# Употреба веб платформе Омека за дигиталне библиотеке из домена рударства

УДК 004.738.5:027.7]:004.420МЕКА

**САЖЕТАК:** У овом раду биће представљена Омека, веб платформа за приказивање дигиталних колекција и систем за управљање њиховим садржајем. Њену примену у области техничких наука, а конкретно у области рударства, приказаћемо на примеру дигиталне библиотеке ROmeka@RGF. За Омеку смо се определили првенствено због чињенице да је једноставна за коришћење, има обимну пратећу документацију и не захтева уско специфичне информатичке вештине што је чини приступачном за већину корисника, а нарочито за рударске инжењере, којима је ова дигитална библиотека првенствено намењена. Докумената прикупљена и ускладиштена у ову дигиталну библиотеку послужиће као подлога за даљи истраживачки рад, екстракцију терминологије, обележавање текста, екстракцију знања и др.

**КЉУЧНЕ РЕЧИ:** Омека, дигиталне библиотеке, рударство.

**РАД ПРИМЉЕН:** 2. септембар 2017.
**РАД ПРИХВАЋЕН:** 11. октобар 2017.

Александра Томашевић
aleksandra.tomasevic@rgf.bg.ac.rs

Биљана Лазић
biljana.lazic@rgf.bg.ac.rs

Далибор Воркапић
dalibor.vorkapic@rgf.bg.ac.rs

Михаило Шкорић
mihailo.skoric@rgf.bg.ac.rs

Љиљана Колоња
ljiljana.kolonja@rgf.bg.ac.rs

*Универзитет у Београду*
*Рударско-геолошки факултет*

## 1. Увод

За потребе истраживања приказаних у овом раду, креирана је дигитална библиотека ROmeka@RGF[1] са циљем да се прикупе, систематизују, обраде и похране стручни текстови из области рударства, чиме би се обезбедила основа не само за различита лингвистичка и

[1] http://romeka.rgf.rs

терминолошка истраживања, већ и за различите задатке из области инжењерства знања (не само екстракцију знања). Прикупљање, обрада и складиштење докумената у дигиталну библиотеку резултат су истраживачког рада на пројектима које финансира Министарство просвете, науке и технолошког развоја Републике Србије. Креирање дигиталне библиотеке ROmeka@RGF настало је из потребе да се текстови из домена рударства обједине и тако учине доступним за даља истраживања како инжењерима рударства, тако и лингвистима. Током истраживања, аутори нису пронашли дигиталне библиотеке за област рударства, осим часописа Подземни радови који је саставни део двојезичне библиотеке Библиша[2], те се ROmeka@RGF може сматрати првом такве врсте у Србији. Према нашем сазнању, једини пример дигиталне библиотеке на веб платформи Омека, а из домена геологије као рударству најближе области, представља дигитална библиотека „Геолошке књиге Филипа Глонжуа” (*Carnets géologiques de Philippe Glangeaud*[3]), развијена на Клермон универзитету.

У одељку 2. овог рада дат је опис развојног окружења веб платформе Омека, као и преглед најкориснијих програмских додатака. У одељку 3. описани су основни елементи дигиталне библиотеке као и начин њиховог креирања. Дигитална библиотека ROmeka@RGF описана је у одељку 4., а у одељку 5. приказане су могућности претраживања похрањених текстуалних ресурса уз коришћење алата и ресурса за српски језик. Могућности имплементације TEI смерница[4] дате су у одељку 6., док су у одељку 7. дати закључци и идеје за даљи развој.

## 2. Веб платформа Омека

За израду дигиталне библиотеке ROmeka@RGF коришћена је Омека, веб платформа за приказивање дигиталних колекција и систем за управљање њиховим садржајем[5]. За њен све развој заслужан је истраживачки тим Центра за историју и нове медије Рој Розенцвајг[6] са Универзитета Џорџ Мејсон (*George Mason*) у Вирџинији. Припада

[2] http://jerteh.rs/biblisha/ListaDokumenata.aspx?JCID=2&lng=en
[3] http://bibliotheque.clermont-universite.fr/glangeaud/
[4] TEI: Text Encoding Initiative http://www.tei-c.org/index.xml
[5] Content Management System (CMS)
[6] Roy Rosenzweig Center for History and New Media (RRCHNM), https://rrchnm.org/

групи софтвера отвореног кôда (*Open Source Software*), под лиценцом GPL v3.0 (*General Public License*)[7], што значи да је изворни кôд јавно доступан, те се платформа може побољшавати и прилагођавати потребама корисника. Изворно је осмишљена за научне институције које изучавају културно наслеђе, али данас је користе истраживачи многих других научних области.

