Univerzitet u Beogradu

Fakultet organizacionih nauka

Seminarski rad iz predmeta

**Inteligentni sistemi**

Tema: Kreiranje RDF repozitorijuma transformacijom postojećih podataka

|  |  |
| --- | --- |
| **Student**:  Lazar Deretić 7/09/I | **Profesor**:  dr Jelena Jovanović  **Mentor**:  Nikola Milikić |

Beograd, 2013.

**Sadržaj**

[1. Opis problema 2](#_Toc360048585)

[2. Domenski model 3](#_Toc360048586)

[3. Rešenje 5](#_Toc360048587)

[4. Tehnička realizacija 7](#_Toc360048588)

# 1. Opis problema

Cilj ovog rada je kreiranje aplikacije koja transformiše u RDF format podatke o pozivima za publikovanje radova (eng. *Call for Papers* -­‐ CfP[[1]](#footnote-1)). Ovi podaci su raspoloživi iz različitih izvora, s tim što su se za potrebe ovog projektnog zadatka koristili CfP podaci raspoloživi sa WikiCfP sajta[[2]](#footnote-2). Konkretno, potrebno je iskoristiti podatke o (preko 5000) CfPs koje ovaj sajt čini dostupnim u XML formatu[[3]](#footnote-3). Transformisani podaci bi trebalo da budu predstavljeni u formi RDF tripleta, a na osnovu RDF vokabulara Call Ontology. Za ovaj vokabular je raspoloživ dijagram[[4]](#footnote-4) i owl fajl[[5]](#footnote-5) u kome je vokabular definisan.

Radi kreiranja aplikacije, potrebno je bilo ispuniti sledeće zahteve:

* analizirati oblik (model) u kome su podaci raspoloživi;
* kreirati mapiranja (grafički predstavljena) između modela/oblika u kojem su podaci izvorno raspoloživi i RDF vokabulara Call Ontology;
* na osnovu definisanih mapiranja, izvršiti transformaciju raspoloživih podataka u odgovarajući RDF oblik i tako transformisane podatke smestiti u RDF bazu;
* omogućiti pristup podacima u bazi preko odgovarajućih REST servisa i SPARQL Endpoint-­‐a.

# 2. Domenski model

Podaci o pozivima za publikovanje radova izvorno su raspoloživi u XML formatu. Primer izgleda jednog takvog XML fajla dat je na slici 1.

<?xml version="1.0"?>

<resultset>

<row>

<field name="eventid">7806</field>

<field name="createdate">2010-01-01 00:00:00</field>

<field name="fullname"> 8th International Airship Convention - Bedford, England - 29th Sept/3rd Oct 2010

</field>

<field name="handle">AIRSHIPS</field>

<field name="year">2010</field>

<field name="location">Bedford, England</field>

<field name="begindate">2010-09-29</field>

<field name="finishdate">2010-10-03</field>

<field name="presubdate">2010-04-30</field>

<field name="submitdate">2010-04-30</field>

<field name="notifydate">2010-05-30</field>

<field name="cameradate">2010-07-15</field>

<field name="weblink"> http://www.airship- association.org/2008web/call.pdf

</field>

<field name="info"> ...... </field>

</row>

<row>

...

</row>

...

</resultset>

**Slika 1. Primer XML fajla sa podacima o CfP**

Podatke u izvornom obliku potrebno je predstaviti, odnosno mapirati u odgovarajuće elemente RDF vokabulara Call Ontology. Grafički prikaz vokabulara kao i mapiranja između njega i podataka u izvornom obliku prikazani su na slici 2. Crvenom bojom su označeni atributi iz izvornog XML fajla, dok crvene strelice ukazuju na mapiranja.



Slika 2. Grafički prikaz vokabulara Call Ontology i mapiranja izvornih podataka

Klasa *Event* odnosi se na događaj (npr. konferenciju, kongres itd.) koji prati konkretan poziv za publikovanje radova. Ima atribute kao što su datum početka, datum završetka. naziv, opis, mesto održavanja, kao i referencu ka samom CfP-u. Ostali atributi nisu značajni sa aspekta ovog rada.

