Analiza in predlog dopolnitve informacijskega sistema o raziskovalni dejavnosti s semantično komponento

[1]Aleš Bošnjak, [2]Vili Podgorelec
[1]Institut informacijskih znanosti, IZUM, Prešernova 17, Maribor;
[2]Univerza v Mariboru, Fakulteta za elektrotehniko in računalništvo, Inštitut za informatiko, Smetanova 17, 2000 Maribor
ales.bosnjak@izum.si; vili.podgorelec@uni-mb.si

Izvleček

V članku je predstavljena problematika dostopa do relevantnih informacij v heterogenem okolju informacijskih sistemov o raziskovalni dejavnosti (sistemov CRIS). Analiza je izdelana v okviru skupnega evropskega metapodatkovnega modela CERIF, kakor tudi njegove slovenske izvedenke, modela SICRIS.

V iskanju optimalne rešitve je predlagan model ontološke infrastrukture. Ta je izvedenka modela Ontobroker. Na nekaterih karakterističnih primerih, ki so bili povzeti iz prakse servisa SICRIS, je pokazano, da je uvedba ontologij smiselna tudi v sistemih, ki črpajo informacije iz različnih podatkovnih baz z znanimi metapodatkovnimi modeli. Primeri so bili rešeni z uporabo predlagane ontološke infrastrukture in poenostavljenim modelom ontologije slovenske znanstvenoraziskovalne domene.

Prednost uvedbe se izraža z večjo pokritostjo metapodatkov pri iskanju uporabnikov sistema SICRIS. Z dodatnimi relevantnimi metapodatki, ki so generirani s pomočjo ustreznih ontologij, se lahko pričakuje tudi večja učinkovitost uporabnikov pri iskanju informacij.

Ključne besede: informacijski sistemi o raziskovalni dejavnosti, sistemi CRIS, metapodatkovni modeli ontologije, ontološko podprto iskanje.

Abstract

The Analysis and Upgrading Proposition of IS on Research Activities with a Semantic Component

The article presents the problems of access to relevant information in a heterogeneous environment of information systems on research activities (CRIS systems).

The analysis was carried out within the framework of the joint European metadata model CERIF, as well as its Slovenian version, the SICRIS model. An optimal solution of the ontological infrastructure model is proposed. This is a version of the Ontobroker model. In some typical cases, which were taken from the practice of the SICRIS service, it is shown that the introduction of ontologies is sensible also in the systems which draw information from various data bases with known meta data models. The cases were solved with the use of the proposed ontological infrastructure and a simplified model of the ontology of the Slovenian scientific –research domain.

The advantage of the introduction is expressed by greater coverage of metadata in the user searches of the SICRIS system. With additional relevant metadata, which is generated by means of appropriate ontologies, an increase in user efficiency in finding information can also be expected.

Key words: Current research information systems, CRIS systems, metadata models for ontology, ontological-basedsearch.

1 UVOD

V članku bo predstavljeno širše problemsko področje informacijskih sistemov o raziskovalni dejavnosti, za katere se je v angleškem govornem področju ustalila kratica CRIS – Current Research Information Systems (Jeffery & Asserson, 2006). Med sistemi CRIS se bomo posebej osredinili na problematiko slovenskega sistema SICRIS (SICRIS – informacijski sistem o raziskovalni dejavnosti v Sloveniji, 2012).

Pod drobnogled bomo postavili ustreznost umeščenosti podatkovnega modela slovenskega sistema CRIS v informacijskih tokovih znanstvenoraziskovalne skupnosti. Pri tem bomo poleg metapodatkovnega modela omenjenega sistema upoštevali tudi poslovne procese, ki tečejo v slovenski znanstvenoraziskovalni sferi, ter lastnosti nekaterih povezanih informacijski sistemov.

Pri obstoječem informacijskem modelu SICRIS bomo izpostavili nekaj scenarijev, ki kažejo na pro-

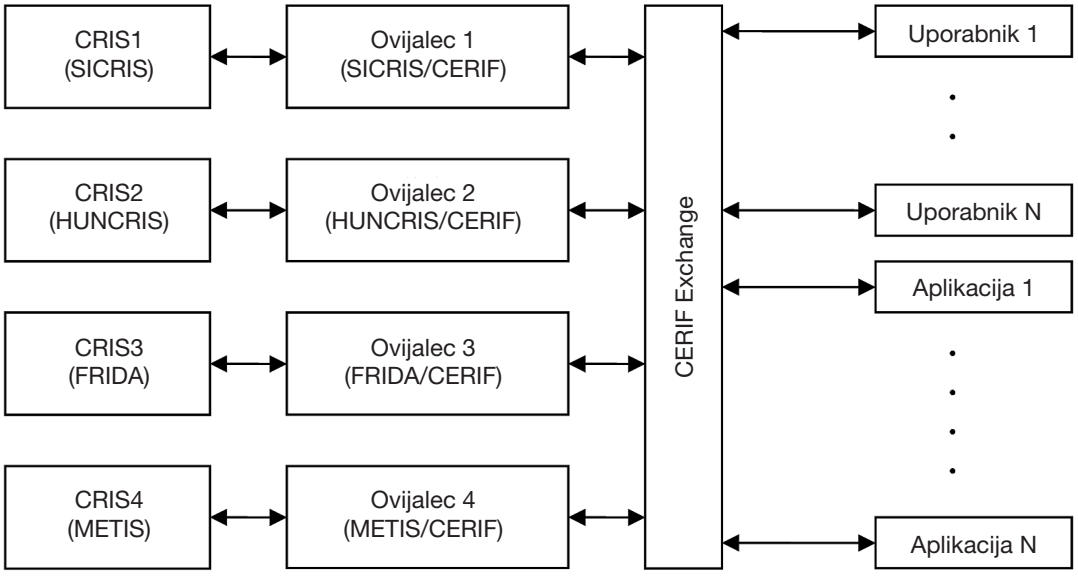
blematičnost dostopa do nekaterih informacij, ki jih potrebujejo uporabniki. Za izboljšanje informacijskega pretoka bomo predlagali uporabo primerne testne ontologije. V članku bo podan tudi predlog ustrezne ontološke infrastrukture. Pri tem bomo skušali dokazati predvsem dvoje trditev:

- trditev 1: z uporabo ontologije se iskanje v bazah podatkov s sicer znano, vendar za uporabnika (vseeno) zelo obsežno metapodatkovno strukturo znatno izboljša v smislu gostejše pokritosti[1] uporabljenih metapodatkov;
- trditev 2: z uvedbo enotne ontologije je mogoče bolj učinkovito kot z uporabo posameznih transformacijskih programov – ovijalcev (s slike 1) zagotoviti interoperabilnost med posameznimi informacijskimi sistemi z različnimi metapodatkovnimi strukturami. (Z učinkovitejšo interoperabilnostjo je v tem primeru mišljena možnost boljšega komuniciranja med posameznimi sistemi, kar bo razloženo v nadaljevanju.)

Dogovorjeni skupni metapodatkovni format evropskih sistemov CRIS se imenuje CERIF (Common European Research Information Format) in je predstavljen v obliki entitetno relacijskega (ER) modela. Baza SICRIS je slovenska izvedenka modela CERIF in se po nekaterih lokalnih posebnostih sicer razlikuje od modela CERIF, vendar je kompatibilna z njim. Ena od bistvenih razlik med modeloma je v tem, da v modelu baze SICRIS manjka osnovna entiteta Results. Ta je pri modelu CERIF namenjena shranjevanju raziskovalnih rezultatov (člankov, knjig, zapisov s konferenc itd.).

