData. URL: https://www.w3.org/wiki/SweoIG/TaskForces/CommunityProjects/LinkingOpenData#LOD_Community_Gatherings (2021-08-08)

3.2. Setovi podataka u LOD oblaku

Setovi podataka i njihove međusobne poveznice su ono što čini bit LOD oblaka. No da bi broj setova podataka u LOD oblaku mogao rasti potrebno je ispuniti određene uvjete koji se postavljaju pred korisnike koji žele uključiti svoje setove podataka u LOD oblak. Dva su osnovna uvjeta koje jedan set podataka mora zadovoljiti kako bi bio uključen u LOD oblak: prvi je da podaci u setu podataka, zbog pretraživanja na SPARQL pristupnoj krajnjoj točci (*SPARQL endpoint*), budu dostupni s dereferenciranim URI/IRI-jima.[22] Dereferencijacija URI/IRI-ja definira se kao dohvaćanje (engl. *retrieving*) reprezentacije izvora na adresi koji URI/IRI identificira.[23] U tom slučaju, dereferencirani URI/IRI svaki puta vodi na stvarnu URL adresu izvora. SPARQL pristupna krajnja točka omogućuje korisnicima pretraživanje seta podataka putem SPARQL upitnog jezika, a rezultati se obično vraćaju u strojno čitljivom obliku.[24] Drugi uvjet dodavanja setova podataka je postojanje barem pedeset RDF poveznica koje vode na druge setove podataka.[25]

Nekada je bilo moguće set podataka uključiti u LOD oblak putem datahub.io mrežne stranice.[26] Međutim, danas je to moguće učiniti jedino putem mrežne stranice samog LOD oblaka. Tražene informacije o setu podataka koje obrazac zahtijeva su jedinstveno ime, naziv, URL, broj tripleta i poveznice na druge setove podataka. Isto tako, kao obavezne informacije potrebno je navesti i opis, domenu seta podataka, kontakt korisnika, SPARQL krajnju točku pristupa te licencu pod kojom se set podataka objavljuje (Slika 3).[27] Međutim, minimalan broj tripleta koji se zahtijeva u setu podataka je 1000, osim toga nužno je da se podaci nalaze u jednom od poznatih RDF formata, poput RDF/XML-a, Turtle-a i sl.[28] Pristup setovima podataka omogućuje *LOD-a-lot*, zbirka indeksiranih datoteka, spremnih za preuzimanje i pretraživanje, koja pruža više od 28 milijardi tripleta iz 650 tisuća setova podataka iz samoindeksirane Header, Dictionary, Triples (HDT) datoteke.[29]

---

[22] Usp. World Wide Web Consortium. TaskForces/CommunityProjects/LinkingOpenData/DataSets/CKANmetainformation. URL: https://www.w3.org/wiki/TaskForces/CommunityProjects/LinkingOpenData/DataSets/CKANmetainformation (2021-08-08)

[23] Farago, Filip; Bosančić, Boris. Povezani podaci i knjižnice. // Vjesnik bibliotekara Hrvatske 54, 4(2013), str. 30. URL: https://hrcak.srce.hr/142376 (2021-08-08)

[24] Usp. European environment agency. SPARQL endpoint, 2017. URL: https://data.europa.eu/euodp/en/data/dataset/european-environment-agency-sparql-endpoint (2021-08-09)

