UNIVERZA V NOVI GORICI
POSLOVNO-TEHNIŠKA FAKULTETA

MODELIRANJE ZNANJA IN PRIKAZ V OBLIKI ONTOLOGIJ S ŠTUDIJO PRIMERA RAZISKAV V OKOLJU

MAGISTRSKO DELO

Jaka Vogrinčič Bizjak

Mentorica: doc. dr. Ingrid Petrič

Nova Gorica, 2014

ZAHVALA

Za strokovno sodelovanje in lepe besede se iskreno zahvaljujem mentorici doc. dr. Ingrid Petrič, ki mi je vedno priskočila na pomoč, predsednici komisije in dekaniji prof. dr. Tanji Urbančič za pomoč pri zaključnem pregledu magistrske naloge in za zaključek še zahvala članu in mentorju moje diplomske naloge prof. dr. Bojanu Cestniku, prav tako za zaključni pregled magistrske naloge.

Zahvala gre tudi moji družini, ženi Maji ter hčerkama Kaji in Jani za podporo v času študija.

Cilj, ki sem si ga zastavil sem ga s podporo vseh dosegel. Ker je znanje potrebno vedno nadgrajevati si moram sedaj postaviti nov cilj in ga uresničiti.

Hvala vsem

NASLOV

Modeliranje znanja in prikaz v obliki ontologij s študijo primera raziskav v okolju

IZVLEČEK

Magistrsko delo se osredotoča na področje modeliranja znanja in njegovega prikaza v obliki terminoloških ontologij. V delu prikazujemo uporabnost tehnologij znanja za analizo večjih besedilnih zbirk. Kot primer predstavljamo analizo objav znanstvenih in strokovnih člankov raziskovalne skupine Laboratorija za raziskave v okolju Univerze v Novi Gorici, ki je eden najstarejših raziskovalnih laboratorijev univerze in največji po številu zaposlenih. Podatke o objavah laboratorija smo pridobili s pomočjo Informacijskega sistema o raziskovalni dejavnosti v Sloveniji - SICRIS. V študijo primera so bili zajeti naslovi objav članov Laboratorija za raziskave v okolju za obdobje 30 let, in sicer od leta 1983, ko se v SICRIS-ovi bazi pojavijo prve objave raziskovalcev, vključenih v laboratorij, do konca leta 2012. Podatke iz informacijskega sistema SICRIS smo uredili v zbirko naslovov objav formata golo besedilo. Besedilno analizo naslovov člankov, ki so jih objavili člani laboratorija, smo izdelali s podporo računalniškega programa OntoGen, ki omogoča polavtomatsko gradnjo ontologij iz besedil. Z modeliranjem domenskega znanja z gradnjo ontologij iz naslovov člankov smo ugotavljali ključna raziskovalna področja, ki se odražajo iz objav laboratorija. Rezultat analize, ki smo jo izvedli, je podroben prikaz področij, na katerih je bila v preučevanem obdobju 30 let raziskovalna dejavnost članov laboratorija najbolj plodna v smislu objav. V obravnavanem obdobju je bilo največ objav s področja spektrometrije na podlagi termičnih leč. Po številčnosti sledijo objave o sol-gel postopkih in tankih premazih. V večini raziskav je bila prisotna tudi problematika obremenjenosti okolja.

KLJUČNE BESEDE

modeliranje znanja, besedilna analiza, ontologija, OntoGen, raziskave v okolju

TITLE

Knowledge modelling and ontology-based knowledge representation: an environmental research case study

ABSTRACT

This master thesis is focusing on knowledge modelling and knowledge representation with terminological ontologies. In the thesis we show the advantages of the usage of knowledge technologies for the analysis of large text collections. As an example, we present the analysis of scientific and professional articles published by the research team of the Laboratory for environmental research of the University of Nova Gorica. The laboratory is one of the oldest research groups of the university with the largest number of employees. We obtained the data about the laboratory's publications from the Slovenian Current Research Information System - SICRIS. In the study case, we included publications' titles of articles published by members of the Laboratory for environmental research in the period of the last 30 years, from 1983, when the first publications of the laboratory's researchers appeared in the SICRIS' database, until the end of 2012. We organized the data from the SICRIS information system into a collection of the articles' titles in a plain text format. We performed the text analysis of the titles of articles published by the laboratory's members, with the support of a computer programme called OntoGen that enables the semi-automatic construction of ontologies from texts. By modelling the domain knowledge in the form of ontologies based on the titles of articles, we aimed to investigate the main research areas, which can be reflected from the laboratory's publications. The result of the analysis, that we performed, is a detailed representation of areas in which the research activity of the laboratory was the most productive in terms of publications during the examined period of the last 30 years. In this period, the majority of publications dealt with thermal lens spectrometry, followed by publications about sol-gel procedures and thin films. One of the most important areas of research was also the environmental pollution.

KEYWORDS

knowledge modelling, text analysis, ontology, OntoGen, environmental research

KAZALO

KAZALO SLIK

1 UVOD

Razvoj družbe tretjega tisočletja temelji na znanju ter sposobnosti njegove implementacije v gospodarstvo in druga področja človekovega delovanja. Zato je pomembno, da znamo lastno ali pridobljeno znanje ustrezno predstaviti ter pritegniti pozornost zanj pri tistih posameznikih in organizacijah, ki lahko to znanje koristno uporabijo pri svojem delu. Modeliranje znanja omogoča celovit pregled nad obravnavano problematiko ali proučevano domeno, zato se magistrsko delo osredotoča na področje modeliranja znanja in njegovega prikaza v obliki terminoloških ontologij.

V magistrskem delu smo pridobljena teoretična spoznanja in študijske izkušnje s področja upravljanja znanja uporabili za izgradnjo ontoloških prikazov na primeru raziskovalnega dela Laboratorija za raziskave v okolju, Univerze v Novi Gorici. Cilj magistrskega dela je bil analizirati znanstvene in strokovne objave skupine raziskovalcev Laboratorija za raziskave v okolju, ki opravlja temeljne in uporabne raziskave na različnih področjih preučevanja in varovanja okolja. Z računalniško podprto besedilno analizo smo omogočili celovit pregled nad objavami del te skupine in podrobneje prikazali področja, na katerih je raziskovalna dejavnost laboratorija najbolj plodna v smislu objav. Hkrati smo identificirali tista področja, kjer bi z vidika raziskovalne dejavnosti skupine sicer pričakovali številnejše objave raziskav, a je objavljenih člankov v obravnavanem obdobju manj.

Z orodji za obdelavo in analizo besedil smo analizirali znanstvene in strokovne objave ter gradili modele znanja na primeru raziskav v okolju. Podatke smo pridobili iz Informacijskega sistema o raziskovalni dejavnosti v Sloveniji - SICRIS. V študijo primera so bili zajeti naslovi objav sodelavcev Laboratorija za raziskave v okolju Univerze v Novi Gorici. Zajeto je bilo obdobje 30 let, in sicer od leta 1983, ko se v SICRIS-ovi bazi pojavijo prve objave raziskovalcev vključenih v laboratorij, do konca leta 2012. Podatke iz informacijskega sistema SICRIS smo z urejevalniki besedil uredili v zbirko naslovov objav formata golo besedilo. Ta je bila osnova za besedilno analizo, ki smo jo izvedli z računalniškim orodjem OntoGen (Fortuna in sod., 2006), ki omogoča polavtomatsko gradnjo modelov znanja v obliki ontologij. OntoGen omogoča rudarjenje besedila in prikaz ključnih konceptov iz besedila v

obliki hierarhičnih struktur. Njegov uporabniški vmesnik je hiter, logičen in preprost za uporabnika programa.

V obravnavanem obdobju 30 let je bilo največ objav sodelavcev Laboratorija za raziskave v okolju na področju spektrometrije na podlagi termičnih leč. Po številčnosti sledijo objave o sol-gel postopkih in tankih premazih. V večini raziskav je bila prisotna tudi problematika obremenjenosti okolja zaradi posledic onesnaževanja.

V magistrskem delu najprej opišemo raziskovalno dejavnost Laboratorija za raziskave v okolju, ki obsega temeljne raziskave, aplikativne raziskave in ekspertno dejavnost. V tretjem poglavju predstavimo modeliranje znanja, od zajema podatkov do pridobivanja informacij in prikazov znanja. V četrtem poglavju opredelimo prikaze znanja v obliki ontologij, predstavimo klasifikacijo ontologij in pomen terminoloških ontologij pri modeliranju znanja. V petem poglavju prikažemo študijo primera raziskav v okolju, kjer podrobneje opišemo pripravo podatkov za besedilno analizo in sam potek rudarjenja besedil z računalniškim orodjem OntoGen. V šestem poglavju predstavimo rezultate besedilne analize naslovov člankov, raziskovalcev Laboratorija za raziskave v okolju in poglavje zaključimo z razpravo o dobljenih rezultatih besedilne analize. Sledi zaključek magistrskega dela, v katerem povzamemo glavne ugotovitve.

2 RAZISKOVALNA DEJAVNOST LABORATORIJA ZA RAZISKAVE V OKOLJU

Laboratorij za raziskave v okolju je začel delovati v letu 1995, ko je bila ustanovljena Fakulteta za znanosti o okolju, predhodnica Politehnike Nova Gorica in današnje Univerze v Novi Gorici. Prva leta delovanja je bila dejavnost laboratorija usmerjena v razvoj novih instrumentalnih metod za meritve onesnaženosti okolja. Dejavnost je bila v veliki meri odvisna od opreme na Institutu Jožef Stefan, ki jo je laboratorij takrat uporabljal. Prav tako je laboratorij sodeloval s tujimi univerzami, predvsem z Univerzo v Wageningenu na Nizozemskem (Bratina (ur.), 2005).

Razvojno prelomnico v delovanju laboratorija predstavlja leto 1997, ko je laboratorij pridobil nove prostore, v katerih je bila postavljena oprema za raziskave kemijskih pojavov v okolju. Raziskovalna dejavnost se je hitro razširila na druga področja raziskav, ki zajemajo študije fotokemijske in mikrobiološke razgradnje ter prevoza onesnaževal v okolju, študije vsebnosti antioksidantov v živilih ter razvoj laserskih in bioanalitskih metod kot tudi ekotoksikoloških testov za ugotavljanje prisotnosti strupenih snovi v okolju in njihovih vplivov na organizme.

Delovanje laboratorija se je kasneje razširilo še na pripravo in karakterizacijo novih materialov, ki imajo možnost uporabe v okolju prijaznih sistemih. V to dejavnost sodijo membrane za gorivne celice, metode za sintezo biodizla ter v zadnjem času raziskave na področju zatiranja škodljivcev na sadnem drevju, vinski trti, krompirju in žitaricah (Univerza v Novi Gorici, 2012).

Laboratorij šteje trenutno skupaj več kot 20 zaposlenih doktorjev znanosti in mladih raziskovalcev iz Slovenije in tujine. Sodelavci laboratorija aktivno sodelujejo pri izvajanju dodiplomskih in podiplomskih študijskih programov na Univerzi v Novi Gorici, kjer imajo poleg predavanj in laboratorijskih vaj tudi pomembno vlogo mentorstva bodočim diplomantom, magistrantom in doktorantom. Laboratorij s svojo opremo nudi študentom priložnost za eksperimentalno delo.

Raziskovalno in eksperimentalno delo v laboratoriju poteka na sodobni raziskovalni opremi, ki obsega plinske kromatografe z detektorji na zajetje elektronov, plamensko ionizacijskimi detektorji in masno-selektivnimi detektorji, tekočinskimi kromatografi

visoke ločljivosti s spektrofotometrijskim detektorjem z diodno matriko in fluorescenčnim detektorjem, ionski kromatograf, osnovne in frekvenčno podvojene Ar-laserje ter ekscimerne in barvilne sunkovne laserje. Sredstva za opremo je laboratorij pridobil predvsem iz mednarodnih projektov Evropske Unije ter s sofinanciranjem Ministrstva za visoko šolstvo, znanost in tehnologijo Republike Slovenije.

Laboratorij je v letu 2013 sodeloval v raziskovalnih projektih in programih s Kemijskim inštitutom Ljubljana, Gozdarskim inštitutom Slovenije, Biotehniško fakulteto Univerze v Ljubljani, Univerzo Blaise Pascal Clermont Ferrand iz Francije, Moskovsko državno Univerzo, Univerzo v Padovi, Univerzo Greenwich ter številnimi drugimi slovenskimi in mednarodnimi ustanovami (Laboratorij za raziskave v okolju, 2013). V nadaljevanju so podrobneje predstavljene temeljne raziskave, aplikativne raziskave in ekspertna dejavnost laboratorija.

2.1 Temeljne raziskave

Na področju temeljnih raziskav je dejavnost Laboratorija za raziskave v okolju usmerjena v študij fotokemijske razgradnje različnih organskih onesnaževal v vodnem okolju in transporta onesnaževal, raziskovanje strupenosti različnih organskih onesnaževal za organizme v vodnem in kopenskem okolju, razvoj laserskih in bioanalitskih metod za ugotavljanje prisotnosti toksičnih snovi v okolju ter razvoj novih materialov za uporabo v okolju prijaznih tehnologijah (Laboratorij za raziskave v okolju, 2013)

Spodbudne rezultate je laboratorij dosegel na področju fotokatalize, kjer so bile raziskave usmerjene predvsem v razvoj novih tankih plasti in prahov titanovega dioksida ter uporabo le teh kot fotokatalizatorjev za razgradnjo substituiranih fenolov, tekstilnih in drugih barvil ter farmacevtskih izdelkov s svetlobo iz ultravijoličnih svetilk in pod vplivom sončne svetlobe (Univerza v Novi Gorici, 2012). Laboratorij deluje tudi na področju naprednih tehnik oksidacijskega čiščenja pitnih in odpadnih voda.