Razlog za to je zgodovinski, saj smo imeli v Sloveniji raziskovalne rezultate že dosti pred uvedbo modela CERIF in informacijskega sistema SICRIS shranjene v slovenski vzajemni bibliografski bazi COBIB.SI in bi bil zato prenos teh zapisov v model CERIF zelo neracionalen. Bazi SICRIS in COBIB.SI sta povezani v entitetah Raziskovalec (baza SICRIS) in Bibliografska enota[2] (baza COBIB.SI) preko enoličnega ključa, ki je v tem primeru številka raziskovalca.

S trditvijo 2 bomo poskušali preveriti, ali je z ontologijami mogoče izboljšati komunikacijo med posameznimi evropskimi sistemi CRIS v primerjavi s predlogom organizacije EUROCRIS, ki ga kaže slika 1. V tem predlogu je bilo zamišljeno, da bi evropski sistemi CRIS med seboj komunicirali preko formata CERIF Exchange (Joerg, Krast, Jeffery, & Van Grootel, 2007). Ta predstavlja dogovorjeno podmnožico podatkov metapodatkovnega formata CERIF. HUNCRIS na sliki predstavlja madžarski sistem CRIS, FRIDA norveški sistem CRIS, METIS pa nizozemski sistem CRIS.



Slika 1: **Združevanje evropskih sistemov CRIS preko formata CERIF Exchange**

[1] Z gostejšo pokritostjo metapodatkov je tu mišljeno, da imamo v primeru ontološko obogatenega iskalnika na voljo določene metapodatke, ki jih sicer ne bi imeli pri običajnem iskalniku.

[2] Bibliografska enota je po bibliotekarski definiciji lahko monografska publikacija, članek v reviji, serijska publikacija, predstavitev na konferenci ipd.

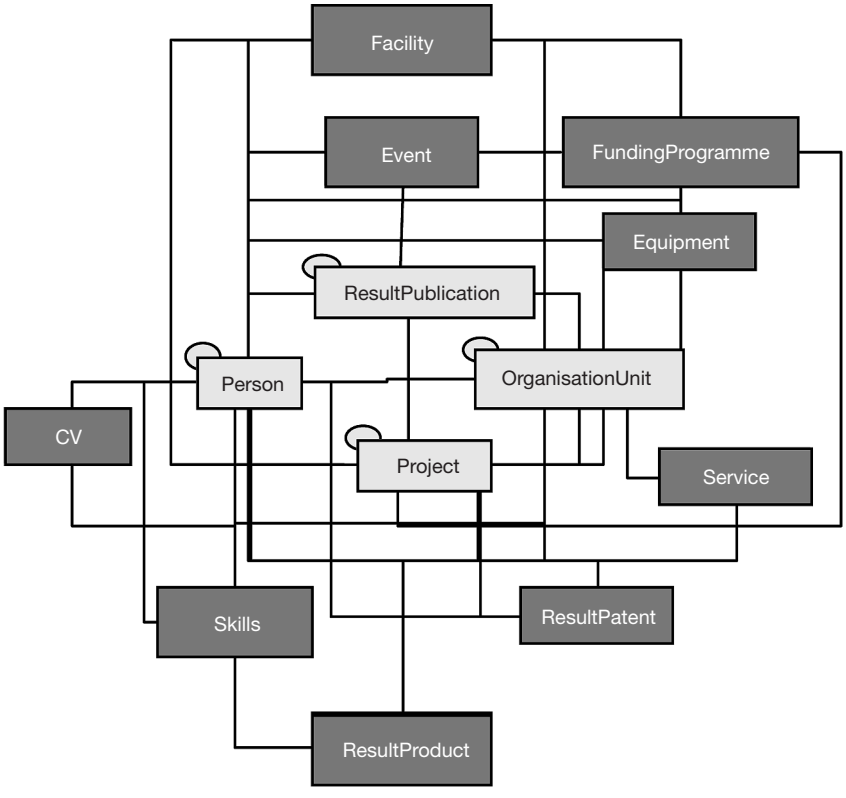
Predlog uvedbe ontologij bo v nadaljevanju zaradi splošnosti rešitve prikazan na modelu CERIF, mogoče pa ga je zlahka razširiti tudi na slovensko različico.

2 METAPODATKOVNI MODEL CERIF

Za natančnejšo razlago povedanega je treba najprej pogledati glavne značilnosti metapodatkovnega modela CERIF. Format je v praksi standard, tehnično gledano pa ima status priporočila EU. CERIF 2000 je komisija priporočila svojim članicam (CERIF: a service hosted by CORDIS, 2012). Po letu 2002 je skrb za vzdrževanje in razvoj formata Evropska komisija poverila neprofitni organizaciji EUROCRIS (EUROCRIS, 2012). CERIF se v svoji zadnji razvojni fazi imenuje CERIF 2008 – 1.2 Full Data Model (Joerg idr., 2010). Opis modela je prirejen relacijskim bazam. V svojih osnovnih tabelah opisuje lastnosti štirih glavnih raziskovalnih entitet:

- *Person* – za opis oseb, vključenih v raziskovalni proces,
- *OrgUnit* – za opis raziskovalnih organizacij ali tudi samo organizacijskih enot,
- *ResultPublication* – za opis raziskovalnih rezultatov v obliki klasičnih ali e-publikacij (knjige, članki, poročila ipd.),
- *Project* – za opis raziskovalnih projektov ali programov.

V procesih znanstvenoraziskovalne srenje predstavljajo osnovne entitete glavne akterje (Person, OrgUnit) in tudi rezultate njihovih aktivnosti (ResultPublication, Project).



Slika 2: **Glavne entitete CERIF v povezi z entitetami druge ravni**

Sivo obarvane entitete na sliki 2 so med seboj povezane vsaka z vsako. Poleg te povezave pa v modelu obstaja še avtorekurzivna povezava, to je povezava vsake glavne entitete s samo seboj.

Primer 1: ***Povezovanje različnih entitet***
Oseba A je relacijsko povezana z organizacijo B (oseba A je zaposlena v organizaciji B, A je lahko direktor organizacije B, A lahko le uporablja opremo organizacije B ipd.).

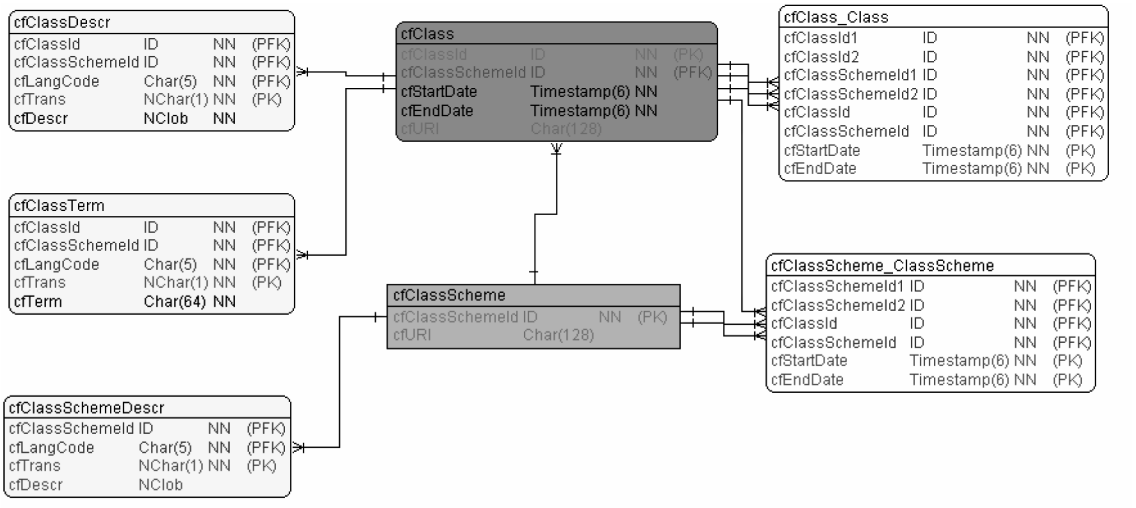
Primer 2: ***Avtorekurzivno povezovanje***
Organizacija C je relacijsko povezana z organizacijo D (organizacija C je financer organizacije D).