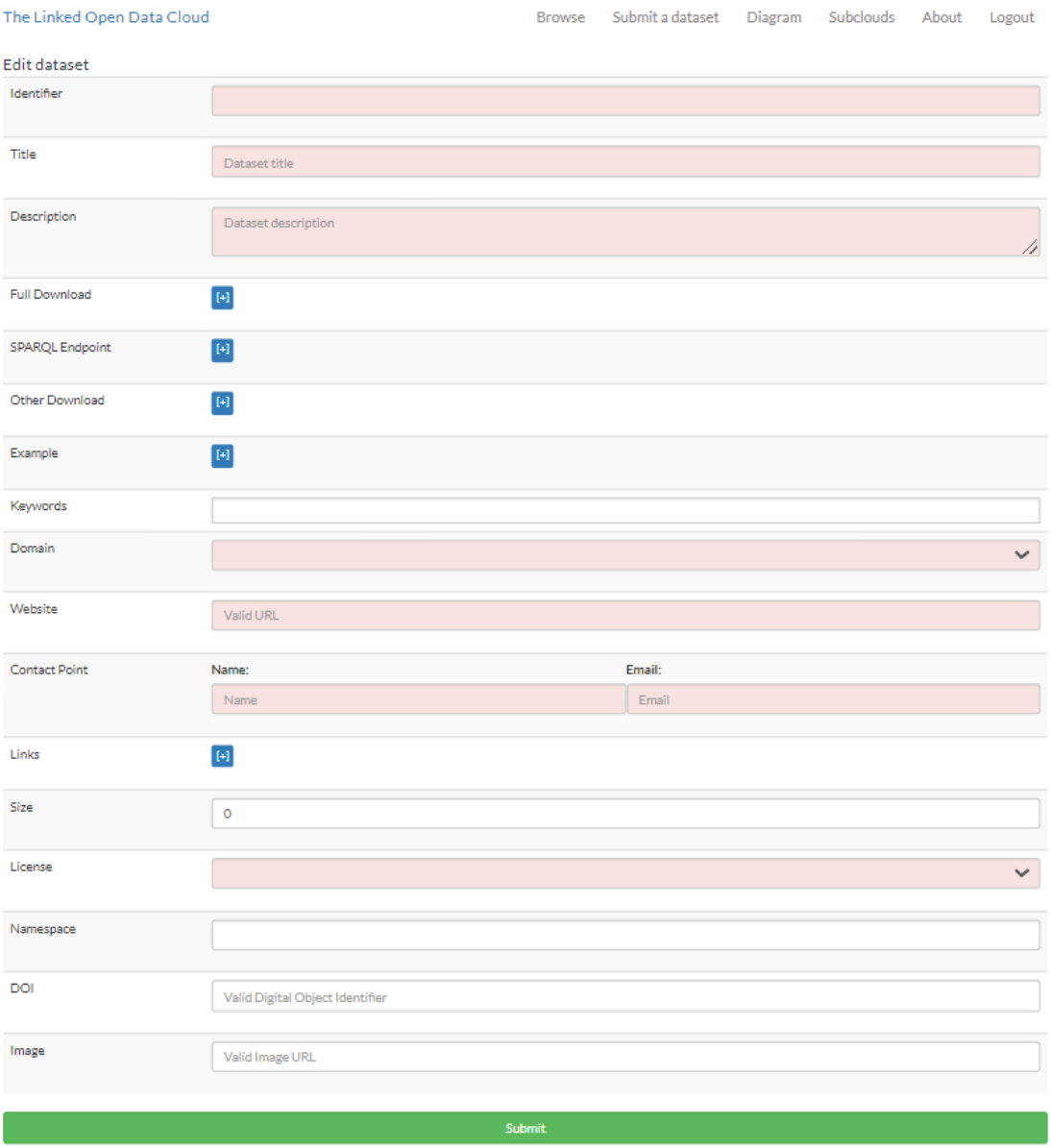
[25] Usp. World Wide Web Consortium. TaskForces/CommunityProjects/LinkingOpenData/DataSets/CKANmetainformation. URL: https://www.w3.org/wiki/TaskForces/CommunityProjects/LinkingOpenData/DataSets/CKANmetainformation (2021-08-09)

[26] Datahub. URL: https://old.datahub.io/ (2021-09-01)

[27] Usp. The linked open data cloud. URL: https://lod-cloud.net/add-dataset (2021-08-12)

[28] Usp. The linked open data cloud. URL: https://lod-cloud.net/#about (2021-08-12)

[29] Usp. Lod-a-lot. URL: http://lod-a-lot.lod.labs.vu.nl/ (2021-08-12)