Pomembno vlogo v mednarodnem merilu ima laboratorij tudi pri razvoju visoko občutljivih laserskih metod kemijske analize na podlagi termičnih leč v kombinaciji z

bioanalitskimi metodami (acetilholinesterazni in metalotioneinski biosenzorji), metodami pretočno injekcijske analize in tekočinske kromatografije. Temeljne raziskave so usmerjene tudi v tehnike resonance površinskih plazmonov za odkrivanje kovinskih ionov in organskih onesnaževalcev v vodah. Pri tem so bile uporabljene različne imobilizirane biomolekule, kot na primer encimi. Poleg odkrivanja strupenih snovi je laboratorij uporabljal navedene metode tudi za merjenje prisotnosti biološko aktivnih snovi, kot na primer antioksidantov (karotenoidi, polifenoli in bilirubin). Raziskave karotenoidov so se nanašale na fotokemisjke cikle nekaterih amfibijskih rastlin in fitoplanktona ter razpada planktonskih celic v času cvetenju morja. Z visoko občutljivo detekcijo bilirubina so potekale raziskave transporta antioksidantov prek celične membrane in vloge transportnih proteinov pri tem.

Raziskave so vključevale tudi ostale za človeka pomembne antioksidante. Proučevali so količino in sestavo antocianov in hidroksicimetnih kislin predvsem v nekaterih lokalno razširjenih sortah češenj in belega grozdja. Raziskovali so vpliv okoljskih dejavnikov, kot so na primer lega, vremenske razmere ter obdelava vinograda in sadovnjaka, na vsebnost polifenolov v teh sortah.

Laboratorij je uporabljal sodobno lasersko tehnologijo pri meritvah onesnaženosti zraka in spremljanju fizikalnih lastnosti atmosfere na daljavo. S tehniko laserskega skeniranja (angl. Light Detection And Ranginig – LIDAR) so preučevali transport aerosolov v okolici avtocest in drugih prometnic. Z dobljeno dvodimenzionalno sliko so lahko določili emisijske vrednosti onesnaževalcev, profil vetra ter difuzijske lastnosti atmosfere.

Z ekotoksiološkimi raziskavami so v laboratoriju preučevali strupenost organofosfatnih in neonikotinoidnih pesticidov za kopenske nevretenčarje in raziskovali možnost uporabe teh organizmov za merjenje obremenjenosti okolja s pesticidi. Vplive pesticidov so merili na osnovi sprememb aktivnosti encimov, kot sta acetlilholinesteraza in glutation-S-transferaza ter sprememb energijskih zalog, kot na primer lipidov in glikogena v preiskovanih organizmih.

Med področja, s katerimi se je laboratorij raziskovalno ukvarjal sodijo tudi raziskave ekološke energetike kopenskih ekosistemov. Na osnovi načel sistemske ekologije, tj.

integritete oz. neokrnjenosti preučevanih sistemov, so izvajali ocenjevanje trajnosti procesov v okolju, predvsem procesov rasti in razvoja ekosistemov ter samo-organizirajočih lastnosti evolucije ekosistemov. Pri tovrstnih raziskavah so uporabili moderne tehnike zaznavanja termalnih značilnosti površine ekosistemov na daljavo in ocenjevanja informacij shranjenih v genomu. Z različnimi orodji za ekološko modeliranje, kot sta STELLA in EcoPath, so ocenjevali odnose med sistemskimi indikatorji ter rastjo in razvojem ekosistemov (Bratina (ur.), 2005).

2.2 Aplikativne raziskave

Za prenos znanstvenih ugotovitev v prakso, se laboratorij ukvarja tudi z aplikativnimi raziskavami, kjer preučujejo predvsem možnosti za uporabo optotermične spektometrije in bioanaliznih metod za uporabo v medicinski diagnostiki ter za kontrolo kakovosti in neoporečnosti živil (Laboratorij za raziskave v okolju, 2013). Z novo razvitimi metodami so uspeli določiti značilne antioksidante iz skupine karotenoidov, ki služijo kot indikatorji pristnosti, čistosti in kakovosti oljčnih in drugih rastlinskih olj ter sadnih sokov (Bratina (ur.), 2005). Za širšo uporabo v proizvodnji sadnih sokov so testirali acetilholinesterazne optotermične biosenzorje, ki so se pokazali kot hitri in cenovno ugodni presejalni testi za ugotavljanje prisotnosti organofosfatnih pesticidov v sadnih kašah.

Laboratorij je intenzivno izvajal aplikativne raziskave tudi na reki Soči. Pri vodnem zajetju Mrzlek, ki je vir pitne vode za novogoriško območje, so poleg onesnaženosti vode in sedimentov s težkimi kovinami in strupenimi organskimi snovmi proučevali tudi možnosti onesnaženja pitne vode pri tem zajetju ter s tem povezanih zdravstvenih posledic za prebivalstvo (Bratina (ur.), 2005).

Posebno področje aplikativnih raziskav laboratorija z veliko potencialno uporabnostjo predstavljajo samočistilne površine. Na tem področju je bilo delo laboratorija usmerjeno v pripravo stabilnih koloidnih raztopin za nanos prozornih delcev titanovega in silicijevega dioksida (TiO_2-SiO_2) v obliki prevlek na steklu ali drugih anorganskih podlagah. Pri teh postopkih po odhlapitvi topil ostane na površini tanka plast samočistilne keramike. Zaradi aktivnosti katalitske tanke plasti TiO_2-SiO_2 se organska umazanija, kot so ostanki maščobe in saje, pod sončno svetlobo razgradi na ogljikov dioksid in vodo, medtem ko se anorganske snovi, kot je na primer pesek

na takšni površini enostavno sperejo z dežjem. Te tanke plasti so v običajnih okoljskih pogojih stabilne več let (Laboratorij za raziskave v okolju, 2013).

2.3 Ekspertna dejavnost

Pomembna dejavnost laboratorija je tudi ekspertna dejavnosti, ki je vključevala predvsem presoje vplivov na okolje. Sodelavci laboratorija so v preteklosti sodelovali pri pripravi strokovnih podlag za postavitev naprave za zmanjšanje emisij odpadnih plinov iz proizvodnje anhidrida ftalne kisline v podjetju Kemiplast Koper. Raziskava je vključevala ocenjevanje primernosti razpoložljivih tehnologij in obstoječega stanja okolja, metod nadzora in spremljanja ter svetovanje za pridobitev soglasja za postavitev naprave (Bratina (ur.), 2005).

Ekspertno sodelovanje je potekalo tudi z upravitelji odlagališča komunalnih odpadkov Stara Gora pri namestitvi naprav za predelavo trdnih komunalnih odpadkov, izkoriščanje deponijskega plina in čiščenje izcednih in odpadnih vod ter s hidroelektrarnama Plave II in Doblar II pri spremljanju vpliva na okolje med doinstalacijo teh dveh hidroelektrarn. Laboratorij sodeluje tudi v številnih mednarodnih projektih povezanih s presojo vpliva na okolje. Tako so na primer v sodelovanju z italijanskimi strokovnjaki sodelovali v strokovni komisiji, ki jo je imenovala Pokrajina Gorica za oceno tveganja severovzhodnega dela mesta italijanske Gorice zaradi onesnaženosti s formaldehidom, cinkom, bakrom in nekaterimi drugimi onesnaževalci (Univerza v Novi Gorici, 2010).

3 MODELIRANJE ZNANJA

3.1 Od podatkov do informacij in znanja

Podatki so neorganizirana, neobdelana dejstva, ki nam sama zase ne povedo ničesar, dokler jih ne ustrezno obdelamo. Šele, ko podatke obdelamo in/ali primerjamo med seboj ugotovimo njihov pomen. Podatki nam ne povedo ničesar o motivaciji, kvaliteti, ali značilnosti analiz, s katerimi so bili pridobljeni. Podatki so predpogoj za informacije, ki jih pridobimo kot rezultat raziskav. Brez ustreznih podatkov ne moremo začeti analize. Problem nastane, ko imamo preveč podatkov in takrat ne vemo, katere odločitve bi morali sprejeti, da bi prišli do želenega cilja.

Ko damo podatkom pomen, ki ga bo razumel prejemnik, kateremu so podatki namenjeni, dobimo informacijo. Informiranje je združevanje podatkov, ki nas pripeljejo do lažje odločitve. Pri neurejenih podatkih informacije pomagajo razumeti razmerja med njimi in dajejo podatkom pomen. Podatke lahko reorganiziramo, statistično analiziramo, odstranimo napake ali drugače obdelamo, da bi dobili njihov pomen. Šele ko ljudje naberemo dovolj informacij in na osnovi zbranih informacij razumemo nek pojav ali problem, govorimo o znanju (Awad in Ghaziri, 2007).

Znanje je bilo vedno zelo pomembno za človeški napredek. Naši predniki so morali veliko raziskovati in preizkušati, da so lahko razkrili vse, kar nam je danes poznano in se še odkriva. Znanje je višji nivo abstraktnega, ki je v človeškem umu. Zato je znanje veliko težje razumeti kot podatke in informacije. Ljudje iščemo znanje, ker nam pomaga pri delu. Znanje ima več pomenov, odvisno od tega kje se uporablja. Znanje se opira na učenje, razmišljanje in iskanje podobnih problemov med študijem. Pomembne so tudi izkušnje, ki jih z leti pridobimo. Znanje ni zgolj informacija kakor tudi informacija niso zgolj podatki. Znanje izvira iz informacije na enak način kot informacija izvira iz podatkov.

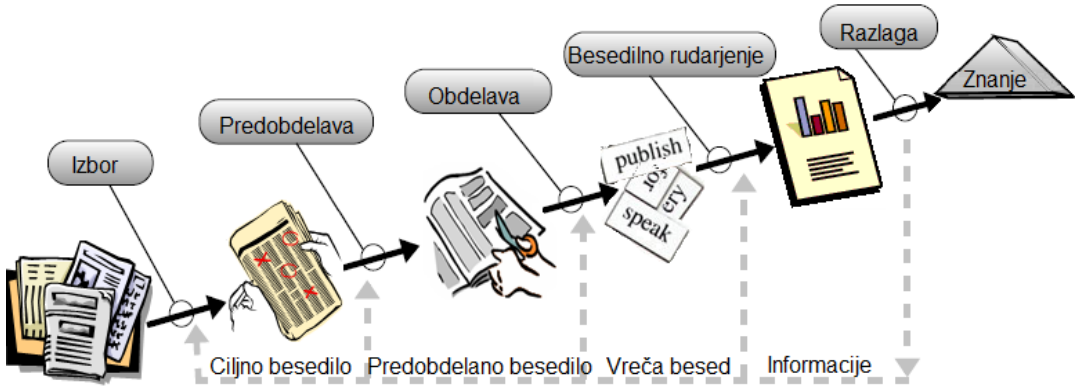
3.2 Prikaz znanja

Sodobne baze podatkov ponujajo hitro rastoče količine besedilnih in drugih formatov podatkov. Iskanje želenih podatkov in pridobivanje informacij na njihovi osnovi je med tako množico podatkov oteženo, še posebej, če si želimo vpogled v področja, na katerih nismo strokovno izobraženi. Zato narašča potreba po metodah in orodjih, ki

bi iz množice podatkov izluščila uporabno znanje na uporabniku čim bolj prijazen način. Procesi odkrivanja uporabnega znanja iz podatkov običajno temeljijo na podatkovnem rudarjenju, ki z uporabo posebnih računalniških algoritmov omogoča pridobivanje in prikaz smiselnih vzorcev iz vhodnih podatkov (Fayyad in sod., 1996).

Procesi odkrivanja znanja lahko temeljijo tudi na zbirkah besedil. Postopku, ki iz vhodnih besedil izlušči strukturirane informacije, pravimo rudarjenje besedila. Običajno se izvaja z besedili, ki so zapisana v obliki naravnega jezika.

Podobno kakor lahko baze podatkov analiziramo s podatkovnim rudarjenjem, se v primeru baz besedil, kakršne so na primer bibliografske baze zapisov o znanstvenih in strokovnih objavah, poslužujemo postopkov besedilnega rudarjenja. Primer procesa besedilnega rudarjenja je prikazan na Sliki 1, kjer je za predstavitev izrazov iz zbirke vhodnih besedil uporabljena t.i. vreča besed (angl. Bag of words), ki predstavi analizirana besedila tako, da izlušči iz njih izraze in pri tem ne upošteva slovničnih pravil in vrstnega reda besed (Sebastiani, 2002).