Poleg glavnih entitet pozna model CERIF tudi entitete druge ravni. Te razširjajo model osnovnih in povezovalnih entitet. Največkrat služijo za opis širšega okolja osnovnih entitet in njihovih interakcij. Nekatere pomembnejše entitete druge ravni (temni okvirčki) in njihove medsebojne povezave so prikazane na sliki 2.

Jezikovno odvisne entitete CERIF: format CERIF vsebuje tudi mnogo jezikovno odvisnih entitet, kot npr. ključne besede, povzetek, ime organizacije, ime projekta, naslov itd.

3 CERIF 2008 IN SEMANTIČNI SLOJ

Bistvena razlika med metapodatkovnimi modeli CERIF 2008 (velja tudi za novejše različice) in njegovimi starejšimi različicami se kaže predvsem v tem, da vsebuje najnovejši model posebej izločen sloj za vnos semantičnih informacij, ki omogoča tudi uporabo ontologij. Sloj je v modelu predstavljen s tabelami s predpono Class.Te so s poljubnimi tabelami modela povezane preko ključev ClassId in ClassSchemeId.



Slika 3: **Entitete CERIF semantičnega sloja v obliki fizičnega modela**

Semantični sloj vsebuje entitete iz slike 3. Osrednja klasifikacijska entiteta s slike (*cfClass*) je preko povezave n:1 povezana z entiteto klasifikacijskih shem (*cfClassScheme*). Entitete semantičnega sloja podpirajo:

- vsebinske klasifikacije samih entitet modela s pomočjo entitet na levi: *cfClassDescr*,*cfClassTerm* in *cfClassSchemeDescr*.

Primer 3:
Za tričrkovno klasifikacijo svetovnih jezikov ISO bi pri slovenskem jeziku imeli v *cfClassTerm* (cfTerm) vrednost »slo«, pri *cfClassDescr* (cfDescr) vrednost »Slovenian language« in v *cfClassSchemeDescr* (cfDescr) »ISO 639-3; 2007 languagecodes«.

- Vsebinske klasifikacije povezovalnih entitet, ki v tem, ko se nanašajo na neko entiteto, običajno opisujejo njeno vlogo (Roles) ali tip (Type). Za te je značilno, da jih lahko klasificiramo znotraj neke določene klasifikacijske sheme.

Primer 4:
Povezovalne tabele vlog (Roles) pri povezavi entitet cfPerson_cfResult_Publication lahko po vnaprej določeni klasifikacijski shemi opišemo vloge oseb v publikaciji z vrednostmi avtor, soavtor, ilustrator, pisec predgovora ipd.

Zelo pomemben del semantičnega sloja sta avtorekurzivni entiteti cfClass_Class in cfClassScheme_ ClassScheme. Ti entiteti, ki na neki način spominjata na že opisano avtorekurzijo glavnih entitet, uvajata tudi v tem primeru nekaj zelo pomembnih konceptov:

- **cfClass_Class** – v primeru povezave *klasifikacija/klasifikacija* lahko opišemo koncepte in trditve, kot

npr. »refošk« je »rdeče vino«. Povezava omogoča da vsaki posamezni klasifikaciji lahko priredimo ustrezno »nadklasifikacijo« (klasifikacijsko shemo, v katero spada ta klasifikacija). V zgornji trditvi je refošk ena od vrednosti v klasifikaciji rdečih vin, medtem ko je rdeče vino ena od vrednosti splošne klasifikacije vin. V taki klasifikaciji se lahko pojavljajo še bela vina, rozéji ipd.;

- **cfClassScheme_ClassScheme**– v primeru povezave klasifikacijska shema/klasifikacijska shema lahko opišemo koncepte in trditve, ki opisujejo povezave in odnose med klasifikacijskimi (taksonomskimi) shemami, kot npr. klasifikacija vin je podskupina klasifikacije alkoholnih pijač, ta pa nato spet naprej podskupina klasifikacije pijač.

Znotraj entitete cfClass_Class lahko opišemo celo paleto različnih konceptov klasifikacijskih struktur: od najpreprostejšega »je« odnosa med taksonomskimi klasifikacijami, do relacij »sinonim od«, »širši pojem od«, »ožji pojem od«, ki se uporabljajo pri terminoloških slovarjih – tezavrih (Broughton, 2006), in nazadnje do pomensko definiranih relacij kot so »starši od« ali »sestra od« med entitetami s pomočjo ontologij (Hendler, Lassila, & Berners-Lee, 2001).

4 POMANJKLJIVOSTI OBSTOJEČEGA MODELA INFORMACIJSKEGA SISTEMA O RAZISKOVALNI DEJAVNOSTI

Pri pomanjkljivostih, ki jih v praksi prinaša tako zastavljeni model informacijskega sistema o raziskovalni dejavnosti se bomo v članku osredinili predvsem na dve vrsti težav, ki sta se v zadnjih nekaj letih na ravni evropskih sistemov CRIS izkazali za pomembnejši (glej opis ciljev delovne skupine BestPractice TG (EUROCRIS Task Groups, 2010)):

- težave, ki jih imajo uporabniki pri iskanju relevantnih podatkov zaradi slabega obvladovanja kompleksne strukture metapodatkov modela CERIF;
- težave, ki jih imajo uporabniki pri iskanju zaradi slabe interoperabilnosti med CRIS sistemi z različnimi strukturami (običajno posledica implementacije različnih verzij modela CERIF).

Omenjene težave, kakor tudi predlog možnosti za njihovo odpravljanje bodo opisani v nadaljevanju.

4.1 Težave pri iskanju po entitetah modela CERIF zaradi nepoznavanja obsežne strukture

V strukturi modela CERIF se nahaja 191 različnih entitet, za katere je malo verjetno, da bi jih poznali uporabniki, ki niso specialisti s tega področja. Za ilustracijo kompleksnosti strukture lahko povemo, da ta čas največji svetovni ponudnik specialnih baz podatkov DIALOG (DIALOG, 2012), ki deluje pod okriljem podjetja ProQUEST, svojim abonentom kot opise baz ponuja t. i. Datasheete (DATASTAR, 2011). Ti za posamezno bazo podatkov vsebujejo opise od 20 do približno 40 različnih iskalnih indeksov. Za manj izkušene iskalce je v takih primerih na voljo tudi servis telefonske ali spletne pomoči.

V nadaljevanju bomo podali nekaj splošnih primerov iz vsakdanje prakse iskanja po kombinaciji baz SICRIS in COBIB.SI, ki predstavlja slovensko implementacijo modela CERIF. Primeri so vsebinsko vzeti iz zapisov telefonskih pogovorov pomoči uporabnikom servisa SICRIS.

Karakteristični primer 1: *V bazah SICRIS in COBIB.SI želimo poiskati vse publikacije, ki jih je v letošnjem letu napisal avtor s priimkom Kovač. Zanj pa vemo le to, da je univerzitetni profesor z nazivom docent, redno zaposlen nekje na Univerzi v Ljubljani.*

Rešitev 1: V primeru iskanja po bazi z znanim metapodatkovnim formatom CERIF bo za rešitev treba postaviti te, z logičnim AND povezane pogoje:

- oseba mora imeti priimek Kovač;
- oseba mora biti avtor publikacije;
- publikacija mora biti izdana v letu 2010;
- oseba mora imeti v organizaciji redno zaposlitev;
- oseba mora imeti znanstveni naziv docent;
- organizacija, v kateri je oseba zaposlena, mora spadati med organizacije Univerze v Ljubljani.