Налази се на самом врху листе сродних софтвера због чињенице да је флексибилна и једноставна за коришћење. Карактерише је атрактиван и прилагодљив визуелни дизајн, једноставна инсталација, могућност проширивања која дозвољава измену постојећих и додавање нових функционалности, флексибилан приступ метаподацима, подршка за веб стандарде (CSS, HTML, RSS), увоз и извоз података у стандардизованим форматима (RDF, CSV, XML, JSON) (Kucsma et al., 2010).

Није осмишљена за ИТ-стручњаке и не захтева специфична информатичка знања, што корисницима омогућава да се фокусирају на садржај дигиталне библиотеке и њено описивање и тумачење, а не на програмирање. Има уграђене функције за каталогизацију и представљање дигиталних објеката, засноване на Даблинском језгру, којим се обезбеђује стандардизовано описивање и организовање дигиталних објеката.

### 2.1 Развојно окружење Омеке

Постоје две основне верзије веб платформе Омека, и то:

- Omeka.net, верзија која не захтева сопствени сервер. Капацитет простора за складиштење датотека ограничен је на 500 MB, величина датотеке оргранична је на 64 MB, број расположивих додатака ограничен је на 15, нема могућност функционалног прилагођавања и могућност измене изгледа је ограничена. Проширивање платформе је могуће, али захтева финансијска средства која, у зависности од врсте корисничког пакета, износе од 35 до 1000 долара годишње,
- Omeka.org, верзија која се инсталирати или на локални диск или као виртуелна машина и има могућност потпуног функционалног прилагођавања.

Пре покретања инсталације Омеке, потребно је извршити одговарајуће припреме на серверу, које подразумевају инсталирање:

[7] https://www.gnu.org/licenses/gpl-3.0.en.html

веб (HTTP) сервера Apache, система за управљање базама података MySQL (верзија 5.0 или новија) и интерпретатора програмског језика PHP (верзија 5.3.2 или новија). Дистрибуције оперативног система Linux, на којима Омека стабилно ради, су: Fedora, OpenSuse и Ubuntu. ROmeka@RGF је инсталирана на виртуелној машини са оперативним системом Ubuntu верзија 15.10.

Инсталирање платформе започиње креирањем MySQL базе са додељеним администраторским привилегијама. Најновија верзије Омеке, преузета са званичног сајта, распакује се и добијени директоријум (у даљем тексту: *omeka-root*) смешта се или у корени директоријум веб сервера или у неки његов поддиректоријум. Коришћење система за управљање базама података MySQL се омогућава изменама у датотеци *omeka-root/db.ini* (вредности поља: адреса рачунара (*database host*), корисничко име (*username*), лозинка (*password*) и име базе података (*database name*). Сви директоријуми, у којима се налази Омека, морају имати дозволу за уписивање. После завршетка инсталације је потребно кориговати привилегије. Инсталација се покреће путем веб читача уношењем ИП адресе или домена на коме се налази дигитална библиотека, где се захтева дефинисање администраторског налога и назива сајта.

Радно окружење веб платформе је вишејезично и преведено је, потпуно или делимично, на 50 језика. Српска верзија је једна од десетак потпуно преведених.

## 2.2 Програмски додаци Омеке

Изглед веб платформе може се мењати и прилагођавати избором неке од десетак понуђених тема или креирањем сопствене теме, а проширивање функционалности омогућено је применом додатака (енг. *plugins*). Развијено је око 80 додатака, прилагођених различитим верзијама Омеке. Корисници Омеке су развили велики број додатака, али како су углавном рађени за старије верзије Омеке, захтевају додатна прилагођавања најновијој инсталираној верзији платформе. У зависности од намене, додаци се могу сврстати у неколико група и то: за масовно креирање колекција и објеката, за организовање садржаја, за преглед датотека, за описивање доприноса заједнице, за геопросторну обраду и приказ на мапи. Овде ће бити приказано неколико додатака који су коришћени приликом израде наше дигиталне библиотеке или који су корисни и важни за сваког корисника Омеке.