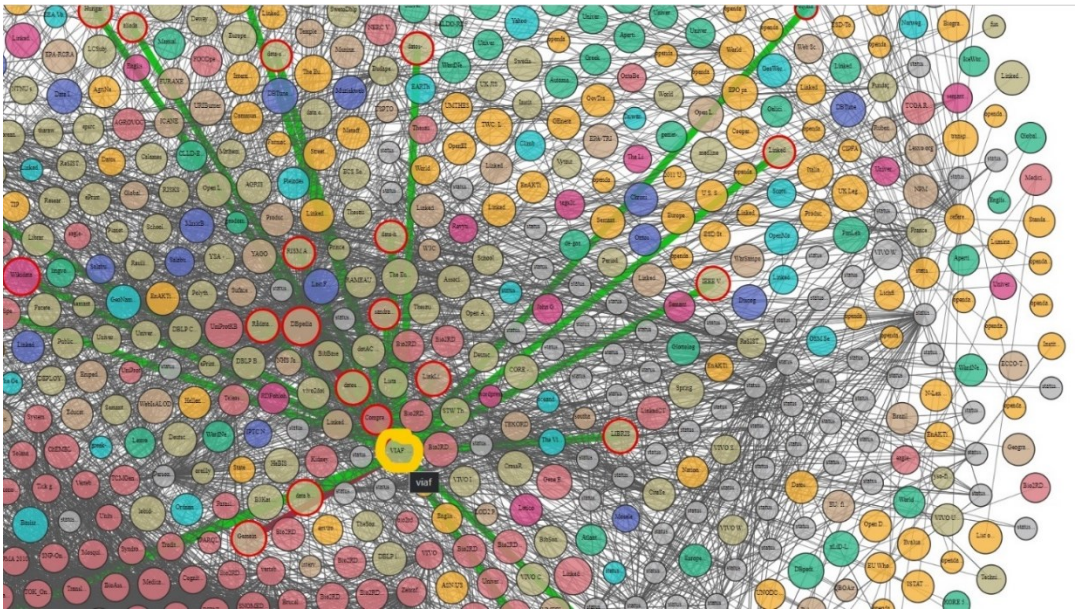
Slika 3 – Obrazac za dodavanje seta podataka LOD oblaku na službenoj stranici.

Setove podataka čine tripleti, koji se nazivaju i RDF izjavama, a koje se pak sastoje od izvora, svojstva i vrijednosti svojstva.[30] Jednostavnije prikazano, RDF izjava je rečenica koja se sastoji od subjekta, predikata i objekta, što ujedno čini gramatički oblik RDF modela podataka. U slučaju grafičkog prikaza RDF dokumenta, element izvora prikazuje se u ovalnom obliku, element svojstva strelicom, a vrijednost svojstva u obliku pravokutnika, ako je riječ o slovnoj vrijednosti.

[30] Usp. Powell, A. Encoding DC in (X)HTML, XML and RDF. URL: http://www.ukoln.ac.uk/metadata/presentations/ecdl-2004/dctutorial/tutorial_files/v3_document.htm (2021-08-14)

Ukoliko se radi o neslovnoj vrijednosti onda je vrijednost svojstva prikazana također u ovalnom obliku.[31]

Na primjeru projekta VIAF (*Virtual International Authority File*), virtualne međunarodne normativne baze seta podataka iz knjižničarske domene, bliže će biti pojašnjena veza između setova podataka. Ulaskom u skalabilnu vektorsku grafiku LOD oblaka, omogućeno je kretanje među setovima podataka koji nosi nazive organizacija i projekata koji iza njih stoje. Klikom na pojedini set podataka u novom prozoru prikazuju se osnovni podaci o istom.[32] Slika 4 prikazuje žutom bojom označen VIAF set podataka, a zelenom njegove poveznice s drugim setovima podataka. Iako se čini kako je poveznica malo, zapravo je taj broj mnogo veći.



Slika 4 - Prikaz VIAF seta podataka u LOD oblak dijagramu.

Odabirom VIAF seta podataka, stranica LOD oblaka daje uvid u njegove osnovne podatke (Slika 5). Na Slici 5 može se vidjeti da se VIAF set podataka sastoji od ukupno 200 milijuna tripleta, i da je najveći broj poveznica ostvario sa setom podataka njemačke knjižničarske zajednice, čak 4 milijuna.

[31] Usp. RDF Primer: W3C Recommendation 10 February 2004. URL: https://www.w3.org/TR/rdf-primer/ (2021-08-14)
[32] Usp. The linked open data cloud. URL: https://lod-cloud.net/versions/2021-05-05/lod-cloud.svg (2021-08-14)

Slika 5 - Prikaz podataka o VIAF setu podataka.