Povpraševalni niz za iskalnik, ki podpira format CERIF, bi moral biti v tem primeru sestavljen iz tehle pogojev:

- Person (FirstName) = KOVAČ
- *Person_ResultPublication (Classification) = Tip avtorstva, ClassTerm = Avtor Person_ResultPublication (StartDate) = 2010-01-01 Person_ResultPublication (EndDate) = 2010-12-31*
- Person_Organisation (Classification) = Zaposlitev, ClassTerm = Redno zaposlen
- Person_Organisation (Classification) = Znanstveni naziv, ClassTerm = Docent
- Organisation (Classification) = Tip organizacij, ClassTerm = Univerza
- Organisation (Classification) = Klasifikacija slovenskih univerz, ClassTerm = Univerza v Ljubljani

Pri slovenski implementaciji modela CERIF pa je stvar malo bolj kompleksna. V tem primeru je za

rešitev naloge treba poznati metapodatkovne modele baz SICRIS in COBIB.SI in povezavo med obema bazama, ki poteka preko enolične številke raziskovalca.

Naloga je zaradi tega rešljiva v dveh zaporednih korakih:

- v prvem je treba s pomočjo zgornjih alinej v bazi SICRIS poiskati avtorja Kovača in njegovo enolično številko raziskovalca,
- v drugem pa je s pomočjo znane številke raziskovalca v bazi COBIB.SI treba poiskati vse njegove bibliografske enote, ki so bile izdane leta 2010.

V spletnem iskalniku sistema COBISS/OPAC sistema COBISS (COBISS – Kooperativni bibliografski sistem in servisi, 2012) v bazi COBIB.SI bi bil za to potreben tale iskalni niz: SELECT AS=08050[3]and PY=2010.

Glede na to, da splošni uporabnik obeh baz ne more poznati tega pravila, se mu ponuja bistveno učinkovitejša rešitev pri iskanju s pomočjo uporabe ontologije, kar bo razloženo v nadaljevanju.

Karakteristični primer 2: Uporabnik baze SICRIS želi ugotoviti, katere znanstvenike z najvišjimi mogočimi znanstvenimi nazivi (redni ali izredni profesor) iz različnih znanstvenih institucij lahko povabi v skupni projekt s tematskega področja robotike. Pri tem si želi, da bi bil iskani raziskovalec hkrati nekdo, ki lahko kandidira za sredstva ARRS (ARRS – Agencija RS za raziskovalno dejavnost, 2012).

Rešitev 2: Za rešitev omenjene naloge bo treba iskalni niz sestaviti iz več pogojev:

1. osebe morajo imeti prvi ali vsaj drugi najvišji mogoči znanstveni naziv;
2. osebe morajo imeti med svojimi podatki zapisan raziskovalni interes s področja robotike;
3. za kandidiranje na programe NRP – Nacionalnega raziskovalnega in razvojnega programa (Nacionalni raziskovalni in razvojni program; Uradni list RS, 3/2006, Resolucija o nacionalnem raziskovalnem in razvojnem programu za obdobje 2006–2010 (ReNRRP), 2006) morajo biti osebe raziskovalci s potrjenim statusom;[4]
4. osebe morajo imeti aktualno zaposlitev v raziskovalni organizaciji s potrjenim statusom.

Če predpostavimo, da ima iskalec dovolj dobro znanje glede strukture vseh entitet, lahko pride do rešitve prve in druge alineje. Zadeva pa vseeno ni čisto preprosta. Uporabnik baze iz naloge 2 je lahko tudi oseba iz tujine. Recimo, da je podatek o znanstvenem nazivu iz druge alineje šifriran in se nahaja v CERIF entiteti cf_Pers_Qual. V šifrantu se nahajajo te vsebine:

- Asistent z doktoratom,
- Docent,
- Izredni profesor,
- Redni profesor itn.

Vsaka od šifer ima v modelu CERIF predviden tudi pripadajoči angleški prevod, ki je s slovenskim izvirnikom povezan preko primarnega ključa:

- PostdoctoralResearchAssistant,
- Juniorprofessor,
- Professorextraordinarius,
- Professorordinarius itn.

Tega iskalec sicer lažje razume, vendar pri poljih, ki vsebujejo daljše šifrante, ni zmeraj nujno, da bo iskalec vsebinsko pravilno izbral šifre, sploh če te niso urejene po nekem logičnem vrstnem redu.

Podatek o raziskovalnem interesu posameznika iz zadnje alineje najdemo hkrati v dveh entitetah CERIF, in sicer v:

- cf_PersResInt, v kateri se nahaja opisno v besedilni obliki,
- cf_Pers_Class, v kateri je isti podatek šifriran v skladu s klasifikacijsko shemo posamezne države. V Sloveniji je to npr. šifrant http://www.arrs.gov.si/sl/gradivo/sifranti/sif-vpp.asp, v katerem je iskana šifra »2.10.04 – robotika«.

Za neveščega iskalca, ki pozna le strukturo ne pa tudi vsebine, se v primeru šifriranih polj cf_Pers_Qual in cf_Pers_Class obeta zamudna iskalna procedura. Najprej je treba po indeksu preiskati vsebine vnesenih šifriranih entitet. Šele v drugem koraku, ko poznamo vsebino, se lahko postavi neki ustrezen iskalni niz. S preiskovanjem po indeksu, ki ga izvajamo z ukazom EXPAND CS=2, bi v tem primeru dobili izpisane vse indekse od 2 naprej (CS je akronim iskalnega indeksa za polje cf_Pers_Class): EXPAND CS = 2.

2	2.10.02	Izdelovalna tehnologija
64	2.10.03	Avtomatizacija
25	2.10.04	Robotika
11	2.10.05	Industrijski inženiring
76	2.10.06	Varilstvo

Slika 4: **Iskalni rezultat funkcije EXPAND**

[3] Številka 08050 je zaradi varovanja osebnih podatkov anonimizirana in ne ustreza dejanskim podatkom.

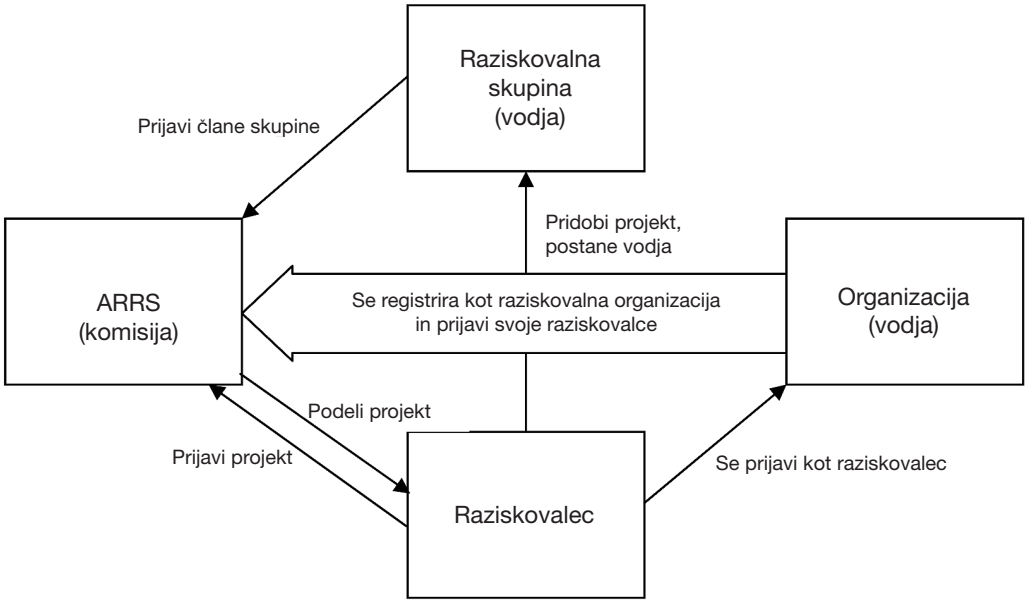
[4] Potrjen status pomeni, da je osebe kot raziskovalce na ARRS prijavila odgovorna oseba iz raziskovalne organizacije, hkrati pa je organizacijo potrdila ARRS.