*Archive Repertory* омогућава да се увезене датотеке на серверу чувају са изворним именима, као и да се групишу у хијерархијску структуру, најпре по објектима, а потом по колекцијама у којима се објекти налазе. Задржавање изворних имена чини URL-адресе датотека читљивијим, а уједно олакшава руковање датотекама.

*Bulk Metadata Editor* омогућава брзо и једноставно претраживање и измену метаподатака за велики број објеката истовремено. У првом кораку врши се избор објеката према различитим критеријумима (према тим критеријумима, могу се бирати: сви објекти дигиталне библиотеке, објекти у одабраној колекцији или објекти чији метаподаци задовољавају један или више задатих критеријума). У другом кораку бира се метаподатак који ће бити измењен. У трећем, завршном кораку, врши се сама измена метаподатка (та измена може подразумевати: претраживање и замену текста, додавање новог метаподатка у одабрано поље, додавање текста у постојеће метаподатке, уклањање дупликата и празних поља у одабраном опису објекта, уклањање дуплираних датотека у одабраним објектима или брисање свих постојећих метаподатака у одабраним пољима).

*Catalog Search* омогућава претраживање других каталога коришћењем поља dc:subject. Понуђени каталози су: *Archive Grid*,*Digital Public Library of America*, *Google Books*, *Google Scholar*, *Hathi Trust*, *JSTOR*, *Library of Congress*, *WorldCat*. Дата је и могућност додавања линкове ка каталозима других институција.

*COinS*[8] уграђује метаподатке о цитирању за сваки објекат (енг. *item*) у веб странице дигиталне библиотеке. Када је активиран, COinS омогућава видљивост објеката на on-line платформама, као што је Зотеро[9], аутоматским уграђивањем метаподатака о цитатима у стране других веб сајтова. Такође, у библиотеку платформе Зотеро могу се додати појединачни објекти било ког Омека сајта, док се додавање више објеката истовремено имплементира скриптама које раде у позадини. Додатак *COinS* олакшава истраживање и поспешује компатибилност са другим системима.

[8] ContextObjects in Spans (COinS), http://omeka.org/codex/Plugins/Coins_2.0, приступљено 28. маја 2017.

[9] Зотеро је некомерцијални софтвер намењен за прикупљање, уређивање и управљање библиографским белешкама и аутоматизовано форматирање библиографских референци. Као и Омека, развијен је у „Центру за историју и нове медије Рој Розенцвајг", https://www.zotero.org/

*Collection Tree* обезбеђује визуални приказ хијерархијске структуре колекција у дигиталној библиотеци, што олакшава њихово прелиставање.

*CSV Import* омогућава масован увоз метаподатака, ознака и датотека, приказаних табеларно у формату CSV. Уколико су називи колона усаглашени са Даблинским језгром, мапирање података врши се аутоматски; у супротном је неопходно податке мапирати ручно.

*Drop Box* омогућава администратору једноставнији увоз датотека које се налазе на серверу, у каталогу */plugins/Dropbox/files*. Преко административног сучеља датотеке се увозе појединачно или масовно одабиром са понуђеног списка.

*Dublin Core Extended* проширује листу метаподатака Даблинског језгра, чиме се обезбеђује потпуна анотација објеката. На 15 основних елемената (title, creator, subject, description, publisher, contributor, date, type, format, identifier, source, language, relation, coverage, rights) додаје скуп од 40 додатних елемената и то: abstract, access rights, accrual method, accrual periodicity, accrual policy, alternative title, audience, date available, bibliographic citation, conforms to, date created, date accepted, date copyrighted, date submitted, audience education level, extent, has format, has part, has version, instructional method, is format of, is part of, is referenced by, is replaced by, is required by, date issued, is version of, license, mediator, medium, date modified, provenance, references, replaces, requires, rights holder, spatial coverage, table of contents, temporal coverage, date valid.

*Geolocation* омогућава да се информације о локацијама у вези са дигиталним објектима додају на географску карту, уз могућност њиховог претраживања.