Klasa *Place* opisuje lokaciju na kojoj se održava događaj i ima atribut ime.

Klasa *Call* opisuje poziv za publikovanje radova atributima datum abstrakta, datum predaje konačne verzije rada, datum donošenja konačne odluke i datum štampanja objavljenih radova. Ima i referencu na događaj koji prati CfP.

Vokabular Call Ontology svojim sadržajem u velikoj meri odgovara postavljenom problemu mapiranja. Međutim, pojedini podaci nisu adekvatno predstavljeni istim. Stoga, potrebno je bilo iskoristiti neke od drugih postojećih vokabulara i predstaviti problematične podatke njihovim elementima. Tako je atribut *createdate* predstavljen pomoću elementa *created* vokabulara DC (Dublin Core[[6]](#footnote-6)). Iz istog vokabulara su iskorišćeni elementi *alternative*, *date* za mapiranje atributa *handle* i *year*, respektivno. Atribut *weblink* je predstavljen elementom *homepage* vokabulara FOAF (Friend Of A Friend[[7]](#footnote-7)).

# 3. Rešenje

U skladu sa postavljenim problemom i specifičnim zahtevima, kreirana je aplikacija koja omogućava kreiranje repozitorijuma transformacijom postojećih podataka.

Prvi korak je transformisanje podataka u XML formatu u odgovarajuće RDF triplete, na osnovu RDF vokabulara i modela sa slike 2. Podaci se potom smeštaju u lokalnu RDF bazu.

Aplikacija omogućava pristup sačuvanim podacima korišćenjem RESTful servisa. Podržana su dva servisa, jedan koji prikazuje sve podatke vezane za određene CfP-ove, i drugi koji prikazuje podatke samo o događajima koji prate CfP-ove. Specifikacija REST servisa data je u nastavku.

* **GET /api/cfp** - servis koji vraća podatke o pozivima za publikovanje radova (CfP). Opcioni parametri su:

- ***abstractFrom*** - donja granica datuma (roka) za predaju abstrakta

- ***abstractTo*** - gornja granica datuma (roka) za predaju abstrakta

- ***fullSubmissionFrom*** - donja granica datuma (roka) za predaju kompletnih radova

- ***fullSubmissionTo*** - gornja granica datuma (roka) za predaju kompletnih radova

- ***finalDecisionFrom*** - donja granica datuma (roka) za donošenje odluke o prihvatanju rada

- ***finalDecisionTo*** - gornja granica datuma (roka) za donošenje odluke o prihvatanju rada

- ***startDateFrom*** - donja granica datuma početka događaja

- ***startDateTo*** - gornja granica datuma početka događaja

- ***endDateFrom*** - donja granica datuma završetka događaja

- ***endDateTo*** - gornja granica datuma završetka događaja

- ***location*** - lokacija održavanja događaja

- ***name*** - naziv događaja

- ***limit*** - ograničenje broja vraćenih rezultata

- ***offset*** - pomeraj

Primer poziva REST servisa:

GET /api/cfp?abstractFrom=2010-02-01&abstractTo=2010-09-30 &location=Italy&startDateFrom=2010-09-01&limit=20

* **GET /api/cfp/events** - servis koji vraća podatke o događajima koji prate pozive za publikovanje radova. Opcioni parametri su:

- ***startDateFrom*** - donja granica datuma početka događaja

- ***startDateTo*** - gornja granica datuma početka događaja

- ***endDateFrom*** - donja granica datuma završetka događaja

- ***endDateTo*** - gornja granica datuma završetka događaja

- ***location*** - lokacija održavanja događaja

- ***name*** - naziv događaja

- ***limit*** - ograničenje broja vraćenih rezultata

- ***offset*** - pomeraj

Primer poziva REST servisa:

GET /api/cfp/events?location=Russia&startDateFrom=2010-06-01& startDateTo=2010-09-01&name=computer&limit=20