Slika 1: Postopek rudarjenja besedil prikazan v Petrič in sod. (2010) skladno z opredelitvijo odkrivanja znanja iz baz podatkov po Fayyad in sod. (1996)

Preden zgradimo uporabne prikaze znanja, se moramo seznaniti s shranjenimi strukturami podatkov, njihovimi medsebojnimi povezavami in pomenom. Prvi korak, ki ga moramo narediti je, da zberemo različne skupine podatkov in iz njih prikažemo statistično pomembne podatke. Raziskovanje podatkov pomeni iskanje prikaza

značilnosti za skupino, ki jo obdelujemo. To vključuje naslednje korake (Awad in Ghaziri, 2007):

- prikaz ključnih atributov,
- identifikacija osamelcev (ang. outliers), ki so izven pričakovane množice rezultatov,
- opredelitev začetnih hipotez in napoved nadaljnjih ukrepov,
- izbor zanimivih skupin podatkov za nadaljnje raziskave.

Sodobna informacijska orodja, ki omogočajo avtomatsko ali polavtomatsko analizo in prikaz podatkov, nam lahko zelo olajšajo delo pri tem. Eno takih orodij, imenovano OntoGen, smo uporabili za besedilno analizo naslovov člankov raziskovalne skupine Laboratorija za raziskave v okolju in za prikaz rezultatov v obliki vsebinskih hierarhij, ki jih imenujemo ontologije.

4 ONTOLOGIJE

4.1 Opredelitev in klasifikacija ontologij

Ontologija je predstavitev množice konceptov nekega področja človeškega znanja, ki jih lahko kategoriziramo. Z ontologijo prikažemo tudi razmerja med koncepti. Za področja, ki jih lahko natančneje opredelimo, tako ontologije združujejo objekte in pravila, ki prikazane objekte povezujejo v ontologijo. Izraz ontologija izhaja iz filozofije in se pogosto uporablja za opredelitev filozofskih disciplin za prikaz osnove, vzrokov in splošnih značilnosti stvarnosti. Z ontologijami so predstavljene tudi nekatere teorije o naravi bitij in njihovem obstoju (Lavbič in Krisper, 2005).

Ontologije se lahko uporabi kot model formalne strukture za enotno predstavitev in semantiko informacij. S sklicevanjem na ontologije se lažje vzpostavi dialog med različnimi agenti in doseže jasno predstavitev informacij, ki se jih ti agenti poslužujejo (Lavbič in Krisper, 2005). Z vidika informacijske rabe ontologij, se ontologija nanaša na formalni opis pomembnih konceptov obravnavanega področja, ki omogoča skupno razumevanje tega področja človeku in računalniškim sistemom (Islovar, 2013). V tem smislu imajo ontologije velik potencial predvsem na področjih obvladovanja znanj, zbiranja informacij in združevanja inteligentnih sistemov ter na področju elektronskega trgovanja (Lavbič in Krisper, 2005).

Glede na stopnjo formalnosti so ontologije deskriptivne, formalne ali formalizirane (Poli in Seibt, 2010). Deskriptivno ontologijo določajo izrazi, ki se pogosto uporabljajo za predstavitev znanja in opis določenega tematskega področja. Formalno ontologijo sestavljajo le formalno sprejeta terminologija deskriptivnih ontologij, ki označuje različne vidike ali tipe nekega obravnavanega področja. Formalizirane ontologije vsebujejo formalne specifikacije nekega področja v najožjem pomenu formalizacije predstavitve (npr. s soglasjem strokovnjakov z obravnavanega področja). Formalizirane ontologije so slovarji pojmov (besed in njihovih sinonimov), ki pomensko strukturirajo določeno področje in hkrati omogočajo prikaz hierarhične strukture relacij med pojmi.

Guarino in Giaretta (1995) navajata več pogostih različnih interpretacij pojma ontologije:

1. Ontologija kot filozofska disciplina, ki se ukvarja z naravo in organizacijo stvarnosti;
2. Ontologija kot neformalni semantični (pomenski) konceptualni sistem določene baze znanja;
3. Ontologija kot prikaz formalne semantične strukture baze znanja;
4. Ontologija kot specifikacija konceptualizacije, pogosto uporabljena na področju umetne inteligence;
5. Ontologija kot logični prikaz konceptualnega sistema s pomočjo sintakse (skladenjskih pravil, ki določajo strukturo stavkov);
6. Ontologija kot slovar logičnih definicij;
7. Ontologija kot specifikacija izrazov z meta oznakami v teoriji določene domene.

V nadaljevanju prikazujemo in obravnavamo ontologije v smislu neformalnih semantičnih (pomenskih) konceptualnih prikazov domenskega znanja. Sowa (2000) opredeljuje take ontologije kot terminološke, ki prikazujejo splošne koncepte in relacije določenega domenskega znanja, za razliko od formalnih ontologij, ki natančno prikazujejo domenske izraze in specifične tipe relacij (kot na primer *je-del*, *je-primerek*) med njimi.

4.2 Pomen terminoloških ontologij pri modeliranju znanja

Vsaka terminološka ontologija predstavlja specifičen pogled na določeno področje znanja (Lacasta in sod., 2010). Terminološke ontologije predstavimo z besedami in besednimi zvezami v naravnem jeziku, medtem ko so formalne ontologije predstavljene v logičnem jeziku, zasnovane na pojmovnih slovarjih in stavkih, ki prikazujejo razmerja med pojmi. Ontologije, ki predstavljajo domensko znanje, omogočajo skupno razumevanje domene (Lavbič in Krisper, 2005), ki olajša sporazumevanje med ljudmi (terminološke ontologije) in računalniškimi sistemi (formalne ontologije). Gruninger in Lee (2002) povzemata uporabnost ontologij še z nekaterih drugih vidikov:

1. za sporazumevanje:

- med računalniškimi sistemi,

- med ljudmi,

- med ljudmi in računalniškimi sistemi;

2. za računalniško sklepanje:

   - za načrtovanje in predstavitev informacij,

   - za analizo notranjih struktur, algoritmov ter vhodnih podatkov in rezultatov implementiranih sistemov v teoretičnem in konceptualnem smislu;

3. za ponovno uporabo in organizacijo znanja:

   - za strukturiranje in organizacijo knjižnic ali repozitorijev načrtovanih informacij in domenskega znanja.

Ontologije, kakršna je na primer ontologija medicinskih izrazov in postopkov, ki je rezultat večletnega projekta GALEN (Rector in Nowlan, 1994), imajo veliko izrazno moč pri zajemu in predstavitvi semantike z obravnavanega področja, omogočajo pa tudi sklepanje na podlagi znanja (Lavbič in Krisper, 2005).

V literaturi lahko najdemo besedilne analize z ontološkimi prikazi znanja na različnih znanstvenih in strokovnih področjih, kot na primer na področju poslovne inteligence, (Drelichowski in sod., 2012), fizike (Ginsparg, 2004), bioinformatike (Janssens in sodelavci, 2007) in ekonomije (Vogrinčič in Bosnić, 2011). V nadaljevanju je predstavljeno strukturiranje znanja s prikazom v obliki ontologij na osnovi besedilne analize naslovov člankov, ki smo jo izvedli za področje raziskav v okolju.

5 ŠTUDIJA PRIMERA RAZISKAV V OKOLJU

Podatki, ki smo jih zajeli v raziskavo, se nanašajo na objave od leta 1983 do konca leta 2012 in vsebuje naslove znanstvenih objav Laboratorija za raziskave v okolju. Z orodji za obdelavo in analizo besedil smo obdelali bibliografijo raziskovalne skupine 1540-001 – Laboratorij za raziskave v okolju za obdobje zadnjih tridesetih let, od 1983 do vključno 2012. Podatke smo pridobili s pomočjo Informacijskega sistema o raziskovalni dejavnosti v Sloveniji - SICRIS. Kriteriji po katerih smo iskali podatke v SICRIS-ovi bazi so prikazani na sliki 2. Zaradi potreb besedilne analize naslovov objav smo izbrali format bibliografske enote ISO in format izpisa XML. Izpis bibliografskih enot je uvrščen v vedo naravoslovja in spada med znanstvena dela. Na ta način smo dobili 366 naslovov znanstvenih bibliografskih enot.

Slika 2: Kriteriji iskanja objav na spletni strani SICRIS

5.1 Priprava podatkov za besedilno analizo

Podatke, ki smo jih pridobili iz informacijskega sistema SICRIS, smo morali najprej urediti v zbirko naslovov objav za besedilno analizo s programom OntoGen. Da bi podatke lahko obdelali s programom OntoGen, smo pripravili datoteko formata golo besedilo, v kateri vsaka posamezna vrstica vsebuje bibliografske podatke posamezne objave. V našem primeru posamezna vrstica vhodne datoteke vsebuje naslov posamezne objave članov Laboratorija za raziskave v okolju. V ta namen smo podatke, pridobljene iz SICRIS-a v obliki XML (Slika 3), najprej uredili s programom za urejanje besedila Microsoft Word.

V Wordu smo naslov vsake objave postavili v svojo vrstico tako, da smo pred in za naslov vsake objave vrinili oznako odstavka. To smo izvedli z avtomatsko zamenjavo XML oznak začetnih bibliografskih podatkov o objavah </AuthorGroup><Title> z oznako odstavka: ^p<Title> ter z zamenjavo oznake končnih pozicij naslovov </Title> z dodatno oznako odstavka: ^p</Title>. Tako spremenjeno besedilo smo razvrstili po abecednem redu odstavkov, da smo besedilo razvrstili po različnih XML oznakah. Na ta način smo dobili skupaj zbrane vse naslove objav. Ohranili smo le naslove objav in brisali ostalo besedilo tako, da smo v besedilni datoteki ohranili samo odstavke, ki se začnejo z XML oznako <Title>. Tako smo dobili zbirko bibliografskih podatkov objav laboratorija, ki je vsebovala naslove objav, med katerimi so se pojavili tudi naslovi zbornikov in elektronskih knjig, kjer so raziskovalci laboratorija objavili svoje prispevke. Ker je v teh primerih šlo za sekundarne naslove objav, smo iz datoteke brisali vse naslove zbornikov in elektronskih knjig.

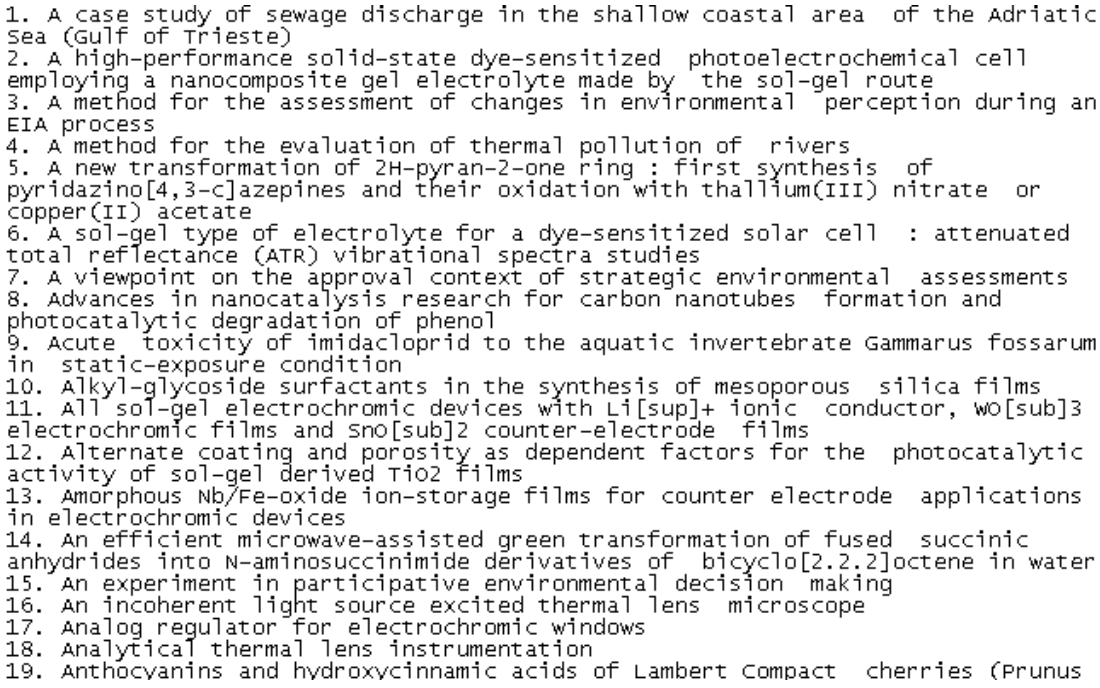
Raziskovalci Laboratorija za raziskave v okolju večino znanstvenih in strokovnih del objavijo v mednarodnih publikacijah, zato je večina njihovih del napisanih v angleškem jeziku. Zaradi podvajanja naslovov v angleškem in slovenskem jeziku smo nazadnje brisali še slovenske prevode angleških objav (kjer so se ti pojavljali) in naslove objav v zgolj slovenskem jeziku, da smo ohranili samo naslove v angleščini. V zadnjem koraku priprave besedilne datoteke smo odstranili tudi XML oznake naslovov: <Title> ter oštevilčili odstavke (naslove člankov). Prečiščeno besedilno

datoteko smo shranili v formatu golo besedilo, ki smo jo nato uporabili za vhodno datoteko pri analizi besedil z računalniškim orodjem OntoGen.

```
  </BiblioEntry>
- <BiblioEntry bno="219" type="article">
  - <AuthorGroup>
    - <Author responsibility="alternative">
        <FirstName>Elias</FirstName>
        <SurName>Stathatos</SurName>
        <Contrib>avtor</Contrib>
      </Author>
    - <Author responsibility="alternative">
        <FirstName>Panagiotis</FirstName>
        <SurName>Lianos</SurName>
        <Contrib>avtor</Contrib>
      </Author>
    - <Author responsibility="alternative">
        <FirstName>Urška</FirstName>
        <SurName>Lavrenčič Štangar</SurName>
        <Code>11873</Code>
        <CodeOrg>1-002.02</CodeOrg>
        <Contrib>avtor</Contrib>
      </Author>
    - <Author responsibility="alternative">
        <FirstName>Boris</FirstName>
        <SurName>Orel</SurName>
        <Code>02565</Code>
        <CodeOrg>1-002.02</CodeOrg>
        <Contrib>avtor</Contrib>
      </Author>
    </AuthorGroup>
  - <Title>
      A high-performance solid-state dye-sensitized photoelectrochemical cell employing a nanocomposite gel electrolyte made by the sol-gel route
    </Title>
  - <TitleISBD>
      A high-performance solid-state dye-sensitized photoelectrochemical cell employing a nanocomposite gel electrolyte made by the sol-gel route / E. Stathatos ... [et al.]
    </TitleISBD>
  - <BiblioSet relation="journal">
```

Slika 3: Podatki, pridobljeni iz sistema SICRIS v neobdelani obliki

Slika 3 prikazuje neobdelane podatke v XML različici preden smo jih obdelali s postopki čiščenja besedila. Ker bi bili podatki v taki obliki neuporabni za delo s programom OntoGen, smo jih morali urediti v tekstovno obliko brez oblikovanja in XML oznak, ki jo prikazuje slika 4.