V levi koloni slike 4 se nahaja število zadetkov v bazi za posamični indeks. Kako bi si v tem primeru lahko pomagali z ontološko obogatenim iskalnikom, bo natančneje razloženo v nadaljevanju.

Nepoznavanje pravil: Zadnjih dveh alinej rešitve karakterističnega primera 2, ki se nanašata na potrjeni status raziskovalca in potrjeni status raziskovalne organizacije, ni mogoče rešiti s poznavanjem same strukture, saj je za njihovo reševanje potrebno poznavanje določenih pravil, ki so poznana samo poznavalcu tega področja.

V slovenski raziskovalni sferi je na primer poznano pravilo, da lahko za projekte slovenskega NRP kandidira le raziskovalna skupina, ki jo je prijavila potrjena raziskovalna organizacija. V praksi to pomeni, da je organizacija registrirana pri ARRS. Vrstni red dogodkov je prikazan na sliki 5 in si mora slediti takole:

- najprej mora raziskovalna organizacija izpolniti določene pogoje, na podlagi katerih se lahko registrira pri ARRS in tam dobi potrjen status raziskovalne organizacije;
- ko je organizacija pridobila status, lahko pri ARRS prijavi svoje raziskovalce;[5]
- ko je to opravljeno, se lahko njeni raziskovalci prijavljajo na projekte pri ARRS;
- ob uspešni pridobitvi projekta si lahko raziskovalec sestavi raziskovalno skupino in jo registrira pri ARRS (člani skupine so lahko tudi iz drugih raziskovalnih organizacij).



Slika 5: **Poslovni procesi pri dodeljevanju statusov na ARRS**

Glede na to, da je za rešitev naloge potrebno poznavanje poslovnih procesov in pravil pri dodelitvi statusa, se v tem primeru z običajnimi iskalnimi nizi ne da ugotoviti, katera organizacija in raziskovalec sta po standardih ARRS primerna za kandidaturo na razpisani projekt NRP in katera ne. V nadaljevanju bomo pokazali, kako je na to vprašanje mogoče odgovoriti z uporabo lokalne ontologije.

5 PREDNOSTI KREIRANJA DOMENSKIH ONTOLOGIJ

Po mnenju avtoric D. McGuinness in N. F. Noy (Noy & McGuinness, 2001) je domenska ontologija dogovor o konceptualizaciji znotraj neke določene vsebinske domene. Običajno vsebuje najprej slovar terminov in oznak ter nato opredelitve konceptov ter njihove medsebojne odvisnosti in omejitve. Opisovanje se lahko nanaša bodisi na razrede lastnosti ali na primerke. Glavna razlika med podatkovnimi modeli, kot je CERIF, in ontologijami je v tem, da se podatkovni modeli ukvarjajo predvsem s strukturo in integriteto podatkov, ontologije pa s formalno predstavitvijo znanja. Opis znanja, ki je najbolj zaželen, je tak,

[5] Tudi raziskovalec mora ustrezati določenim predpisanim pogojem (minimalna izobrazba in znanstveni naziv).

da je čim bolj splošen in ga zaradi te splošnosti lahko souporabljamo v čim več različnih aplikacijah.

Tehnološko gledano sestavlja vsako ontologijo niz trditev RDF (Lassila & Swick, 1999). Glede na to, da je ontologija eden od osnovnih gradnikov semantičnega spleta, bi bilo dobro za začetek vsaj v osnovi razdelati tudi problematiko v zvezi s tem. Tehnologije, ki so omogočile nastanek semantičnega spleta, so URI – uniform resource identifier (URI, 2001), XML – extensible mark up language (XML, 2000) in RDF – (Resorce Description Framework) (RDF, 2004).

Trditve RDF so za namene strojnega procesiranja zapisane v formatu XML, sicer pa za človeško branje pretežno uporabljamo predstavitev v obliki grafa.

Uporaba ontologij in jezik OWL: kot je bilo že omenjeno, za predstavitev ontologij uporabljamo različne jezike. Nekateri avtorji (Lavbič & Krisper, 2005) jih kronološko in tudi vsebinsko delijo na tiste, ki so nastali pred semantičnim spletom (KL-ONE, Ontolingua, LOOM, KIF, CycL in UML), in tiste, ki so nastali v tem obdobju (RDF/S, DAML+OIL, OWL).

Ocenjujemo, da je za opis in uporabo ontologije, ki bi pomagala rešiti naše težave pri iskanju, najprimernejši izbor jezika OWL. Razlog za ta izbor je, da je njegova izrazna moč med vsemi omenjenimi največja in tudi, da je njegova uporaba najbolj razširjena. Pomembno je tudi dejstvo, da za ontologije, napisane v jeziku OWL, obstaja kar nekaj sklepalnikov (angl. reasoner).

Jezik OWL je mogoče uporabiti v treh različno izrazno močnih variantah: OWL Lite, OWL DL in OWL Full. V sklepalnikih je najbolje podprt jezik OWL DL, zato se bomo pri prikazu rešitev osredinili nanj.

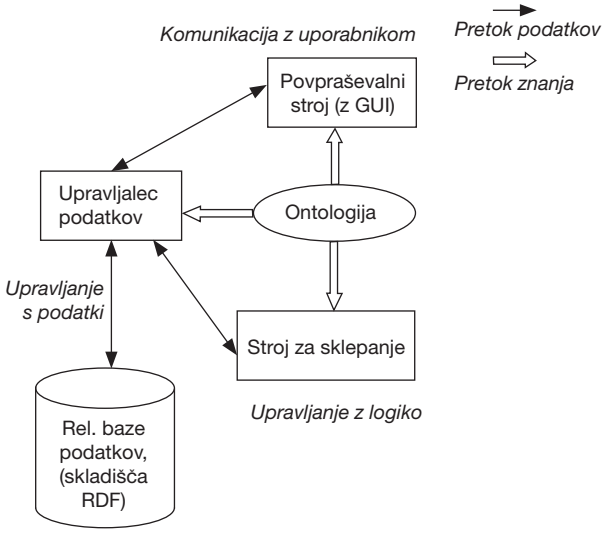
Rešitev, ki jo predlaga EUROCRIS z uvedbo semantičnega sloja v model ERM, CERIF 2008 (slika 3), ima po naših ocenah predvsem to pomanjkljivost, da omogoča le opise relacij med različnimi taksonomskimi razredi in primerki te domene. Izrazno močnejše ontologije nam poleg tega omogočajo tudi opise pravil, ki veljajo v domeni in brez katerih ni možnosti za »inteligentno« iskanje. Izkoriščanje domenskih pravil pri iskanju izvajamo s pomočjo procesa sklepanja, pri katerem iz obstoječih pravil izvajamo nova pravila (dobra podlaga za mogoče sklepanje v OWL so npr. opisane karakteristike definiranih lastnosti, kot so tranzitivnost, simetričnost ipd.) (OWL, 2012).

5.1 Iskanje po relacijskih bazah z ontološko obogatenim iskalnikom

Pri reševanju problematike, ki jo nakazujeta karakteristična primera 1 in 2, sta nas vodili predvsem dve zahtevi:

- treba je zagotoviti možnost iskanja po različnih relacijskih bazah z metapodatkovnim formatom CERIF (ali kakšno njegovo izvedenko);
- iskanje mora biti izvedeno na učinkovitejši način kot s klasičnim iskalnikom, saj bo moralo upoštevati tudi koncepte in pravila, ki so uveljavljeni v domeni sistemov CRIS.