Primjer jednog tripleta iz seta podataka u RDF/XML formatu prikazan je u nastavku. U navedenom primjeru, radi se o književnom djelu Macbeth čiji je autor William Shakespeare. Subjekt u ovom tripletu je djelo Macbeth, prikazano kao poveznica na kontrolni broj Kongresne knjižnice http://id.loc.gov/authorities/names/n82011242. Predikat je autor, a prikazuje se Dublin Core metapodatkovnim svojstvom http://purl.org/dc/terms/creator. Naposljetku, objekt je autor William Shakespeare opisan poveznicom na VIAF, http://viaf.org/viaf/96994048[33] (Slika 6).

```
<?xml version="1.0"?>
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
         xmlns:dc="http://purl.org/dc/terms/">
  <rdf:Description rdf:about="http://id.loc.gov/authorities/names/n82011242">
    <dc:creator rdf:resource="http://viaf.org/viaf/96994048"/>
  </rdf:Description>
</rdf:RDF>
```

Slika 6 - Prikaz tripleta u RDF/XML dokumentu.

S obzirom da se radi o povezanim podacima, objekt će u većini slučajeva biti novi izvor, pa se u grafičkom prikazu, umjesto pravokutnim oblikom, naznačuje ovalnim oblikom, kao i subjekt. (Slika 7)

Slika 7 - Grafički prikaz RDF/XML dokumenta.

[33] Usp. Blaney, Jonathan. Introduction to the principles of linked open data, 2020. URL: https://programminghistorian.org/en/lessons/intro-to-linked-data#rdf-and-data-formats (2021-08-14)

3.3. Problematika LOD oblaka

U teoriji, otvorenost i povezanost podataka postiže se jednostavnim praćenjem načela, ispunjavanjem uvjeta i sl. Međutim, u praksi je to malo složenije, stoga se LOD oblak u najvećem broju slučajeva susreće s problemima vezanim upravo uz otvorenost i povezanost podataka. Pojedine domene unutar LOD oblaka su bitno ograničene po pitanju otvorenosti podataka. S jedne strane, imamo DBpediju, koja svoje setove podataka gradi na temelju Wikipedije, te ima jednostavne podatke čija otvorenost ne ovisi o privatnim osobama i javnim tvrtkama, a s druge, mrežnu stranica s podacima vlade Sjedinjenih Američkih Država (https://www.data.gov/) koja se susreće s preprekama poput dobivanja otvorene licence za podatke, razvijanja metapodatkovnih standarda i sl. te onom koja se čini gotovo nepremostivom, ali i najvažnijom - omogućavanje transparentnosti podataka uz očuvanje privatnosti korisnika.[34] To se, recimo, odnosi na bankovne i medicinske podatke čija je upotreba, sigurnost i privatnost na taj način ugrožena, a zloupotreba gotovo sigurna.[35] Iako je jednostavno objaviti novi set podataka u RDF formatu, upravo tu nastupaju izazovi: povezanost podataka je teško održavati, tripleti postaju nevjerodostojni, sintakse pogrešne. S obzirom na to da prilikom dodavanja novih setova podataka neki podaci nisu važni, većina setova nema dovoljno povezanih metapodataka koji bi u suprotnom bili od velike pomoći za slučajna otkrivanja relevantnih informacija prilikom pretraživanja.[36] No to nisu jedini problemi LOD zajednice, pouzdanost tj. kvaliteta poveznica također muči korisnike kod pretraživanja. Kvaliteta povezanih podataka ogleda se u korisnosti podataka u skladu sa setovima podataka. Sukladnost podataka i setova podataka je dobra ukoliko RDF set podataka prati sva načela povezanih podataka. Set podataka kvalitetniji je ukoliko nema pogrešnih podataka, grešaka u sintaksi, i nedostupnih SPARQL krajnjih pristupnih točaka za pretraživanje ili pogrešno spojene poveznice.[37] U istraživanju iz 2016. godine na temelju poveznica među setovima podataka iz 2014. godine, ustanovljeno je 62% neprovjerenih, 35% provjerenih i 3% mrtvih poveznica u LOD oblaku.[38] Razlog tomu nalazi se u neprestanom stvaranju novih poveznica, gdje jedna poveznica među dvjesto drugih postaje nedostupna na tjednoj razini. Važno je osigurati minimalnu kvalitetu i pouzdanost poveznica, jer se sve više aplikacija gradi na setovima podataka, a nedostupne