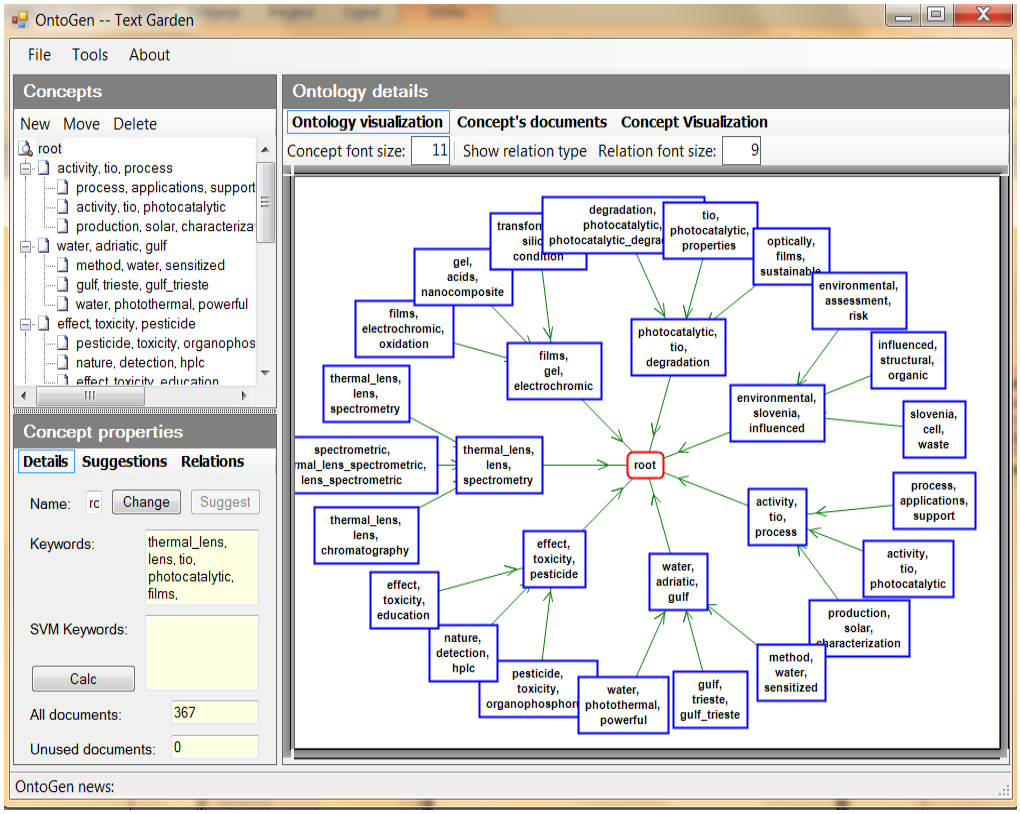
Slika 4: Urejeni naslovi člankov

5.2 Rudarjenje besedil s programom OntoGen

OntoGen je polavtomatsko voden program, ki omogoča analizo besedil s pomočjo jezikovnih tehnologij in prikaz besedil v obliki terminoloških ontologij. Program OntoGen omogoča rudarjenje besedil s pomočjo učinkovitega uporabniškega vmesnika in različnih tehnik odkrivanja zakonitosti v besedilih. V OntoGenu so besedila iz vhodne datoteke predstavljena kot vreče besed in opisana z vektorji, pri čemer vsak element vektorja ustreza pogostosti pojavitve posamezne besede v besedilu. OntoGen je hkrati interaktivno orodje, ki pomaga uporabniku v procesu gradnje ontologij. Program sam predlaga ključne besede, samodejno dodeli primere besedil, ki vsebinsko ustrezajo predlaganim ključnim besedam, omogoča prikaz primerov besedil združenih v t.i. koncepte in zagotavlja hierarhičen pregled nad koncepti in podkoncepti ontologije. Uporabnik lahko deloma ali v celoti sprejme ali zavrne predlagane rezultate analize besedil, ki mu jih sistem avtomatsko ponudi.

Slika 5 prikazuje uporabniški vmesnik OntoGena, ki se prikaže ob zagonu programa in je izhodišče za nadaljnje delo. Začetno okno je razdeljeno v dve manjši in eno večje okno, od katerih je večje namenjeno prikazu ontologij in upravljanju z vhodnimi besedili, ki so v programu poimenovani dokumenti. V zgornjem levem

oknu se prikaže hierarhično drevo z vsemi koncepti in podkoncepti ontologije. V spodnjem levem oknu lahko uporabnik dodaja, spreminja ali ureja koncepte in podkoncepte pri nadaljnji analizi besedil.



Slika 5: Uporabniški vmesnik programa OntoGen verzija 2.0.0.0

Funkcionalnost sistema, ki zajema učenje algoritmov, analizo in upravljanje besedil, se naslanja na zbirko programov imenovano Text Garden library (Fortuna in sod., 2007). Uporabniški vmesnik je implementiran v programskem jeziku C# in zahteva namestitev ogrodja Microsoft NET framework 2.0 za delovanje programa.

6 REZULTATI IN RAZPRAVA

V študiji primera objav Laboratorija za raziskave v okolju smo iz objav od leta 1983 do konca leta 2012, po postopku opisanem v poglavju 5, pridobili 366 naslovov člankov, ki skupno vsebujejo 5126 besed. Na seznam blokiranih besed (t.i. stop listo) smo poleg običajnih praznih besed, ki pri računalniški statistični analizi besedil ne nosijo posebnega pomena (npr. vezniki, števniki, zaimki), zaradi pogoste pojavitve v naslovih člankov, uvrstili naslednje angleške izraze: *study* (študija), *studies* (študije) in *determination* (določitev), ki nimajo posebne vsebinske teže. S programom OntoGen smo izvedli več poskusov gradnje ontologij na osnovi terminološkega razvrščanja naslovov člankov v skupine.

Pri gradnji ontologij s programom OntoGen smo uporabili tehniko razvrščanja v skupine z algoritmom *k-means*, ki razvrsti vhodne elemente v k skupin na osnovi podobnosti elementov in pri tem iskali najprimernejšo vrednost parametra k za obravnavano študijo primera. Algoritem *k-means* razvrsti enote iz vhodne zbirke podatkov po skupinah tako, da je medsebojna podobnost primerkov znotraj posamezne skupine večja v primerjavi s podobnostjo med primerki iz različnih skupin. Zaradi preprostosti in učinkovitosti je algoritem *k-means* eden od najpogosteje uporabljenih algoritmov razvrščanja v praksi (Duda in sod., 2000).

6.1 Rezultati analize naslovov člankov

Pri gradnji ontologij s programom OntoGen smo parameter k (število skupin tekstov) spreminjali od nizke vrednosti (k=3) do razmeroma visoke (k=7) glede na vhodno velikost datoteke, ki je v našem primeru vsebovala 366 naslovov člankov. Med tema dvema vrednostma smo nato iskali vmesne nastavitve, s katerimi bi nam program prikazal vsebinsko čim bolj homogene in hkrati dovolj reprezentativne skupine naslovov člankov, ki jih program poimenuje s koncepti oz. ključnimi besedami.

V prvem poskusu smo z OntoGenom zgradili ontologijo s tremi avtomatsko predlaganimi koncepti (vrednost parametra k=3) na prvem nivoju razvrščanja naslovov člankov. Do izraza so prišle tri večje skupine naslovov člankov, ki so bile vsebinsko zelo nehomogene. Prva je vsebovala 154 naslovov člankov, katerih vsebina je bila opisana z naslednjimi ključnimi besedami: thermal_lens (toplotne

leče), lens (leče), environmental (okoljski), slovenia (Slovenija), spectrometry (spektrometrija), adriatic (jadranski), gulf (zaliv), trieste (Trst), gulf_trieste (Tržaški zaliv), assessment (ocenjevanje). V drugo skupino je bilo razvrščenih 91 naslovov člankov, opisanih z naslednjimi ključnimi besedami: films (premazi) (premazi), gel, TiO[1] (titanov oksid), sol (delci v tekoči fazi), sol_gel (sol-gel metoda), electrochromic (elektrokromatski), effect (učinek), properties (lastnosti), optically (optično), coating (prevleka). V tretjo skupino je program razvrstil 122 naslovov člankov in skupino opisal s ključnimi besedami: degradation (razgradnja), photocatalytic (fotokatalitski), activity (dejavnost), influenced (vplival), pesticide (pesticid), method (metoda), water (voda), TiO (titanov oksid), compounds (spojine), evaluation (vrednotenje).

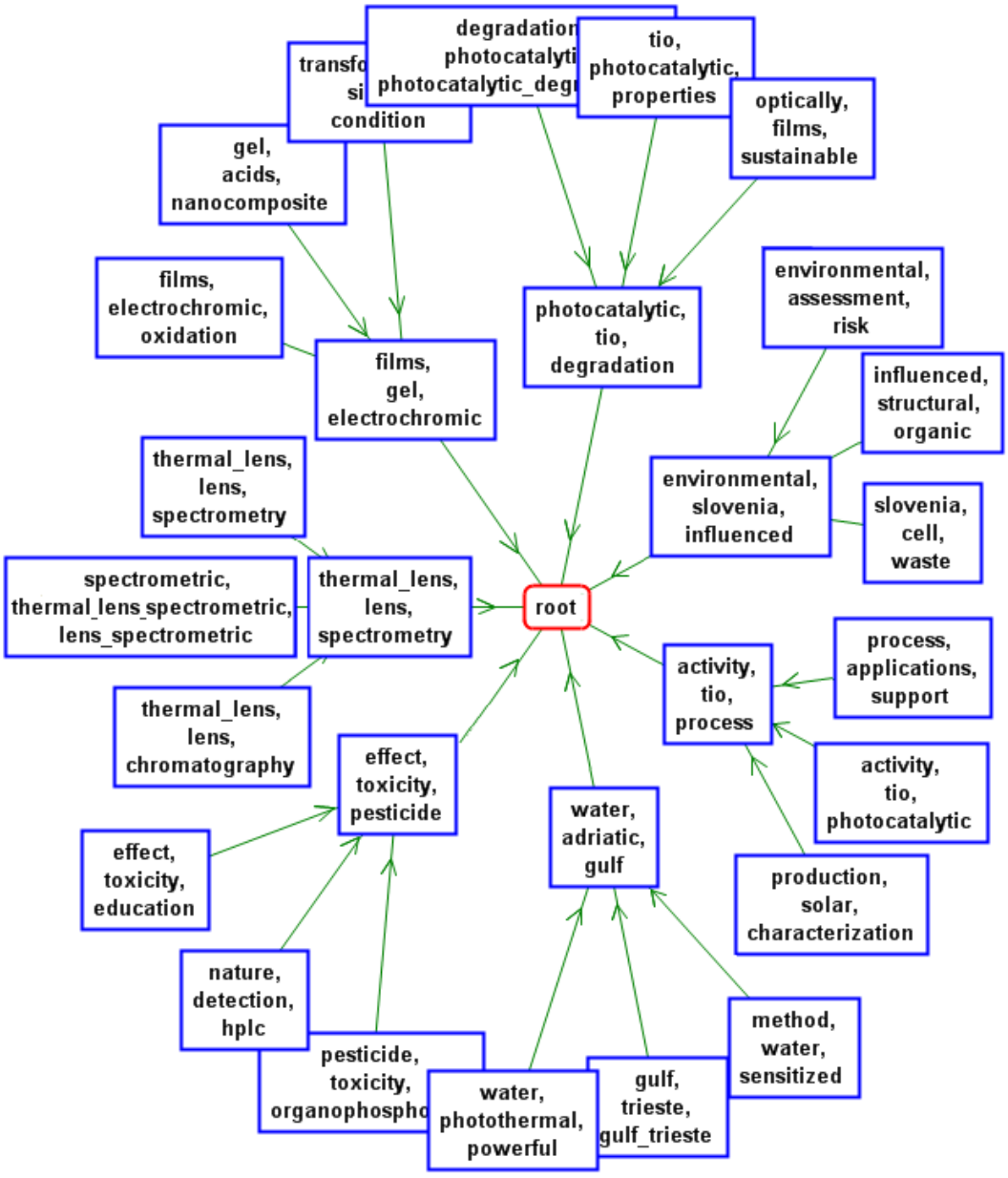
V drugem poskusu smo povečali parameter *k* na 7 in vsakemu od 7-ih dobljenih konceptov dodali po 3 podkoncepte, ki so bili avtomatsko predlagani. V prvi koncept je bilo razvrščenih 52 naslovov člankov, ki jim je program dodelil naslednje ključne besede: thermal_lens (toplotne leče), lens (leče), spectrometry (spektometrija), thermal_lens_spectrometry (spektometrija s toplotnimi lečami), lens_spectrometry (spektometrija z lečami), chromatography (kromatografija), thermal (toplotna), liquid (tekočina), detection (odkrivanje), spectrometric (spektrofotometrijska). V drugi koncept je bilo razvrščenih 55 naslovov člankov, ki so bili opisani s ključnimi besedami: films (premazi), gel, electrochromic (elektrokromatski), oxidation (oksidacija), acids (kisline), synthesis (sinteza), sol (delci v tekoči fazi), sol_gel (sol-gel metoda), transformation (sprememba), benzopyran (benzopiran). V tretjem konceptu je bilo 44 naslovov člankov s ključnimi besedami: photocatalytic (fotokatalitska), TiO (titanov oksid), degradation (razgradnja), properties (lastnosti), sustainable (trajnostno), photocatalytic_degradation (fotokatalitska razgradnja), optically (optično), coating (prevleka), films (premazi), development (razvoj). V četrtem konceptu je bilo 63 naslovov člankov s ključnimi besedami: environmental (okoljski), slovenia (Slovenija), influenced (vplival), assessment (ocenjevanje), organic (organski), structural (strukturni), cell (celica), risk (tveganje), process (proces), characteristics (značilnosti). V petem konceptu je bilo 41 naslovov člankov s ključnimi besedami: activity (dejavnost), TiO (titanov oksid), process (proces),