Prednosti ontologij lahko najbolje preverimo z uporabo ene od primernih ontoloških infrastruktur.[6] Na poti do predloga rešitve je bilo treba najprej preučiti obstoječe ontološke infrastrukture, saj smo predpostavljali, da bomo zastavljene cilje skušali doseči z uporabo ontologije. Kot ena zelo primernih infrastruktur se je izkazala infrastruktura Ontobroker[7] (Ontobroker, 2012). Infrastruktura ima vse funkcionalnosti, ki jih potrebujemo, poleg tega pa omogoča še iskanje virov na spletu, njihovo indeksiranje, semantično anotacijo in še nekatere druge funkcionalnosti, ki jih v našem primeru ne potrebujemo. Ko smo zreducirali infrastrukturo na tiste gradnike, ki so za nas ključnega pomena, ter izvedli nekatere manjše spremembe, smo dobili infrastrukturo na sliki 6.



Slika 6: **Infrastruktura za iskanje z ontološko obogatenim iskalnikom**

[6] Z ontološko infrastrukturo je mišljena struktura, ki jo sestavljajo gradniki (običajno aplikacije), z namenom, da omogočajo čim učinkovitejšo uporabo ontologij na posameznem področju uporabe.

[7] Ontobroker je danes komercialno orodje, ki ga kot middleware ponuja podjetje Ontoprise. Uporabljajo ga nemški Telekom, podjetje Audi, robotski sistemi Kuka in še nekateri drugi.

Povpraševalni stroj je enota, ki vsebuje grafični vmesnik za uporabnika, ki prikazuje hkrati razrede in njihove lastnosti ter tudi vsebine pripadajočih tabel v relacijski bazi, če te obstajajo. Stroj pošlje iskalno zahtevo upravljalcu podatkov. Zahteva lahko v splošnem vsebuje specifikacijo razredov in lastnosti iz ontologije in tudi omejitev vsebin relacijskih tabel.

Pri preslikavi posamezne ontologije v relacijsko bazo je značilno, da ni nujno, da imamo za vsak poljubni razred ali lastnost v ontologiji relacijsko tabelo v bazi. V primeru naloge 2 imamo na primer v ontologiji razred RRK and id ARRS, ki predstavlja raziskovalca, kandidata za projekte NRP. Takega polja ni v relacijski bazi, vseeno pa lahko iskanje izvedemo s pomočjo uporabe koncepta, ki je zapisan v ontologiji.

Upravljalec podatkov sprejme iskalno zahtevo, in tisti del, ki je vezan na ontologije, posreduje stroju za sklepanje. Ta potem izvede povpraševanje po ontologiji in vrne rezultat upravljalcu podatkov. V infrastrukturi je zamišljeno, da je lahko relacijskih baz tudi več, če so v njih shranjeni podatki, lahko pa imamo namesto relacijskih baz tudi skladišče RDF trojčkov, po katerih lahko povprašujemo z orodjem SPARQL (SPARQL, 2012).

Ko sistem za upravljanje podatkovnih baz vrne rezultat, upravljalec podatke posreduje povpraševalnemu stroju, ki jih potem prikaže v ustreznem uporabniškem vmesniku.

Stroj za sklepanje izvede povpraševanja po eni ali več ontologijah in posreduje podatke upravljalcu podatkov. Danes večina tovrstnih strojev uporablja algoritem Tableau (Haehnle, 2001), ki ima težave pri večjih ontologijah. Nekateri, npr. KAON2 (KAON2, 2012) ali Hermi T Reasoner,[8] izvajajo sklepanje na podlagi opisne logike. V tem primeru je treba imeti tudi zapis ontologije v ustreznem jeziku, kot je npr. OWL DL.

6 REŠEVANJE KARAKTERISTIČNIH PRIMEROV S POMOČJO ONTOLOŠKE INFRASTRUKTURE

Poglejmo si ponovno dva primera povpraševanja, ki smo ju obravnavali na začetku, in sta se pri uporabnikih sistema SICRIS izkazala kot karakteristična.

Karakteristični primer 1*:* V prvem primeru je bila želja poiskati vse publikacije, ki jih je v letošnjem letu napisal avtor s priimkom Kovač. Zanj vemo, da je univerzitetni profesor docent, redno zaposlen na Univerzi v Ljubljani. Pri nalogi, ki smo jo sicer lahko rešili, se je izkazalo za problematično slabo poznavanje strukture in pravila, da so podatki raziskovalca, vpisani v bazi SICRIS, in njegove publikacije v bazi COBIB.SI, povezani preko enolične številke raziskovalca. Kako lahko v tem primeru pomagajo ontologije?

V jeziku OWL zapišemo karakteristiko lastnosti »imaARRSst« raziskovalca, kar pomeni, da ima obvezno enolično številko ARRS (ki deluje podobno kot primarni ključ v relacijskih bazah) takole:

```
<owl:ObjectPropertyrdf:ID=«imaARRSst«>
<rdf:typerdf:resource=«&owl;InverseFunctionalProperty« />
<rdfs:domainrdf:resource=«#Raziskovalec« />
<rdfs:range rdf:resource=«#ARRSstRR« />
</owl:ObjectProperty>
```

Koda 1: **Karakteristika lastnosti »imaARRSst«**

Z zapisom karakteristike lastnosti »imaARRSst« in s pomočjo zgornje ontološke infrastrukture se naloga izvede povsem drugače. Uporabnik bi v tem primeru z navigacijo označil dva razreda: »Raziskovalec« in »Publikacije« (ki je podrazred razreda »Raziskovalni rezultati« – ta predstavlja osnovno entiteto CERIF Results). Iskanje bi določil kot konjunkcijo obeh posameznih pogojev. Pri razredu »Raziskovalec« bi v oknu za dodatno omejevanje pod priimek vpisal »Kovač«, pod znanstvenimi nazivi bi iz seznama vnaprej določenih vrednosti izbral »Docent« ter pod zaposlitev »Univerza v Ljubljani«. Pri razredu »Publikacije« pa bi v oknu za dodatne kriterije omejil letnico izdaje na 2010. Povpraševalni stroj bi prenesel iskalne zahteve upravljalcu podatkov. Ta bi preko stroja za sklepanje najprej povprašal po ontologiji. Tam bi zaradi konceptov:

- da ima vsak raziskovalec svojo raziskovalno številko in

[8] Integriran v orodju Protege, verzija 4.1.0.

- da je vsaka ARRS raziskovalna šifra enolična (koncept je v OWL v izpisu kode 1),
- vsaka publikacija raziskovalca pa ima vpisano vsaj eno številko ARRS,