[34] Usp. Fayyaz, Nosheen; Ullah, Irfan; Khusro, Shah. On the current state of linked open data : Issues, challenges and future directions. // International journal on semantic web and information systems, 14, 4(2018), str. 113. URL: https://www.researchgate.net/publication/323969399_On_the_Current_State_of_Linked_Open_Data_Issues_Challenges_and_Future_Directions (2021-08-16)
[35] Isto.
[36] Isto, str. 112.
[37] Isto, str. 116.
[38] Usp. Neto, Ciro Baron… [et. al.] Assessing quantity and quality of links between link data datasets, 2016. URL: http://ceur-ws.org/Vol-1593/article-07.pdf (2021-08-16)

poveznice mogu širiti greške kroz aplikacije.[39] Važna karakteristika LOD oblaka njegova je održivost. Tako je između 2014. i 2017. godine, zbog povećanja povezanih setova podataka temeljenih na RDF-u, LOD oblak zastario. Kako se to ne bi događalo važno je osigurati stalno ažuriranje setova podataka, ali i metapodataka u samom LOD oblaku, jer setove podataka ažurira korisnik koji je objavio podatke, a metapodatke održavatelji LOD oblaka. Bilo kako bilo, setovi podataka postaju zastarjeli, te korisnici zbog toga imaju poteškoća kod njihova pretraživanja.[40] Korisnici povezanim podacima mogu pristupiti na dva načina: putem LOD preglednika ili LOD tražilica. LOD preglednici omogućuju pregledavanje mreže podataka kroz RDF poveznice dok LOD tražilice prikupljaju, integriraju i pretražuju srodne upite, slično kao Google. Međutim, kod pretraživanja korisnik nailazi na poteškoće zbog nepoznavanja načina funkcioniranja semantičkog web-a, RDF-a, SPARQL-a i sl., a time je onemogućeno kvalitetno pretraživanje i pronalazak željenog sadržaja.[41] Treba uzeti u obzir da LOD oblak preuzima sve povezane podatke, pa tako i tok podataka senzora, npr. GPS-a i društvenih mreža čiji su podaci dinamični za razliku od drugih setova podataka koji jednom postavljeni u LOD oblak ne zahtijevaju često ažuriranje. Sve u svemu, potrebno je u stvarnom vremenu pratiti nove nadolazeće podatke te odrediti njihovu učinkovitost, definirati podrijetlo, izvući poveznice, prepoznati pripadnu ontologiju kojom su podaci opisani i ponovno objaviti podatke.[42]

U novom razdoblju društvenih mreža zbog sve većeg broja korisnika, broj objava i dijeljenja podataka je u značajnom porastu. Kada se radi o LOD oblaku i dodavanju dodatnih semantičkih informacija, ono je omogućeno povezivanjem društvenih mreža, koje sadrže navedene dodatne informacije, i LOD oblaka. Pored problema dinamičnih podataka LOD oblak se susreće i s problemima privatnosti korisnika, poput podataka s data.gov domene. Ukoliko bi podaci s društvenih mreža postali dijelom LOD oblaka, bili bi nepotpuni i neprikladni za organiziranje, jer bi se isključile privatne informacije kao što su identifikacija korisnika, zaštićeni razgovori i sl. Samim time bi preuzimanje novosti vezanih uz katastrofe, događaje, uzbune i sl. bile izvan konteksta, s nedostatkom dinamičnog rječnika koji bi ih bio u mogućnosti opisati.[43] Koristeći se novim tehnologijama, trenutni problemi mogli bi ubrzo pronaći svoja rješenja, ali i stvoriti nove prepreke u radu.

[39] Isto.
[40] Usp. Debattista, Jeremy… [et. al.] Is the LOD cloud at risk of becoming a museum for datasets : looking ahead towards a fully collaborative and sustainable LOD cloud, 2019. URL: http://doras.dcu.ie/24657/1/main.pdf (2021-08-17)
[41] Usp. Fayyaz, Nosheen; Ullah, Irfan; Khusro, Shah. Nav. dj., str. 115.
[42] Isto, str. 114.
[43] Isto, str. 114.-115.