[1] V naslovih člankov se pojavlja izraz TiO_2 (titanov dioksid), OntoGen pa prikaže izraz brez števila 2

production (proizvodnja), applications (nanosi), photocatalytic (fotokatalitski), analysis (analiza), sol (delci v tekoči fazi), sol_gel (sol-gel metoda), gel. V šestem konceptu je 49 naslovov člankov in ključne besede: water (voda), adriatic (jadranski), gulf (zaliv), trieste (Trst), gulf_trieste (Tržaški zaliv), sea (morje), method (metoda), photothermal (fototermični), adriatic_sea (Jadransko morje), marine (morski). V zadnjem sedmem konceptu je bilo 63 naslovov člankov s ključnimi besedami: effect (učinek), toxicity (strupenost), pesticide (pesticid), detection (odkrivanje), nature (narava), photothermal (fototermični), imidacloprid (imidakloprid), hplc (High-performance liquid chromatography – HPLC - tekočinska kromatografija visoke ločljivosti), teachers (učitelji), education (izobraževanje).

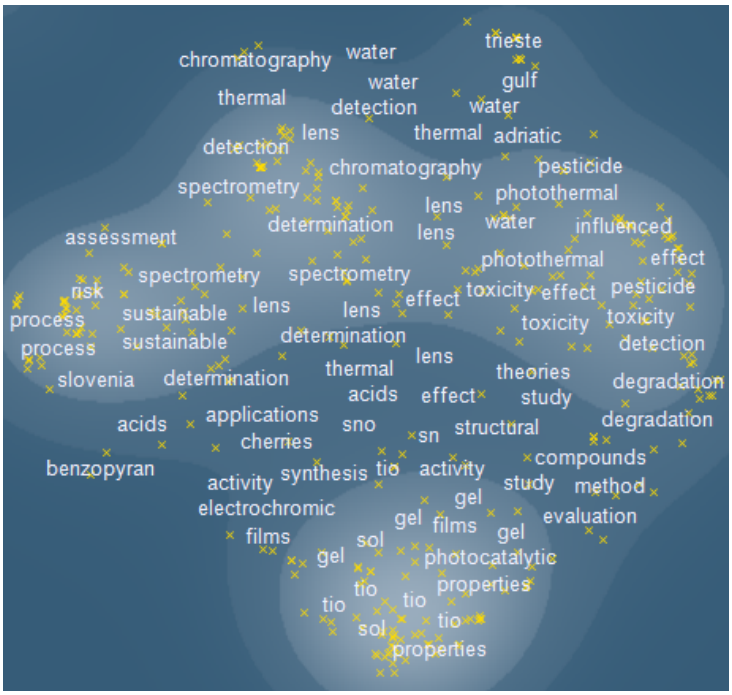
Iz dobljenih rezultatov je razvidno, da največja skupina objav zajema področje strupenosti pesticidov. Vsebinsko zelo homogena je skupina, v kateri se nahajajo naslovi objav o vodi, Jadranskem morju in Tržaškem zalivu. Pri tem poskusu je tako že vidna glavna dejavnost laboratorija za raziskave v okolju, vendar se ključne besede v nekaterih skupinah ponavljajo, kar kaže na preveliko drobljenje vhodnih besedil po skupinah oz. previsok parameter *k*.

V tretjem koraku smo poskusili, kako bi lahko izboljšali vsebinsko razdelitev naslovov člankov v skupine oz. koncepte in podkoncepte, zato smo preizkusili še vrednosti parametra *k* med 3 in 7. Predvsem smo želeli dobiti čim bolj nazoren prikaz vsebin, s katerimi se pretežno ukvarja Laboratorij za raziskave v okolju in koliko je teh objav po posameznih vsebinsko najpomembnejših področjih.



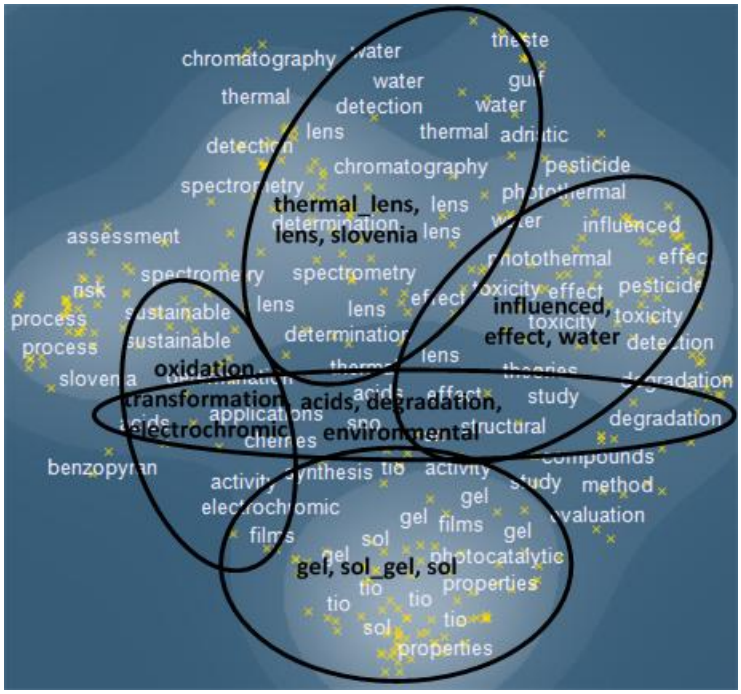
Slika 6: Ontologija s sedmimi koncepti in tremi podkoncepti

Kot najboljša se je izkazala razdelitev vhodnih besedil v 5 skupin (konceptov) na prvem hierarhičnem nivoju ontologije s 3 podskupinami (podkoncepti) na drugem nivoju. To je razvidno tudi iz terminološkega zemljevida, dobljenega s funkcijo *Concept Visualization* v OntoGenu. Na zemljevidu (slika 7) lahko opazimo 4 izrazitejše skupine ključnih besed iz naslovov člankov Laboratorija za raziskave v okolju (4 belo obarvana področja) in skupino ključnih besed v temnejšem, osrednjem delu zemljevida, ki se pojavljajo izven 4 vsebinsko homogenejših skupin naslovov.



Slika 7: Zemljevid prvega poskusa

Slika 9 prikazuje ontologijo vsebinskih področij s petimi koncepti, od katerih je vsak razdeljen še v tri podkoncepte. Relacije med izrazitejšimi skupinami ključnih besed (slika 7) in koncepti ontologije (slika 9) so označene na zemljevidu na sliki 8.



Slika 8: Zemljevid prvega poskusa z oznakami konceptov ontologije

Slika 9: Ontologija s petimi koncepti in tremi podkoncepti

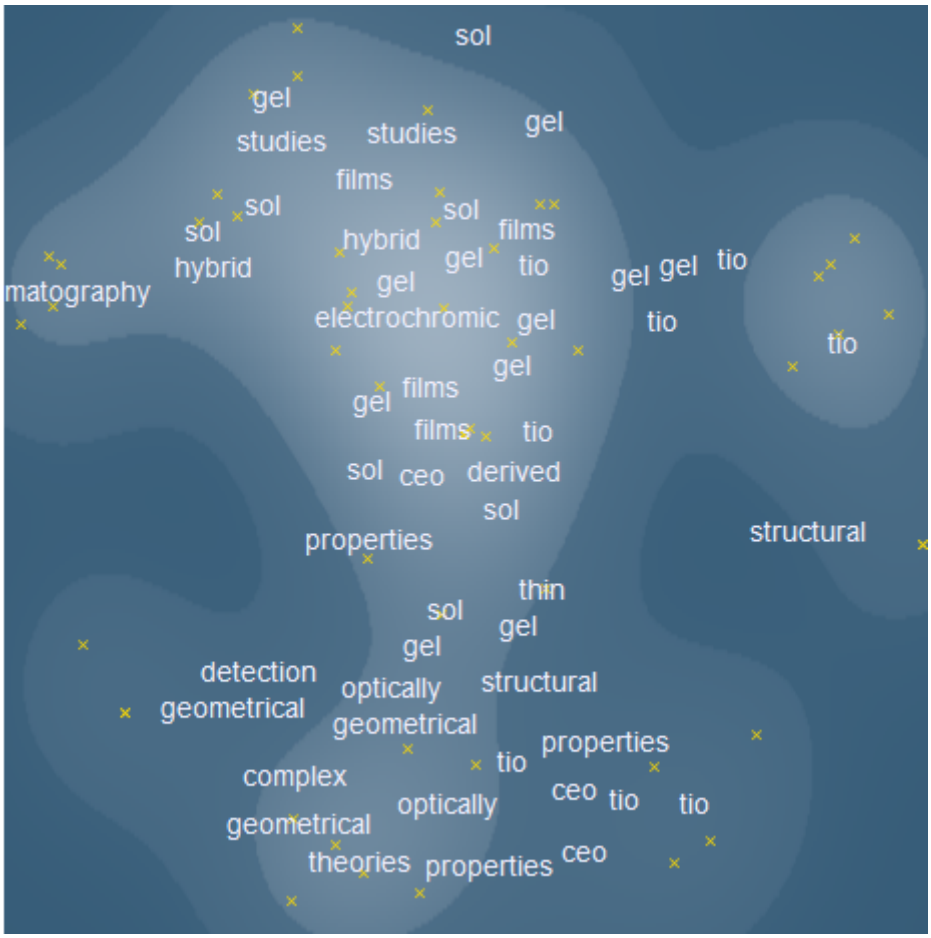
V prvem konceptu (slika 10) je razvrščenih 56 naslovov člankov, ki so vsebinsko opisani s ključnimi besedami: gel, sol_gel (sol-gel metoda), sol (delci v tekoči fazi), TiO (titanov oksid), optically (optično), films (premazi), properties (lastnosti), coating (prevleka), electrochemical (elektrokromni), sn (kositer). Prvi koncept smo razdelili še v tri podkoncepte:

- Prvi podkoncept, ki smo ga označili s tremi najpogostejšimi ključnimi besedami: gel, (sol-gel metoda), sol (delci v tekoči fazi), vključuje 21 naslovov člankov, ki jim je program OntoGen dodelil ključne besede: gel, sol_gel (sol-gel metoda), sol (delci v tekoči fazi), TiO (titanov oksid), films

(premazi), hybrid, properties (lastnosti), electrochemical (elektrokemični), thin (tanek), activity (dejavnost).

- Drugi podkoncept, ki smo ga označili s tremi najpogostejšimi ključnimi besedami: optically (optičen), coating (prevleka), properties (lastnosti), vključuje 22 naslovov člankov, ki jim je program OntoGen dodelil ključne besede: optically (optično), coating (prevleka), properties (lastnosti), sno (kositrov oksid), trieste (Trst), gulf_trieste (Tržaški zaliv), gulf (zaliv), sb (antimon), electrochemical (elektrokemični), detection (odkrivanje).

- Tretji podkoncept, ki smo ga označili s tremi najpogostejšimi ključnimi besedami: TiO (titanov oksid), films (premazi), electrochromic (elektrkromatski), vključuje 13 naslovov člankov, ki jim je program OntoGen dodelil ključne besede: TiO (titanov oksid), films (premazi), electrochromic (elektrokromatski), ion (ion), devices (naprave), synthesis (sinteza), applications (nanosi), optically (optično), chromatography (kromatografija), photocatalysts (fotokataliza).