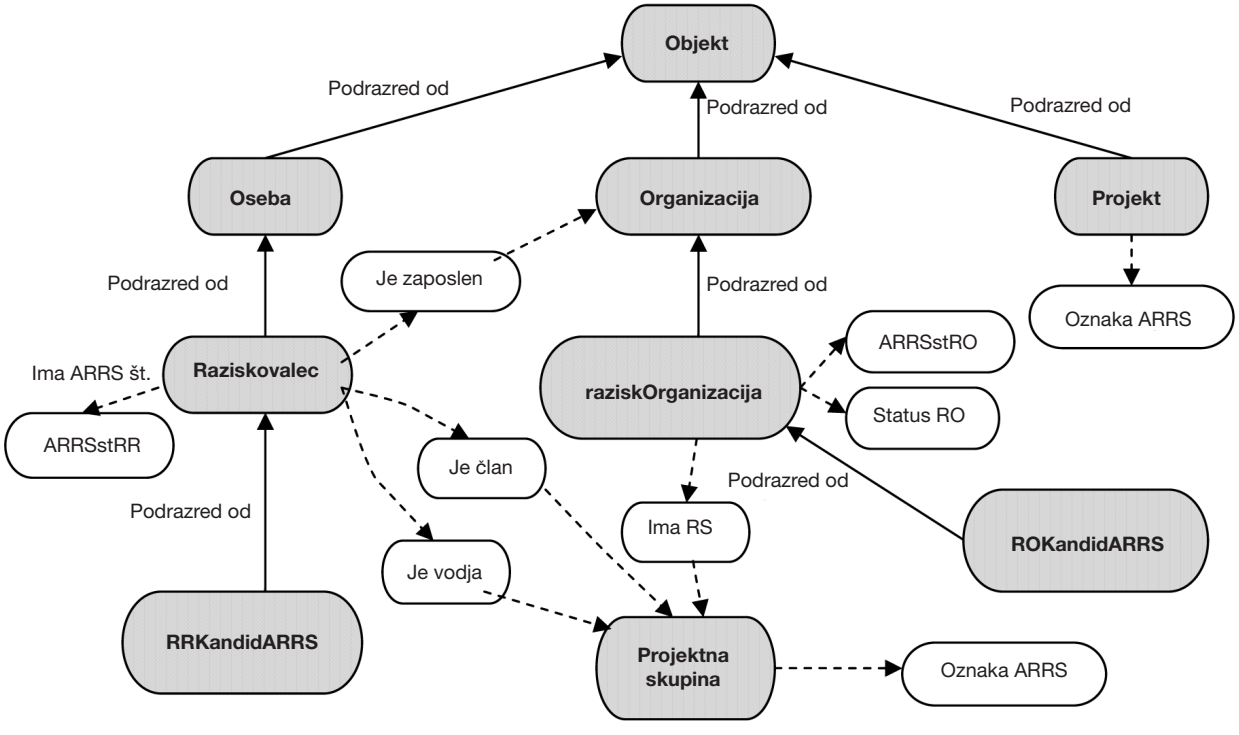
prišel do sklepa, da so raziskovalčeve publikacije prav tiste, ki imajo vpisano njegovo številko ARRS. Rezultat, ki bi ga stroj za sklepanje v obliki ontoloških razredov posredoval upravljalcu podatkov, bi ta potem pretvoril v ustrezne povpraševalne nize za pripadajoče baze podatkov. V primeru slovenskega sistema CRIS bi za tiste ontološke razrede, pri katerih se pripadajoče relacijske tabele nahajajo v SICRIS-u, izvedel povpraševanje v SICRIS-u, za druge, ki se nahajajo v bazi COBIB.SI, pa bi izvedel povpraševanje tam.

Karakteristični primer 2*:* Pri reševanju drugega karakterističnega primera se nam je zataknilo pri zadnjih alinejah:

- za kandidiranje na programe NRP morajo biti osebe raziskovalci s potrjenim statusom;
- osebe morajo imeti aktualno zaposlitev v raziskovalni organizaciji s potrjenim statusom.

Poglejmo, kako nam bi s pomočjo poznavanja prej omenjenih konceptov dodeljevanja potrjenega statusa lahko pomagale ontologije. Pri rešitvi bodo predstavljeni samo tisti deli, ki nas bodo pripeljali do rešitve problema.

Na sliki 7 vidimo del ontologije slovenske znanstvenoraziskovalne sfere, predstavljene z grafom RDF. Razredi so označeni s potemnjenimi entitetami. S povezavo »subclassof« med razredi je označeno, da je entiteta na začetku puščice podrazred entitete na koncu. Z RRKandidARRS je označena podmnožica raziskovalcev, ki lahko na ARRS kandidira za projekte. Z ROKandidARRS je na sliki označena posebna podmnožica raziskovalnih organizacij – tistih, katerih raziskovalci lahko kandidirajo za projekte. Nepotemnjene entitete, povezane s črtkanimi puščicami, predstavljajo atribute ali lastnosti razredov.



Slika 7: **Del ontologije slovenske znanstvenoraziskovalne sfere**

Za vsako lastnost je v ontologiji mogoče napisati pravila, ki služijo za podporo sklepanju odločitvenega sistema. OWL DL pozna med drugim te karakteristike za lastnosti:

- tranzitivnost,
- simetričnost,
- funkcionalnost,
- inverzija,
- inverzna funkcionalnost.

Lastnost »je zaposlen«, ki jo ima v našem primeru raziskovalec, je inverzna, z lastnostjo »Zaposluje«, ki jo ima organizacija.

```
<owl:ObjectPropertyrdf:ID=«jeZaposlenV«>
<owl:inverseOfrdf:resource=«#Zaposluje« />
</owl:ObjectProperty>
```

Koda 2: **Lastnost »jeZaposlen«**

Pravilo, ki rešuje našo nalogo, pravi, da lahko raziskovalec kandidira za projekt NRP le, če ima potrjen status in je hkrati v danem trenutku tudi zaposlen v raziskovalni organizaciji, ki ima prav tako potrjen status, kar se izraža z atributom »Status RO«, ki mora v tem primeru imeti vrednost »raziskovalna organizacija«.

V OWL jeziku se pravilo v skrajšani obliki zapiše takole:

```
<owl:Classrdf:about=« #RRKandidARRS«>
<owl:intersectionOfrdf:parseType=«Collection«>
<owl:Classrdf:about=«#Raziskovalec« />

<owl:Restriction>
<owl:onProperty>rdf:resource=«jeZaposlenV« />
<owl:hasValuerdf:resource=«#ROKandidARRS« />
</owl:Restriction>
</owl:intersectionOf>
</owl:Class>


<owl:Classrdf:about=«ROKandidARRS«>
<owl:intersectionOfrdf:parseType=«Collection«>
<owl:Classrdf:about=«#raziskOrganizacija« />

<owl:Restriction>
<owl:onPropertyrdf:resource=«#Status« />
<owl:hasValuerdf:resource=”#Raziskovalna organizacija”>
</owl:Restriction>
</owl:intersectionOf>
</owl:Class>
```

Koda 3: **Pravilo glede kandidature za projekte NRP, zapisano v OWL**

V prvem delu zgornjega izpisa kode (koda 3) smo definirali razred RRKandidARRS, kar predstavlja raziskovalca, ki se lahko poteguje za projekte NRP. Ta mora zadostiti dvema pogojema (konstrukta owl:intersectionOf in rdf:parseType=«Collection«). Prvi pogoj zahteva, da spada v razred Raziskovalec. Drugi pogoj pravi, da je zaposlen v organizaciji razreda ROKandidARRS. Pogoje, ki jim mora zadostiti taka organizacija, smo napisali v spodnjem delu zapisa kode, kjer smo definirali razred ROKandidARRS. To so organizacije, ki lahko registrirajo svoje raziskovalce kot bodoče kandidate za projekte NRP. Pogoj, ki ga mora izpolniti ta razred, je, da mora imeti vrednost statusa enako »Raziskovalna organizacija«.

V ontologiji slovenske znanstvenoraziskovalne sfere pa bi moralo biti zapisanih še več pravil. Eno teh, ki je vezano na raziskovalno organizacijo, med drugim pravi, da ima lahko raziskovalna organizacija status »raziskovalna organizacija« le, če ima vpisano šifro ARRS. To pravilo lahko služi pri preverjanju konsistentnosti zapisov v bazi, lahko pa služi tudi v primeru, če obstajajo zapisi organizacij, ki nimajo vpisanega podatka o statusu, imajo pa vpisano šifro ARRS.

7 SKLEP

Skupna ugotovitev karakterističnih primerov je, da bi pri reševanju iskalnih problemov najbolj pomagala ontologija, ki bi čim bolj generalno opisovala lastnosti v znanstvenoraziskovalni sferi. Lastnost, da mora biti »potrjeni« raziskovalec zaposlen v raziskovalni organizaciji, vsekakor ni generalna v nekem širšem evropskem prostoru, ampak velja le lokalno za Slovenijo, zato bi bila lahko takšna lastnost aktualna le v lokalni slovenski ontologiji.

Iskanje skupne evropske ontologije, ki bi v čim večji meri obsegala vse lastnosti lokalnih ontologij posameznih držav članic, je vsekakor eden od ciljev, ki se mu bo v prihodnosti morala posvetiti organizacija EUROCRIS.