*Item Relations* омогућава креирање релације између дигиталних објеката. Додатак следи RDF модел дефинисања односа између објеката у виду RDF графа сачињеног од RDF тројки: субјекат-предикат-објекат. На пример, уколико је један дигитални објекат (RDF субјекат) део неког другог дигиталног објекта (RDF објекат), између њих се успоставља веза *isPartOf* (RDF предикат). Слично, ако је један дигитални објекат верзија неког документа, између тих дигиталних објеката се успоставља веза *isVersionOf*. На овај начин формирају се RDF тројке, које омогућавају касније истраживање текста коришћењем техника семантичког веба.

*Hide Elements* омогућава избор метаподатака који ће бити прикривени на форми за унос, на веб страни администратора или

на јавној доступној веб страни, као и на форми за претраживање метаподатака.

*METS Export*[10] омогућава извоз дигиталних објеката као METS XML датотека, појединачних докумената, колекција или дигиталне библиотеке у целини. Подржава га Иницијатива Федерације дигиталних библиотека[11] која предлаже XML шему метаподатака за управљање објектима дигиталне библиотеке и за размену тих објеката међу репозиторијумима или између репозиторијума и корисника. Има нарочиту улогу у окупљању и одржавању докумената који чине један дигитални објекат, с обзиром на њихов број и разноврсност. *METS Export* повезује више дигиталних докумената и омогућава навигацију кроз њих. Такође, укључује техничке информације неопходне за управљање дигиталним објектима: формат документа, технолошке карактеристике, начин скенирања, дигиталне трансформације. *METS Export* не прописује обавезни скуп описних метаподатака који ће се унети за дигитални документ, већ креатору метаподатака оставља одлуку о томе које ће описне метаподатке у ту сврху унети. Описни метаподаци за METS се могу на једноставан начин преузети из записа у Даблинском језгру (Тртовац, 2016).

*Neatline*, *NeatlineFeatures*, *NeatlineSmile*, *NeatlineText*, *NeatlineTime* и *NeatlineWaypoints* представљају серију додатака који омогућавају повезивање просторних и временских тачака на мапи са објектима у дигиталној библиотеци, као и повезивање докумената са Neatline изложбом. Ови додаци нису активирани у дигиталној библиотеци ROmeka@RGF због некомпатибилности са додатком *Geolocation*, који је за инжењерске потребе далеко значајнији.

*OAI-PMH Harvester* прикупља метаподатке од OAI-PMH[12] добављача података, мапира их у локалну базу и увози. Може се позивати једнократно или перманентно за ажурирање и синхронизацију дистрибуираних система. Тренутно може да увезе формате Даблинско језгро и METS.

*OAI-PMH Repository* припрема објекте за размену, представљајући инверзну функцију претходном додатку. Подржава Даблинско језгро, MODS и METS.

[10] The Metadata Encoding and Transmission Standard, https://www.loc.gov/standards/mets/METSOverview.v2.html

[11] Digital Library Federation, https://www.diglib.org/?s=mets

[12] Open Archives Initiative Protocol for Metadata Harvesting, https://www.openarchives.org/pmh/

*PDF Text* омогућава оптичко препознавање карактера текста[13], његову екстракцију из PDF формата и претраживање. Уколико текст није очитан на задовољавајући начин, дата је могућност његове корекције или увоз новог текстуалног документа у предвиђено поље.

*Reference* додаје стране са абецедним, односно азбучним, индексом унапред дефинисаних елемената на којима је могуће прелиставање по унапред задатим метаподацима.

*Search By Metadata* омогућава администратору да дефинише метаподатке за напредно претраживање на HTML страни.

*SimplePages* омогућава администратору да креира динамичке PHP странице не захтевајући притом специфична информатичка знања.

*SimpleVocab* и *SimpleVocabPlus* омогућавају креирање контролисаних речника и њихово синхронизовање на облаку. У дигиталној библиотеци ROmeka@RGF креиран је контролисани речник аутора, чиме су обезбеђени доследан унос и лакше претраживање.

*Text Analysis* повезује дигиталну библиотеку са *Watson Natural Language Understanding*[14] и *Mallet*[15] како би се омогућила анализа корпуса креираног од објеката коришћењем додатка *Ngram*.