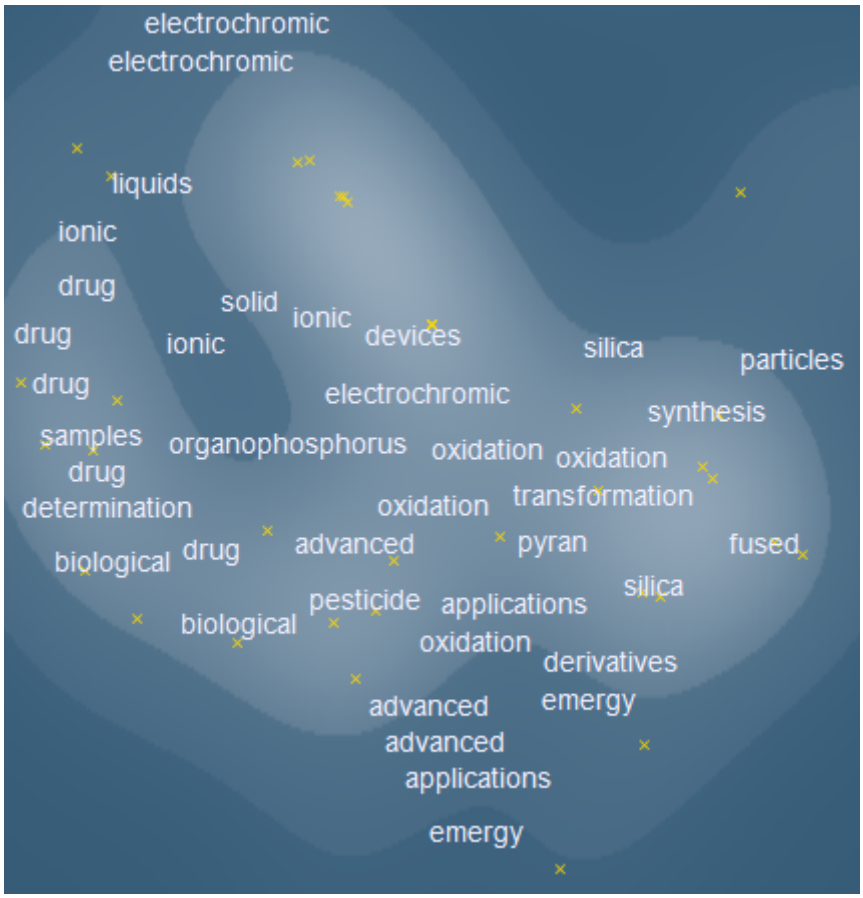
Slika 10: Zemljevid prvega koncepta

V drugem konceptu (slika 11) je 47 naslovov člankov, ki so vsebinsko opisani s ključnimi besedami: oxidation (oksidacija), transformation (sprememba), electrochromic (elektrokromatski), films (premazi), silica (silicijev dioksid), sustainable (trajnostno), solid (trden), synthesis (sinteza), drugs (zdravila), sampling (vzorčenje). Drugi koncept smo razdelili v tri vsebinske podkoncepte:

- Prvi podkoncept, ki smo ga označili s tremi najpogostejšimi ključnimi besedami: education (izobraževanje), applications (nanosi), methodology (metodologija), vključuje 9 naslovov člankov, ki jim je program OntoGen dodelil ključne besede: education (izobraževanje), applications (nanosi), methodology (metodologija), nanocomposite (nanosestavi), sustainable (trajnostni), nature (narava), oxidation (oksidacija), complex (zapleten), biological (biološki), diazinon (Diazinon).

- Drugi podkoncept, ki smo ga označili s tremi najpogostejšimi ključnimi besedami: oxidation (oksidacija), electrochromic (elektrokromatski), films (premazi), vključuje 24 naslovov člankov, ki jim je program OntoGen dodelil ključne besede: oxidation (oksidacija), electrochromic (elektrokromatski), films (premazi), solid (trden), silica (silicijev dioksid), liquid (tekočina), synthesis (sinteza), thin (tanek), acids (kisline), degradation (razpad).

- Tretji podkoncept, ki smo ga označili s tremi najpogostejšimi ključnimi besedami: transformation (sprememba), drugs (zdravila), sustainable (trajnostno), vključuje 14 naslovov člankov, ki jim je program OntoGen dodelil ključne besede: transformation (sprememba), drugs (zdravila), sustainable (trajnostno), condition (pogoj), silica (silicijev dioksid), sampling (vzorčenje), hplc (High-performance liquid chromatography – HPLC - tekočinska kromatografija visoke ločljivosti), formation (oblikovanje), water (voda), derived (izpeljan).



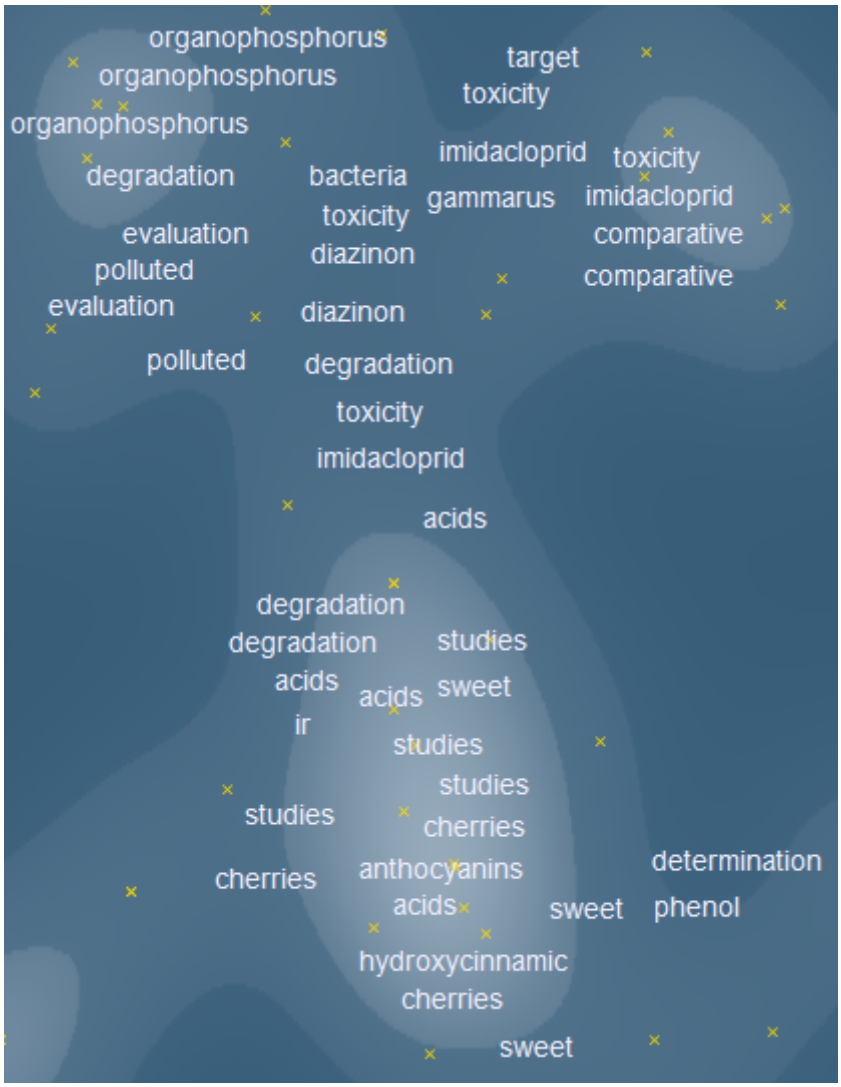
Slika 11: Zemljevid drugega koncepta

V tretjem konceptu (slika 12) je 44 naslovov člankov, ki so vsebinsko opisani s ključnimi besedami: acids (kisline), degradation (razpad), environmental (okoljski), toxicity (strupenost), imidacloprid (sistemski insekticid), organic (organski), cherries (češnje), aquatic (vodni), bacteria (bakterija), sweet (sladek). Tretji koncept smo razdelili v tri podkoncepte:

- Prvi podkoncept, ki smo ga označili s tremi najpogostejšimi ključnimi besedami: acids (kisline), imidacloprid (sistemski insekticid), cherries (češnje), vključuje 23 naslovov člankov, ki jim je program OntoGen dodelil ključne besede: acids (kisline), imidacloprid (sistemski insekticid), cherries (češnje), toxicity (strupenost), organic (organski), sweet (sledek), sweet_cherries (sladke češnje), spectroscopy (spektroskopija), comparative (primerjalni), photoacoustic (fotoakustičen).

- Drugi podkoncept, ki smo ga označili s tremi najpogostejšimi ključnimi besedami: degradation (razpad), bacteria (bakterija), aquatic (vodni), vključuje 13 naslovov člankov, ki jim je program OntoGen dodelil ključne besede: degradation (razpad), bacteria (bakterija), aquatic (vodni), evaluation (vrednotenje), toxicity (strupenost), laser (laser), compounds (sestavine), organophosphorus (organofosforni), high (visok), adriatic (jadransko).

- Tretji podkoncept, ki smo ga označili s tremi najpogostejšimi ključnimi besedami: environmental (okoljski), microbial (mikrobni), development (razvoj), vključuje 8 naslovov člankov, ki jim je program OntoGen dodelil ključne besede: environmental (okoljski), microbial (mikrobni), development (razvoj), sustainable (trajnostni), experiment (poskus), risk (tveganje).



Slika 12: Zemljevid tretjega koncepta

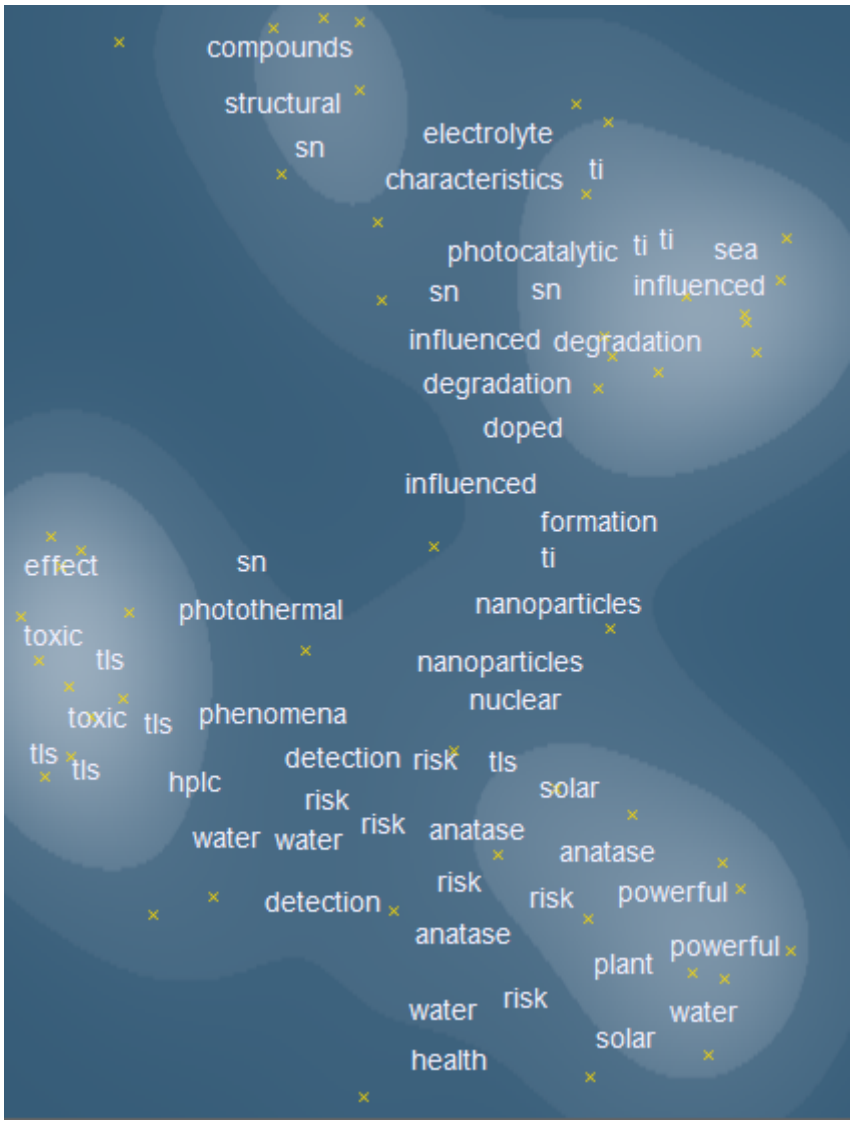
V četrtem konceptu (slika 13) je 52 naslovov člankov, ki so vsebinsko opisani s ključnimi besedami: influenced (vplival), effect (učinek), water (voda), structural (strukturni), powerful (zmogljiv), detection (odkrivanje), sn (kositer), photothermal (fototermičen), tls (Thermal_Lens_Spectrometry – TLS - spektrometrija na osnovi termičnih leč), characteristics (značilnosti). Četrti koncept smo razdelili v tri podkoncepte:

- Prvi podkoncept, ki smo ga označili s tremi najpogostejšimi ključnimi besedami: effect (učinek), powerful (zmogljiv), risk (tveganje), vključuje 15 naslovov člankov, ki jim je program OntoGen dodelil ključne besede: effect (učinek), powerful (zmogljiv), risk (tveganje), environmental (okoljski), plant

(rastlina), toxicity (strupenost), photothermal (fototermičen), investigations (preiskave), compounds (spojine), deflective (sposoben spremeniti smer).