# 4. Tehnička realizacija

Aplikacija je rađena u programskom jeziku Java, u razvojnom okruženju Eclipse. Radi transformisanja izvornih podataka izvršeno je parsiranje XML dokumenta korišćenjem *dom4j[[8]](#footnote-8)* biblioteke. Ona omogućava laku manipulaciju XML, XPath i XSLT dokumentima i pruža punu podršku za DOM, SAX (korišćen u aplikaciji) i JAXP. Parsiranje se vrši tako što se iterira kroz dokument i za svaki element unutar korenog elementa proveravaju se vrednosti atributa njegove dece. Na osnovu tih atributa, postavljaju se vrednosti odgovorajaćih polja klasa.

U daljem radu korišćen je *Jena* *framework*[[9]](#footnote-9), koji predstavlja Java framework i pruža skup alata i biblioteka za izradu aplikacija semantičkog veba. U okviru tog skupa nalazi se API za čitanje, obradu i upisivanje RDF podataka u XML, N-triplet i Turtle formatu; mehanizme koji omogućavaju da se veliki broj RDF tripleta efikasno sačuva na disku kao i query engine koji je u skladu sa najnovijom SPARQL specifikacijom.

Za čitanje i upisivanje RDF podataka korišćena je *Jenabean*[[10]](#footnote-10) biblioteka. Ona rešava problem prevođenja RDF tripleta u Java objekte i obrnuto. Prevođenje se zasniva na anotacijama, koji mapiraju i povezuju polja Java klase i RDF propertije.

Za skladištenje podataka u RDF repozitorijum korišćena je *TDB*[[11]](#footnote-11) komponenta Jena framework-a, koja, pored skladištenja, omogućava i izvršavanje različitih vrsta SPARQL upita nad podacima u RDF formatu. TDB omogućava i pojedine napredne mehanizme, poput transakcija, koji se mogu koristiti prilikom manipulacije podacima.

Pristup podacima je omogućen preko RESTful servisa, kao što je opisano u prethodnom poglavlju. Za rad sa servisima iz Java-e korišćen je *Jersey*[[12]](#footnote-12)framework koji predstavlja implementaciju JAX-RS, Java-inog API-a za RESTful veb servise. JAX-RS koristi anotacije radi pojednostavljenja razvoja i isporuke servisa i endpoint-a. Na osnovu anotacija, određene klase su predstavljene kao resursi, koji sadrže anotacijama označene metode koje prihvataju i obrađuju HTTP zahteve (GET, POST...). Aplikacija na ovaj način isporučuje podatke (u JSON formatu) klijentima, na osnovu različitih parametara koje oni proslede preko HTTP zahteva, a u skladu sa specifikacijom servisa datom u prethodnom poglavlju.

1. <http://en.wikipedia.org/wiki/Call_for_papers> [↑](#footnote-ref-1)
2. <http://www.wikicfp.com/cfp/> [↑](#footnote-ref-2)
3. Podaci se mogu preuzeti sa stranice: <http://www.wikicfp.com/cfp/data.jsp> [↑](#footnote-ref-3)
4. <https://dl.dropbox.com/u/7030190/donotdelete/inteligentnisistemi/seminarski/Call_ontology_v.2.0.jpg> [↑](#footnote-ref-4)
5. <https://dl.dropbox.com/u/7030190/donotdelete/inteligentnisistemi/seminarski/call_v.2.0.owl> [↑](#footnote-ref-5)
6. <http://dublincore.org/> [↑](#footnote-ref-6)
7. <http://www.foaf-project.org/> [↑](#footnote-ref-7)
8. <http://dom4j.sourceforge.net/> [↑](#footnote-ref-8)
9. <http://jena.apache.org/> [↑](#footnote-ref-9)
10. <https://code.google.com/p/jenabean/> [↑](#footnote-ref-10)
11. <http://jena.apache.org/documentation/tdb/> [↑](#footnote-ref-11)
12. <https://jersey.java.net/> [↑](#footnote-ref-12)