3.4. Budućnost LOD oblaka

Uzimajući u obzir sve nedostatke LOD oblaka i prepreke na koje u svom razvoju nailazi, budućnost istog ovisi o rješenjima tih problema i novim idejama koje će se temeljiti na tehnologiji semantičkog weba. Setovi podataka koji na mreži više ne postoje, a samim time nisu od koristi, zadržavaju se u LOD oblak dijagramu. Kako LOD oblak ne bi postao muzej za setove podataka koji više ne postoje, potrebno ih je uklanjati koristeći se *LOD laundromat*-om.[44] *LOD laundromat* predstavlja platformu koja čisti, usklađuje i ponovno objavljuje podatke iz setova podataka u LOD oblaku, a sve to u razdoblju od samo nekoliko dana.[45] Uz mogućnost da jednog dana završi kao muzej nepostojećih setova podataka, još jedna prijetnja se nadvila nad LOD oblakom. Riječ je o nepoštivanju načela povezanih podataka, kao i nedosljednosti u korištenju ontologija, sintaksi i sl. Nabrojane prijetnje mogu dovesti do pada ugleda LOD oblaka u široj zajednici i prestanku njegova korištenja. Zbog toga bi LOD oblak trebao biti neprestano aktivan te ažuran u uklanjanju onih korisnika koji tijekom određenog vremena nisu bili dostupni niti su ažurirali i nadopunili svoje setove podataka. Nadalje, veličina setova podataka od 2007. godine u značajnoj mjeri je narasla, povećao se i broj udvostručenih tripleta odnosno tripleta duplikata koji onda bespotrebno zauzimaju prostor i usporavaju upite i preuzimanja. Za rješenje tog problema preporuča se korištenje *dump* datoteka (set podataka podijeljen u više datoteka) u HDT obliku dokumenta i komprimiranih RDF formata.[46] Bilo kako bilo, sami korisnici organizirani u zajednicu jedino mogu održati LOD oblak na životu, stvaranjem kvalitetnih povezanih podataka na temelju načela povezanih podataka.[47] Ukoliko LOD oblak preživi, postoje prijedlozi da se uvedu i neke nove usluge. Tako bi se u budućnosti LOD oblaka, jedna od usluga za kreatore setova podataka odnosila na slanje obavijesti o objavi povezanih podataka uz mogućnost pohranjivanja recenzija setova podataka poput komentara, glasanja i sl. Ta usluga mogla bi se provesti koristeći se platformom povezanih podataka koja dopušta slanje obavijesti povezanih podataka.[48] Platformu povezanih podataka činio bi skup pravila za HTTP operacije na mrežnim izvorima, kako bi se osigurala arhitektura za čitanje i pisanje povezanih podataka na mreži.[49] Arhitektura ove usluge zamišljena je da počiva na sposobnosti otkrivanja, razumijevanja i socijalnog povezivanja. Socijalno povezivanje odnosi se na uslugu LOD oblaka koja dopušta svima da sudjeluju u razvijanju LOD oblaka. Sposobnost otkrivanja odnosi se na davanje usluge pretraživanja, istraživanja i

[44] Usp. Debattista, Jeremy… [et. al.] Nav. dj.
[45] Usp. Semantics. LOD laundromat. URL: https://2018.semantics.cc/lod-laundromat (2021-08-17)
[46] Usp. Polleres, Axel… [et. al.] A more decentralized vision for linked data. // Semantic Web, 11, 1(2020), str. 4-6. http://semantic-web-journal.net/system/files/swj2308.pdf (2021-08-19)
[47] Isto, str. 9.
[48] Usp. Debattista, Jeremy… [et. al.] Nav. dj.
[49] World Wide Web Consortium. Linked data platform 1.0. URL: https://www.w3.org/TR/ldp/ (2021-08-19)

identificiranja skupova svim korisnicima na najbrži i najučinkovitiji način. Konačno, sposobnost razumijevanja odnosi se na pružanje izlaznih podataka u RDF modelu s obzirom da je LOD oblak usluga građena na interoperabilnom semantičkom sloju različitih taksonomija.[50] U praksi bi kreator objavio set podataka te bi putem nove usluge mogao postaviti obavijest o navedenom setu podataka na LOD oblaku zajedno s poveznicom na isti, a nakon toga bi usluga LOD oblaka pronašla izvor te odobrila poslani set podataka. Zatim bi se URI/IRI seta pohranio u popisu, a grafički oblik prikazao u dijagramu LOD oblaka. Samim time bi konzumenti podataka i pružatelji usluga mogli izravno komunicirati s korisnikom i kreatorom seta podataka putem dereferenciranog URI/IRI-ja metapodataka koji su pohranjeni u LOD oblaku. Svaki komentar kreatora ili konzumenta seta podataka pohranili bi se u *triplestore* uslugu LOD oblaka, a onda bi korisnici mogli sortirati i pretraživati setove po željenim kriterijima i također ostavljati svoje komentare.[51]