- Drugi podkoncept, ki smo ga označili s tremi najpogostejšimi ključnimi besedami: influenced (vplival), structural (strukturni), sn (kositer), vključuje 21 naslovov člankov, ki jim je program OntoGen dodelil ključne besede: influenced (vplival), structural (strukturni), sn (kositer), characteristics (značilnosti), organic (organski), electrolyte (elektrolit), ti (titan), solar (sončni), compounds (spojine), photocatalytic (fotokatalitski).

- Tretji podkoncept, ki smo ga označili s tremi najpogostejšimi ključnimi besedami: water (voda), detection (odkrivanje), photothermal (fototermični), vključuje 16 naslovov člankov, ki jim je program OntoGen dodelil ključne besede: water (voda), detection (odkrivanje), photothermal (fototermični), tls (Thermal_Lens_Spectrometry – TLS - spektrometrija na osnovi termičnih leč), hplc (High-performance liquid chromatography – HPLC - tekočinska kromatografija visoke ločljivosti), effect (učinek), pesticide (pesticid), selected (izbran), anatase (anatas - eden od treh mineralnih oblik titanovega oksida), powerful (zmogljiv).



Slika 13: Zemljevid četrtega koncepta

V petem konceptu (slika 14) je 105 naslovov člankov, ki so vsebinsko opisani s ključnimi besedami: thermal_lens (termične leče), lens (leče), slovenia (Slovenija), spectrometry (spektrometrija), adriatic (jadranski), assessment (ocena), lens_spectrometry (spektrometrija na osnovi leč), thermal_lens_spectrometry (spektometrija na osnovi termičnih leč), process (postopek), measurements (meritve). Peti koncept smo razdelili v tri podkoncepte:

- Prvi podkoncept, ki smo ga označili s tremi najpogostejšimi ključnimi besedami: gulf (zaliv), trieste (Trst), gulf_trieste (Tržaški zaliv), vključuje 28 naslovov člankov, ki jim je program OntoGen dodelil ključne besede: gulf (zaliv), trieste (Trst), gulf_trieste (Tržaški zaliv), thermal_lens (termične

leče), lens (leče), adriatic (jadranski), chromium (krom), slovenia (Slovenija), marine (pristanišče), spectroscopy (spetroskopija).

- Drugi podkoncept, ki smo ga označili s tremi najpogostejšimi ključnimi besedami: spectrometry (spektrometrija), thermal_lens_spectrometry (spektrometrija na osnovi termičnih leč), lens_spectrometry (spektrometrija na osnovi leč), vključuje 34 naslovov člankov, ki jim je program OntoGen dodelil ključne besede: spectrometry (spektrometrija), thermal_lens_spectrometry (spektometrija na osnovi termičnih leč), lens_spectrometry (spektometrija na osnoi leč), thermal_lens (termične leče), lens (leče), adriatic (jadranski), nature (narava), pesticide (pesticid), directive (direktiva), analysis (analiza).

- Tretji podkoncept, ki smo ga označili s tremi najpogostejšimi ključnimi besedami: thermal_lens (termične leče), lens (leče), slovenia (Slovenija), vključuje 43 naslovov člankov, ki jim je program OntoGen dodelil ključne besede: thermal_lens (termične leče), lens (leče), slovenia (Slovenija), measurements (meritve), assessment (ocena), environmental (okoljski), detection (odkrivanje), process (postopek), chromatography (kromatografija), sensitized (občutljiv).

Slika 14: Zemljevid petega koncepta

6.2 Razprava

V prvi koncept ontologije, ki ga označujejo ključne besede gel, sol_gel, sol je vključenih 56 naslovov člankov, ki opisujejo strukturo ter optične in elektrokemične lastnosti premazov pripravljenih po sol-gel postopku in elektrokemičnih naprav izdelanih po sol-gel postopkih. Sol-gel postopek se izvaja v raztopinah oz. suspenzijah, kjer iz začetnih reaktantov med kemijsko reakcijo nastanejo zelo majhni koloidni delci. Tako so na primer raziskovalci Laboratorija za okolje po sol-gel postopku pripravili tanke keramične optičnoselektivne prevleke $CoAl_2O_4$ za uporabo na halogenskih avtomobilskih žarnicah (Lavrenčič Štangar in sod., 2002). V isto skupino naslovov so uvrščeni tudi naslovi objav, ki obravnavajo pripravo tankih

TiO_2 premazov, kot na primer članek (Armelao in sod., 2009), ki obravnava način priprave Ag/TiO_2 nanosistemov na osnovi radiofrekvenčne razpršitve delcev srebra na tanke premaze pripravljene po sol-gel metodi. V tej skupini naslovov so opisane tudi posebne lastnosti tankih Ag/TiO_2 premazov.

V drugem konceptu ontologije je uvrščenih 47 naslovov člankov, ki vključujejo elektrokromatske in strukturne študije tankih premazov. V tej skupini najdemo tudi naslove člankov, ki opisujejo potek oksidacijskih eksperimentov, na primer z uporabo plinske kromatografije na masno detekcijo (Bavcon Kralj in sod., 2006), kjer so člani Laboratorija za raziskave v okolju proučevali učinkovitost kemijske oksidacije žveplovih organofosfatnih pesticidov v vodnih raztopinah in vzorcih sadnega soka. Pri tem so ugotavljali nastale kisikove produkte s pomočjo imobilizirane acetilholinesterazne in spektometrije na osnovi termičnih AChE-TLS leč. AChE-TLS bioanalitske tehnike omogočajo enostavnejšo in temeljitejšo kontrolo in odkrivanje vzorcev, ki so onesnaženi s holinesteraznimi inhibitorji, kot na primer žveplovimi analogi.

V tretji koncept ontologije je razvrščenih 44 naslovov člankov predvsem s področja biokemije pesticidov in drugih zdravju škodljivih snovi, tako na primer raziskava akutne strupenosti imidakloprida, ki je aktivna snov mnogih insekticidnih pripravkov s širokim spektrom delovanja tudi na vodne nevretenčarje (Malev in sod., 2011). Vodne ekosisteme so obravnavali tudi v študiji odmrlih meduz (Tinta in sod., 2010), ki jih bakterije zelo hitro razgradijo, pri čemer nastajajo velike količine amonija, posledično nastane pomanjkanje kisika v vodi, kar lahko povzroča spremembe v delovanju ekosistema. V isto skupino naslovov člankov so uvrščene tudi raziskave o vsebnosti antocianov in hidroksicimetnih kislin v različnih lokalno razširjenih sortah češenj na Goriškem (Mozetič in Trebše, 2004). Različnim kultivarjem češenj je bila s kromatografsko metodo HPLC-DAD določena vsebnost glavnih predstavnikov fenolov in njihove antioksidativne lastnosti. Pri tem so pokazali, da se različni kultivarji ne razlikujejo po vrsti posameznih fenolov, ampak le po njihovi količini v sadju.

V četrtem konceptu ontologije je 52 naslovov člankov, med katerimi so študije ekoloških značilnosti voda, predvsem obalnega morja. Ena od študij (Mozetič in sod., 1999) na primer obravnava vpliv odplak čistilne naprave, ki se pred Piranom

stekajo vzdolž dveh podvodnih izpustov v obalno morje. Oceanografske meritve, analize hranilnih snovi in fekalnih koliformnih bakterij so pokazale širjenje odplak v morski vodi, nekaj metrov nad dnom morja. Podobne molekularne metode so uporabljene tudi v raziskavah posledic vnosa hranilnih snovi, na primer v obliki komunalnih odplak, na sestavo planktonske združbe (Mozetič in sod., 1998). Raziskovalci laboratorija so za analize voda uporabili tudi tehniko spektrometrije na osnovi termičnih leč (TLS), s katero so ugotavljali srebrove ione v vodi (Korte in sod., 2011). Navedena tehnika je bila uporabljena v več raziskavah, katerih naslovi so vključeni v četrti koncept ontologije. Še več tovrstnih raziskav je vključenih v peti koncept ontologije, kar je razvidno tudi iz ključnih besed petega koncepta.

Peti koncept ontologije je številčno najbolj obsežen, saj vključuje 105 naslovov člankov. Članki, ki so vsebinsko najbolj značilni za peti koncept ontologije, obravnavajo metode kemijske analize na osnovi termičnih leč in določanje organskih spojin s spektrometrijo na osnovi termičnih leč. Velik del člankov opisuje razvoj visoko občutljivih laserskih metod analize na osnovi termičnih leč in njihovo kombinacijo z drugimi analitskimi metodami. Spektrometrija TLS omogoča zelo natančno določanje različnih svetlobno občutljivih spojin, ki so lahko prisotne tudi v zelo nizkih koncentracijah. Tako na primer Budal in Franko (2009) opisujeta določanje biogenih aminov na osnovi termičnih leč. Tehnika TLS je bila uporabljena tudi za visoko občutljivo odkrivanje bilirubina v študiji transporta antioksidantov prek celične membrane in vloge transportnih proteinov pri tem procesu (Margon in sod., 2005).

Posebno področje raziskav, ki ga je računalniško orodje OntoGen uvrstilo v isti koncept ontologije, kamor so uvrščeni naslovi objav s področja spektrometrije na osnovi termičnih leč, pa se nanaša na raziskave Jadranskega morja v Tržaškem zalivu in okoljske raziskave na drugih območjih v Sloveniji. Tako na primer Malej in sodelavci objavijo vrsto raziskav, med njimi tudi študijo (1997), v kateri pokažejo, da lokalni pritoki meteornih voda, ki se izlivajo v morje, pomembno spodbujajo nastanek fitoplanktona v priobalnih območjih Jadranskega morja. Desetletje kasneje predstavijo problematiko onesnaževanja Jadranskega morja v Tržaškem zalivu z odpadnimi vodami iz gospodinjstev in industrijskih odplak (Mozetič in sod., 2008). Pri nekaterih od raziskav fitoplanktona je za namene analize uporabljen dvojni

dvožarkovni spektrometer s termičnimi lečami (na primer v Kožar Logar in sod, 2006). To nam pojasni smiselnost uvrstitve objav s področja ekologije morja v peti koncept ontologije, ki obravnava predvsem področje raziskav s spektrometrijo na osnovi termičnih leč.

Med rezultati smo pričakovali tudi imena naprav in instrumentov, ki so jih člani laboratorija uporabljali pri raziskovanju, vendar se naprave ne pojavijo med ključnimi besedami, dobljenimi iz naslovov člankov. Poleg tega med dobljenimi ključnimi besedami ne zasledimo omemb rek, npr. Soče ali Vipave, čeprav bi pričakovali, da laboratorij objavlja tudi rezultate analiz rek in drugih vodotokov.

Rezultati besedilne analize so bili predstavljeni eni od raziskovalk Laboratorija za raziskave v okolju. Po njenem mnenju so rezultati reprezentativna slika vsebin, ki so bile predmet raziskav laboratorija. Opozorila pa je na tri ključne besede: education, drugs in cherries, ki so se ji zdele manj tipične za koncepte, v katerih so bile predlagane. Razložila je, da se navedene ključne besede nanašajo na raziskave o izobraževanju s trajnostnimi vidiki razvoja, na raziskave strupenosti statinov – zdravil za zniževanje ravni holesterola v krvi ter na raziskave o rastlinskih fenolih v slovenskih kultivarjih sadja, ki so se v okviru laboratorija izvajale le občasno ali le krajši čas.

OntoGen se je pri besedilni analizi naslovov člankov raziskovalcev Laboratorija za raziskave v okolju izkazal kot uporabno orodje za modeliranje domenskega znanja. Pri podrobnem pregledu ključnih besed, ki so bile s pomočjo orodja samodejno identificirane za posamezne skupine naslovov glede na obravnavano vsebino, smo ugotovili, da bi lahko prikaz ključnih besed izboljšali z ročnim dopolnjevanjem ali preimenovanjem. Eden najbolj očitnih primerov v študiji Laboratorija za raziskave v okolju je ključna beseda *TiO*, ki se sicer v naslovih člankov pojavlja kot TiO_2 (titanov dioksid). Kot neustrezna se je po podrobnem pregledu analiziranih naslovov izkazala tudi ključna beseda *education* (izobraževanje), ki se v celotnem naboru vhodnih naslovov člankov pojavi le v petih naslovih, OntoGen jo je predlagal kot prvo ključno besedo enega od podkonceptov terminološke ontologije s petimi koncepti.

Sicer smo z gradnjo terminoloških ontologij na osnovi naslovov člankov raziskovalcev Laboratorija za raziskave v okolju pridobili celovit pregled nad obravnavano domeno. S študijo primera raziskav v okolju smo pokazali, da imajo terminološke ontologije veliko izrazno moč pri predstavitvi besedišča z obravnavanega strokovnega področja. Za še podrobnejši prikaz domenskega znanja bi bilo smiselno zajeti tudi povzetke člankov iz obravnavanega obdobja in tako dobiti še podrobnejše prikaze tematik objav Laboratorija za raziskave v okolju.