V članku je bilo do sedaj na dveh karakterističnih primerih prikazano, kako lahko s pomočjo ontologij (ontološko podprtega iskalnika, stroja za sklepanje ipd.) rešujemo težave pri iskanju v bazah podatkov s kompleksnim metapodatkovnim modelom. Prikazano je bilo, kako lahko nepoznavanje kompleksne podatkovne sheme, pa tudi vsebovanih podatkov in njihove povezave s posameznimi predpisanimi procesi, rešimo z uporabo ustreznega zapisa v ontologiji.

Izkazalo se je, da vsaj za karakteristična primera držita trditvi 1 in 2, zapisani na začetku članka. Pri opisu iskanja prvega karakterističnega primera s pomočjo predlagane ontološke infrastrukture s slike 6 so bila v iskanje vključena tudi iskalna polja, ki jih baza podatkov ne pozna, saj jih je uvedel šele sklepalni stroj na podlagi pravila, ki je bilo zapisano v ontologiji. S tem je bila potrjena prva trditev, ki je govorila o povečani pokritosti uporabljenih metapodatkov pri iskanju. Druga trditev je bila prav tako potrjena z možnostjo iskanja po dveh različnih bazah: COBIB.SI in SICRIS.

V nadaljnjem delu se želimo posvetiti iskanju optimalne ontologije znanstvenoraziskovalne dejavnosti na podlagi natančne analize že obstoječih tovrstnih ontologij. Poleg tega želimo preizkusiti uporabo katerega od strojev za sklepanje, da bomo lahko na podlagi testnih primerov preizkusili, ali ontološka infrastruktura ustreza potrebam uporabnikov. Pridobiti in analizirati bo treba večjo količino izvedenih iskanj z uporabo ontološko obogatenih iskalnikov in brez nje. Poleg tega je koristen tudi razmislek, kako bi lahko v nekaj iteracijah iz obstoječe ontologije izvedli prehod v neke vrste »optimalno« ontologijo. Pri prehodu iz ene v drugo se nujno postavlja tudi vprašanje, kako olajšati postopke v zvezi s spremembo ontologije, ki je za znanstvenoraziskovalno sfero vse prej kot statična.

8 VIRI IN LITERATURA

[1] *XML*. (2000). Pridobljeno iz http://www.w3.org/TR/2000/REC-xml-20001006.

[2] *URI*. (2001). Pridobljeno iz http://www.w3.org/TR/uri-clarification/.

[3] *RDF*. (2004). Pridobljeno iz http://www.w3.org/TR/REC-rdf-syntax/.

[4] *Nacionalni Raziskovalni in Razvojni Program: Uradni list RS 3/2006 Resolucija o nacionalne raziskovalnem in razvojnem programu za obdobje 2006 – 2010 (ReNRRP)*. (2006). Pridobljeno iz http://www.uradni-list.si/1/ulonline.jsp?urlid=20063&dhid=80293.

[5] *EUROCRIS Task Groups* . (2010). Pridobljeno iz http://www.eurocris.org/public/about-eurocris/organisation/taskgroups/.

[6] *DATASTAR*. (2011). Pridobljeno iz http://ds.datastarweb.com/datasheets/.

[7] *ARRS – Agencija RS za raziskovalno dejavnost*. (2012). Pridobljeno iz http://www.arrs.gov.si/sl/.

[8] *CERIF: a service hosted by CORDIS*. (2012). Pridobljeno iz http://cordis.europa.eu/cerif/src/copyright.htm.

[9] *COBISS – Kooperativni bibliografski sistem in servisi*. (2012). Pridobljeno iz http://www.cobiss.si.

[10] *DIALOG*. (2012). Pridobljeno iz http://www.dialog.com/.

[11] *EUROCRIS*. (2012). Pridobljeno iz http://www.eurocris.org/Index.php?page=homepage&t=1.

[12] *KAON2*. (2012). Pridobljeno iz http://kaon2.semanticweb.org/.

[13] *Ontobroker*. (2012). Pridobljeno iz http://www.ontoprise.de/de/en/home/products/ontobroker.html

[14] *OWL*. (2012). Pridobljeno iz http://www.w3.org/TR/2004/REC-owl-guide-20040210/#Privacy.

[15] *SICRIS – Informacijski sistem o raziskovalni dejavnosti v Sloveniji*. (2012). Pridobljeno iz http://sicris.izum.si/.

[16] *SPARQL*. (2012). Pridobljeno iz http://www.w3.org/TR/rdf-sparql-query/.

[17] Broughton, V. (2006). *Essential thesaurus construction (ISBN: 1-85604-565-X).*

[18] Haehnle, R. (2001). *Tableaux and Related Methods.*

[19] Hendler, J., Lassila, O. & Berners-Lee, T. (2001). The Semantic Web. Scientific American. *284(5)*, str. 34–43.

[20] Jeffery, K. & Asserson, A. (2006). Supporting the research process with a CRIS. Proceedings of the 8th International conference on current research information systems. Bergen.

[21] Joerg, B., Jeffery, K., Van Grootel, G., Assesron, A., Dvorak, J. & Rasmussen, H. (2010). *CERIF 2008 – 1.2 Full Data Model (FDM) Introduction and Specification*. Pridobljeno iz http://www.eurocris.org/Uploads/Web%20pages/CERIF2008/Release_1.2/CERIF2008_1.2_FDM.pdf.

[22] Joerg, B., Krast, O., Jeffery, K. & Van Grootel, G. (2007). *CERIF2006XML-1.1 Data Exchange Format Specification.*

[23] Lassila, O. & Swick, R. (1999). *Resource Description Framework (RDF) model and syntax specification.* Pridobljeno iz http://www.w3.org/TR/PR-rdf-syntax/.

[24] Lavbič, D. & Krisper, M. (2005). Semantika podatkov in ontologije. *Uporabna informatika, XIII* (3, julij/avgust/september).

[25] Noy, N. F. & McGuinness, D. (2001). *Ontology Development 101: A Guide to Creating Your First Ontology. Stanford Knowledge Systems Laboratory Technical Report KSL-01-05 and Stanford Medical Informatics Technical Report SMI-2001-0880.*

▪

Aleš Bošnjak je vodja oddelka za izobraževanje na Institutu informacijskih znanosti v Mariboru. V preteklosti je vodil oddelek SICRIS (Slovenian Current Research Information System) in bil v letih 2007–2009 član upravnega odbora evropske organizacije EUROCRIS (European Current Research Information Systems). Je avtor nekaterih člankov s tega področja, bile je član programskega odbora, vodja delovne skupine Best Practice in soorganizator konference EUROCRIS leta 2008. Trenutno je študent doktorskega študija na Fakulteti za elektrotehniko, računalništvo in informatiko Univerze v Mariboru, kjer se ukvarja z raziskovalnimi področji, kot so informacijski sistemi o raziskovalni dejavnosti, področja uporabe ontologij, modeliranje s pomočjo omrežij, semantično obogateni iskalniki.

▪

Vili Podgorelec je izredni profesor s področja informatike na Fakulteti za elektrotehniko, računalništvo in informatiko Univerze v Mariboru, kjer predava na programih Računalništvo in informatika, Informatika in tehnologije komuniciranja, Medijske komunikacije in Bioinformatika. Raziskovalno se ukvarja predvsem s področji inteligentnih sistemov, inovativnih informacijskih rešitev, semantičnih tehnologij in teorije kompleksnosti, ki jih aplicira predvsem v programskem inženirstvu in medicinski informatiki. Je avtor mnogih člankov z omenjenih raziskovalnih področij v uglednih mednarodnih revijah, vabljeni predavatelj na več konferencah ter predsednik oz. član programskih odborov in soorganizator nekaj mednarodnih konferenc. Sodeloval je v več domačih in mednarodnih znanstveno-raziskovalnih projektih ter v aplikativnih projektih za industrijo.