Kao što je spomenuto, jedan od uočenih problema u razvoju LOD oblaka odnosio se na povezivanja podataka s društvenih mreža s LOD oblakom. Ukoliko bi se omogućilo ostavljanje povratnih informacija konzumenata podataka, bilo bi moguće i spajanje društvenih mreža s LOD oblakom čiji bi podaci, iako zaštićeni radi privatnosti, mogli pružati veći kontekst podacima u setu podataka. Također, potrebno je pronaći brži način pregledavanja dinamičnih protoka podataka za potpuno povezivanje. Možda bi dodavanje društvenih mreža kao domena LOD oblaka, moglo inicirati povezivanje cijelog Interneta u jednu smislenu točku.

[50] Usp. Debattista, Jeremy… [et. al.] Nav. dj.
[51] Isto.

4. Zaključak

Svrha ovog rada bila je prikazati LOD oblak i objasniti koncept povezanih otvorenih podataka na temelju njihovih načela. Također, svrha rada je bila i objasniti i prikazati strukturu LOD oblaka, njegove setove podataka i triplete na primjeru seta podataka iz knjižničarske struke. Tijekom svog razvoja LOD oblak nailazio je na poteškoće koji su u radu opisani, ali su opisana i moguća rješenja tih problema koja bi se mogla primijeniti u budućem razvoju LOD oblaka. Kao moguća budućnost LOD oblaka predstavljene su između ostalog i moguće nove usluge LOD oblaka.

Na TED konferenciji 2009. godine, Tim Berners-Lee rekao je kako su povezani podaci ideja vrijedna širenja, danas je ta ideja postala uvjet prilikom objavljivanja svake informacije na mreži. Iako je u to vrijeme LOD oblak već dvije godine stvarao svoj LOD oblak dijagram povezanih setova podataka, ono što je postao danas rezultat je korištenja četiriju načela povezanih podataka. Kao mali LOD oblak dijagram nastao u projektu DBpedije, kasnije i Bio2RDF-a, LOD zajednica je koristeći se tehnologijama semantičkog weba uspjela u svom naumu stvaranja povezane mreže. Tijekom godina, LOD oblak se oslanjao na inovativnost LOD zajednice koja je doprinijela, primjerice, stvaranju domena područja setova podataka za potrebe lakšeg snalaženja korisnika. Skalabilnu vektorsku grafiku vrlo je rano implementirao u svom razvoju i omogućio interaktivnost LOD oblak dijagrama. Od 12 do 1031 seta podataka bilo je potrebno puno objavljivanja, stvaranja dokumenata na temelju načela povezanih podataka kako bi se osiguralo da setovi podataka u LOD oblak dijagramu dobiju „pet zvjezdica“ u implementacijskoj shemi Pet zvjezdica otvorenih podataka Tim Berners-Lee-ja. No, katkad se takva načela ne poštuju pa dovode u pitanje otvorenost, održivost, pa tako i povezanost LOD oblaka. LOD oblak se također u svom razvijanju susreće s problematikom otvorenosti podataka i njihovom kvalitetom, te otežanim korištenjem. Premda je budućnost LOD oblaka krhka, rješavanjem problema i korištenjem novih tehnologija semantičkog weba LOD oblak može postati kvalitetno mjesto za namjerno ili slučajno pronalaženje odgovora željenih upita. Ostalo su samo pretpostavke, pa preostaje očekivati nove ideje vezane uz povezane otvorene podatke i njihovo korištenje, a samim time i neke inovacije koje će se moći iskoristiti za daljnji razvoj LOD oblaka.