7 ZAKLJUČEK

Naraščajoče količine besedilnih in drugih formatov zapisa podatkov, ki so shranjeni v sodobnih bazah podatkov, so dragocen vir informacij in znanja. Pridobivanje informacij iz ogromnih baz raznolikih podatkov je zapleteno, še posebej, ko želimo raziskovati področja, na katerih nismo strokovno izobraženi. V pomoč pri tem so nam sodobna računalniška orodja, ki lahko iz množice podatkov izluščijo uporabno znanje in ga prikažejo v obliki smiselnih vzorcev na uporabniku prijazen način. Eno takih orodij je OntoGen, ki smo ga v magistrskem delu uporabili za besedilno analizo ter prikaz terminoloških zemljevidov in ontologij iz naslovov člankov Laboratorija za raziskave v okolju Univerze v Novi Gorici. V ta namen smo za obdobje od leta 1983 do vključno leta 2012 iz informacijskega sistema o raziskovalni dejavnosti v Sloveniji - SICRIS pridobili 366 naslovov člankov. S pomočjo računalniškega orodja OntoGen smo naslove člankov vsebinsko razvrščali v skupine in iz skupin gradili terminološke ontologije. Pri tem smo za razvrščanje naslovov člankov v skupine uporabili algoritem *k-means*, ki razvrsti vhodne elemente v k skupin na osnovi njihove vsebinske podobnosti. Pri tem smo iskali najprimernejšo vrednost parametra k za obravnavano študijo primera. Da bi z besedilno analizo dobili vsebinsko smiselno razdelitev dokumentov, smo poskušali z različnimi vrednostmi parametra k, od $k=3$ do $k=7$. Če bi imeli večjo množico vhodnih besedil (tj. naslovov člankov) npr. 1000 ali več, bi lahko bil parameter k še višji od $k=7$. Pri 366 naslovih pa so se ob visokih vrednostih parametra k začele nekatere ključne besede pojavljati v dveh ali več skupinah naslovov člankov, kar je pomenilo, da smo vhodno množico besedil preveč drobili. Tako smo z magistrskim delom metodološko pokazali, da je vrednost parametra k in s tem število konceptov v ontološkem prikazu domenskega znanja, pogojena z velikostjo množice vhodnih besedil.

Analiza naslovov objav raziskovalcev Laboratorija za raziskave v okolju, ki smo jo opravili v magistrskem delu, kaže na vsebinsko raznolikost raziskovalnih področij obravnavanega laboratorija Univerze v Novi Gorici. Pri besedilni analizi in gradnji terminoloških ontologij so do izraza prišle predvsem raziskovalne tematike, ki uporabljajo sol-gel postopke za pripravo tankih, na primer TiO_2 premazov ali prevlek in za izdelavo elektrokemičnih naprav. Nezanemarljiv je tudi delež raziskav s področja biokemije pesticidov in drugih zdravju škodljivih snovi, kot na primer

imidakloprida ter raziskav, ki obravnavajo snovi, ki imajo pozitivne učinke na zdravje. V to skupino raziskav sodijo na primer študije o vsebnosti antocianov in hidroksicimetnih kislin v različnih lokalno razširjenih sortah češenj na Goriškem.

Vsebinsko največji delež raziskav se uvršča v področje kemijske analize na osnovi termičnih leč in določanje organskih spojin s spektrometrijo na osnovi termičnih leč. Podroben pregled te skupine naslovov člankov pokaže izstopanje manjše skupine naslovov, ki obravnavajo raziskave Jadranskega morja v Tržaškem zalivu in okoljske raziskave na drugih območjih v Sloveniji.

Pri gradnji terminoloških ontologij na osnovi naslovov člankov na primeru raziskav v okolju smo ugotovili, da je natančnost modeliranja in prikazov domenskega znanja v veliki meri odvisna od nastavitev parametrov besedilne analize. V našem primeru je bila ključnega pomena vrednost parametra *k*, po katerem računalniško orodje OntoGen z algoritmom *k-means* razvrsti vhodne elemente (naslove člankov v našem primeru) v *k* skupin na osnovi vsebinske podobnosti elementov. Kot najboljša izbira v obravnavani študiji primera se je izkazala razdelitev vhodnih besedil v 5 skupin (konceptov) na prvem hierarhičnem nivoju ontologije s 3 podskupinami (podkoncepti) na drugem nivoju. Pomembna je tudi velikost vhodne datoteke, v našem primeru: število naslovov, ki so bili zajeti v raziskavo. Vhodnih podatkov mora biti količinsko dovolj, da jih lahko analiziramo ter na osnovi analize dobimo smiselne in reprezentativne rezultate.

OntoGen se je v opisanih poskusih izkazal za praktično orodje, s katerim bi bilo smiselno zajeti tudi povzetke člankov iz obravnavanega obdobja, da bi lahko prikazali še podrobnejše rezultate, ki bi odražali tudi morebitne posebne tematike objav Laboratorija za raziskave v okolju.

8 LITERATURA

Armelao, L., Barreca, D., Bottaro, G., Gasparotto, A., Maccato, C., Tondello, E., Lebedev, O.I., Turner, S., Van Tendeloo, G., Sada, C., Lavrenčič Štangar, U. (2009). Rational design of Ag/TiO2 nanosystems by a combined RF-sputtering/sol-gel approach. Chemphyschem, 10(18), str. 3249-3259.

Awad, E. M., Ghaziri, H. M. (2007). Knowledge Management, New Delhi: Dorling Kindersley, str. 60, 307.

Bavcon Kralj, M., Trebše, P., Franko, M. (2006). Oxidation as a pre-step in determination of organophosphorus compounds by the AChE-TLS bioassay. Acta Chimica Slovenica, 53(1), str. 43-51.

Bratina, G. (ur.) (2005). Poročilo 1995-2005. Nova Gorica: Politehnika.

Budal, S., Franko, M. (2009). Določanje biogenih aminov na osnovi termičnih leč = Determination of biogenic amines with thermal lens spectrometry. Slovenski kemijski dnevi 2009, Maribor, 24. in 25. september 2009, 9 str.

Drelichowski, L., Bobek, S., Bojar, W. Chęsy, W., Cilski, B., Czechumski, W., Feoli, E., Fronczak, E., Ganis, P., Graul, c. in sod. (2012). Methodological aspects and case studies of business intelligence applications tools in knowledge management. V: Drelichowski, L. (Ur.), Studies & Proceedings of Polish Association for Knowledge Management, 59. Bydgoszcz: Polish Association for Knowledge Management, 227 str.

Duda, R. O., Hart, P. E., Stork, D. G. (2000) Pattern classification. New York: John Wiley & Sons, str. 526-528.

Fayyad, U., Piatetsky-Shapiro, G., Smyth, P. (1996) Knowledge discovery and data mining: towards a unifying framework. V: Proceeding of the Second International Conference on Knowledge Discovery and Data Mining. Menlo Park: AAAI Press, str. 82-88.

Fortuna, B., Mladenić, D., Grobelnik. M. (2006). Semi-automatic data-driven ontology construction system. V: Bohanec, M. in sod. (ur.), Zbornik 9. mednarodne

multikonference Informacijska družba IS 2006, 9. do 14. oktober 2006, Ljubljana, Slovenia. Ljubljana: Institut Jožef Stefan, str. 223-226.

Fortuna, B., Mladenić, D., Grobelnik. M. (2007). Text garden – skupek orodij za analizo besedil. Novice IJS, 131, str. 17-18.

Ginsparg, P., et al. (2004). Mapping subsets of scholarly information. Proceedings of the National Academy of Sciences of the United States of America, 101(Sup. 1), str. 5236-5240.

Gruninger, M., Lee, L. (2002). Ontology: Applications and design. Communications of the ACM, 45 (2), str. 39-41.

Guarino, N., Giaretta, P. (1995). Ontologies and knowledge bases: Towards a terminological clarification. V: Mars, N.J.I. (ur.), Towards very large knowledge bases: Knowledge building and knowledge sharing. Amsterdam: IOS Press, str. 25-32.

Islovar. Slovar informatike (2013). Ontologija. Pridobljeno 5.4.2013 s svetovnega spleta: http://www.islovar.org

Janssens, F., Glänzel, W., De Moor, B. (2007). Dynamic hybrid clustering of bioinformatics by incorporating text mining and citation analysis. V: Berkhin, P. (Ur.), KDD 2007: Proceedings of the Thirteenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Jose, California, str. 360–369.

Korte, D., Bruzzoniti, M.C., Sarzanini, C., Franko, M. (2011). Influence of foreign ions on determination of ionic Ag in water by formation of nanoparticles in a FIA-TLS system. Analytical Letters 44, str. 2901-2910.

Kožar Logar, J., Malej, A., Franko, M. (2006). Double dual beam thermal lens spectrometer for monitoring of phytoplankton cell lysis. Instrumentation science & technology, 34(1-2), str. 23-31.

Laboratorij za raziskave v okolju (2013). Projekti. Pridobljeno 13.3.2013 s svetovnega spleta: http://www.ung.si/sl/raziskave/laboratorij-za-raziskave-v-okolju/projekti

Lacasta, J., Nogueras-Iso, J., Zarazaga-Soria, F. J. (2010). Terminological Ontologies: Design, Management and Practical Applications. New York: Springer. Pridobljeno 5.4.2013 s svetovnega spleta: http://books.google.com

Lavbič, D., Krisper, M. (2005). Semantika podatkov in ontologije. Uporabna informatika, 13(3), str. 121-135.

Lavrenčič Štangar, U., Orel, B., Krajnc, M., Cerc Korošec, R., Bukovec, P. (2002). Sol-gel-derived thin ceramic CoAl2O4 coatings for optical applications. Materiali in tehnologije 36(6), str. 387-394.

Malej, A., Mozetič, P., Malačič, V., Turk, V. (1997). Response of Summer Phytoplankton to Episodic Meteorological Events (Gulf of Trieste, Adriatic Sea) Marine Ecology, 18(3), str. 273–288.

Malev, O., Fabbretti, E., Sauerborn Klobučar, R., Trebše, P. (2011). Akutna strupenost imidakloprida za vodne nevretenčarje = Acute toxicity of imidacloprid to the aquatic invertebrate Gammarus fossarum in static-exposure condition. V: Kravanja, Z., Brodnjak-Vončina, D., Bogataj, M. (ur.). Slovenski kemijski dnevi 2011, Portorož, 14-16 september 2011. Maribor: Fakulteta za kemijo in kemijsko tehnologijo. 8 str.

Margon, A., Terdoslavich, M., Cocolo, A., Decorti, G., Passamonti, S., Franko, M. (2005). Determination of bilirubin by thermal lens spectrometry and studies of its transport into hepatic cells. Journal de Physique IV France, 125, str. 717-720

Mozetič, B., Trebše, P. (2004). Identification of sweet cherry anthocyanins and hydroxycinnamic acids using HPLC coupled with DAD and MS detector. Acta Chimica Slovenica, 51(1), str. 151-158.

Mozetič, P., Malačič, V., Turk, V. (1999). Ecological characteristics of seawater influenced by sewage outfall. Annales. Series historia naturalis, 2(17), str. 177-190.

Mozetič, P., Malačič, V., Turk, V. (2008). A case study of sewage discharge in the shallow coastal area of the Adriatic Sea (Gulf of Trieste). Marine ecology, 29(4), str. 483-494.

Mozetič, P., Turk, V., Malej, A. (1999). Nutrient-enrichment effect on plankton composition. Annales. Series historia naturalis, 8(13), str. 31-42.

Petrič, I., Urbančič, T., Cestnik, B. (2011). Ontological representation of virtual business communities: how to find right business partners. V: Cruz-Cunha, M. M., Varajao, J. (ur.), Innovations in SMEs and conducting e-business: technologies, trends and solutions. Hershey: Information Science Reference: Igi Global, str. 263-277.

Poli, R., Seibt, J. (ur.) (2010). Theory and Applications of Ontology: Philosophical Perspectives. Dordrecht: Springer. Pridobljeno 5.4.2013 s svetovnega spleta: http://books.google.com

Rector, A. L., Nowlan, W. A. (1994). The GALEN project. Computer methods and programs in biomedicine, 45(1-2), str. 75-78.

Sebastiani, F. (2002). Machine learning in automated text categorization. ACM Computing Surveys 34(1), str. 1-47.

SICRIS (2013). Bibliografija raziskovalne skupine Laboratorija za raziskave v okolju. Pridobljeno 31.8.2013 s svetovnega spleta: http://sicris.izum.si/search/grp_search1.aspx?lang=slv

Sowa, J. F. (2000). Knowledge representation: Logical, philosophical, and computational foundations. Pacific Grove: Brooks Cole Publishing.

Tinta, T., Malej, A., Kos, M., Turk, V. (2010). Degradation of the Adriatic medusa *Aurelia* sp. by ambient bacteria. Hydrobiologia, 645(1), str. 179-191.

Univerza v Novi Gorici (2010). Univerza v Novi Gorici: Splošna brošura. Nova Gorica: Univerza v Novi Gorici.

Univerza v Novi Gorici (2012). Poročilo o delu Univerze v Novi Gorici v letu 2011. Nova Gorica: Univerza v Novi Gorici.

Vogrinčič, S., Bosnić, Z. (2011). Ontology-based multi-label classification of economic articles. Computer science and information systems, 8(1), str. 101-119.