Literatura

1. Berners-Lee, Tim. Linked data, 2009. URL: https://www.w3.org/DesignIssues/LinkedData.html (2021-05-20)
2. Berners-Lee, Tim. The next web. TED Talks. Edinburgh, 2009. [Predavanje] URL: https://www.ted.com/talks/tim_berners_lee_the_next_web/reading-list (2021-08-23)
3. Best NoSQL Databases Software. URL: https://www.g2.com/categories/nosql-databases (2021-05-20)
4. Blaney, Jonathan. Introduction to the principles of linked open data, 2020. URL: https://programminghistorian.org/en/lessons/intro-to-linked-data#rdf-and-data-formats (2021-08-14)
5. DBpedia. URL: https://www.dbpedia.org/about/ (2021-08-03)
6. Debattista, Jeremy… [et. al.] Is the LOD cloud at risk of becoming a museum for datasets : looking ahead towards a fully collaborative and sustainable LOD cloud, 2019. URL: http://doras.dcu.ie/24657/1/main.pdf (2021-08-17)
7. Dumontier, Michel. Bio2rdf, 2017. URL: https://github.com/bio2rdf/bio2rdf-scripts/wiki (2021-08-03)
8. European environment agency. SPARQL endpoint, 2017. URL: https://data.europa.eu/euodp/en/data/dataset/european-environment-agency-sparql-endpoint (2021-08-09)
9. Farago, Filip; Bosančić, Boris. Povezani podaci i knjižnice. // Vjesnik bibliotekara Hrvatske 54, 4(2013), str. 25-52. URL: https://hrcak.srce.hr/142376 (2021-08-08)
10. Fayyaz, Nosheen; Ullah, Irfan; Khusro, Shah. On the current state of linked open data : Issues, challenges and future directions. // International journal on semantic web and information systems, 14, 4(2018), str. 110-128. URL: https://www.researchgate.net/publication/323969399_On_the_Current_State_of_Linked_Open_Data_Issues_Challenges_and_Future_Directions (2021-08-16)
11. Fileformat. URL: https://docs.fileformat.com/spreadsheet/csv/ (2021-05-20)
12. Idehen, Kingsley Uyi. What is the linked open data cloud, and why is it important, 2019. URL: https://medium.com/virtuoso-blog/what-is-the-linked-open-data-cloud-and-why-is-it-important-1901a7cb7b1f (2021-08-03)
13. Lod-a-lot. URL: http://lod-a-lot.lod.labs.vu.nl/ (2021-08-12)
14. Neto, Ciro Baron… [et. al.] Assessing quantity and quality of links between link data datasets, 2016. URL: http://ceur-ws.org/Vol-1593/article-07.pdf (2021-08-16)

15. Open link software. Linked open data (LOD) cloud. URL: https://www.openlinksw.com/describe/?url=http%3A%2F%2Fdata.openlinksw.com%2Foplweb%2Fglossary-term%2FLODCloud%23this&graph=urn%3Adata%3Aopenlink%3Aglossary (2021-08-03)
16. Polleres, Axel… [et. al.] A more decentralized vision for linked data. // Semantic Web, 11, 1(2020), str. 4-6. http://semantic-web-journal.net/system/files/swj2308.pdf (2021-08-19)
17. Powell, A. Encoding DC in (X)HTML, XML and RDF. URL: http://www.ukoln.ac.uk/metadata/presentations/ecdl-2004/dctutorial/tutorial_files/v3_document.htm (2021-08-14)
18. RDF Primer: W3C Recommendation 10 February 2004. URL: https://www.w3.org/TR/rdf-primer/ (2021-08-14)
19. Semantics. LOD laundromat. URL: https://2018.semantics.cc/lod-laundromat (2021-08-17)
20. Siebes, Ronald…[et al.] Top 10 fair data & software things : Linked open data. URL: https://librarycarpentry.org/Top-10-FAIR/2019/09/05/linked-open-data/ (2021-05-20)
21. The linked open data cloud. URL: https://lod-cloud.net/ (2021-08-06)
22. What is five-star linked open data? URL: https://www.ontotext.com/knowledgehub/fundamentals/five-star-linked-open-data/ (2021-05-20)
23. World Wide Web Consortium. Linked data. URL: https://www.w3.org/wiki/LinkedData (2021-05-20)
24. World Wide Web Consortium. Linked data platform 1.0. URL: https://www.w3.org/TR/ldp/ (2021-08-19)
25. World Wide Web Consortium. SweoIG/TaskForces/CommunityProjects/LinkingOpenData. URL: https://www.w3.org/wiki/SweoIG/TaskForces/CommunityProjects/LinkingOpenData#LOD_Community_Gatherings (2021-08-08)
26. World Wide Web Consortium. TaskForces/CommunityProjects/LinkingOpenData/DataSets/CKANmetainformation. URL: https://www.w3.org/wiki/TaskForces/CommunityProjects/LinkingOpenData/DataSets/CKANmetainformation (2021-08-08)