*TEI Display* претвара (*render*) постављену TEI датотеку у визуелно јасан облик. Подразумевана XSLT трансформација омогућава два типа приказивања: приказује се или цео документ или његове појединачне целине. Први подразумева да се цео документ трансформише у HTML док други приказује садржај документа (div1 или div2), што је погодно за веће документе. Начин приказивања и XSLT трансформације се могу прилагодити, док се метаподаци из TEI заглавља могу аутоматски мапирати у поља Даблинског језгра за објекте и датотеке.

## 3. Креирање дигиталне библиотеке

Према једној од најчешће цитираних дефиниција дигиталних библиотека, коју је срочио Вилијам Армс (*William Y. Arms*), каже да су дигиталне библиотеке контролисане, систематски организоване колекције информација са придруженим сервисима, ускладиштене у дигиталном формату, којима се приступа преко мреже. Заједничко свим дигиталним библиотекама је да су информације организоване

[13] Optical character recognition (OCR)

[14] https://www.ibm.com/watson/services/natural-language-understanding/

[15] **MA**chine **L**earning for **L**anguag**E** **T**oolkit, http://mallet.cs.umass.edu/

на рачунарима и доступне корисницима преко интернет мреже, са процедурама за њихов избор, организовање и архивирање ради повећања доступности (Arms, 2000).

Дигиталне библиотеке су колекције дигиталних објеката, који су у виду различитих дигиталних података (текст, слика, звук, видео, анимација) или њихових комбинација (мултимедија) ускладиштени на мрежи, описани различитим метаподацима и повезани са другим информационим сервисима. Крајњи корисници им могу приступати и користити их без временског и просторног ограничења, а могу и сами да креирају нове или ажурирају постојеће објекте (Тртовац, 2016).

Основне елементе дигиталне библиотеке у Омеки чине:

– објекти (*items*),
– колекције (*collections*) и
– веб стране (*web pages*).

Дигитална библиотека може садржати неограничен број објеката, докумената, ознака и колекција. Једино ограничење је да један објекат може бити придружен искључиво једној колекцији (слика 1)[16]. За сваки од поменутих елемената дигиталне библиотеке је могућа потпуна контрола видљивости на вебу, почевши од појединачних метаподатака, па до сваког елемента у целини.


**Слика 1.** Дијаграм елемената дигиталне библиотеке у Омеки

[16] https://omeka.org/codex/Managing_Items, приступљено 19. маја 2017.

### 3.1 Објекти

Веб платформа Омека дизајнирана је за приказивање објеката који су основни елемент сваке дигиталне библиотеке. Због тога, креирање дигиталне библиотеке управо и започиње креирањем објеката.

Према врсти, објекти (односно: ставке, архиве, извори или ресурси) могу бити различити. Кориснику је понуђена листа од 15 основних врста објеката, уз могућност додавања нових врста, у складу са потребама. Доступне врсте објеката су:

- покретне слике (*Moving Image*), све форме видео записа: анимације, филмови, телевизијски програми;
- звучни записи (*Sound*), све форме аудио записа: аудио компакт дискови, снимљени говор или звукови;
- усмена предања (*Oral History*), информације добијене у интервјуима са особама које поседују знање из прве руке;
- статичне слике (*Still Image*), визуелни приказ текстова, слика, цртежа, графичког дизајна, планова и карата);
- веб стране (*Website*), HTML странице и повезане слике, аудио и видео датотеке, итд.;
- догађаји (*Event*), временски ограничене појаве, на пример: изложба, веб конференција, радионица, пожар, битка, суђење);
- електронска пошта (*Email*), текстуалне поруке са необавезним прилогом, које једна особа шаље другој особи или другим особама;
- план наставе (*Lesson Plan*), детаљан опис тока наставног процеса на курсу;
- особа (*Person*);
- интерактивни ресурси (*Interactive Resource*), веб странице, мултимедијални наставни објекти, сервиси за ћаскање;
- скупови података (*Dataset*), кодирани подаци у дефинисаној структури: листе, табеле и базе података;
- физички објекти (*Physical Object*), неживи, тродимензионални чврсти објекти, које у дигиталним библиотекама репрезентују типови као што су: покретне слике или статичне слике и др.;
- сервиси (*Service*), на пример: сервис за фотокопирање, банкарски сервис, међубиблиотечка размена или веб сервери;
- софтвери (*Software*);
- хипервезе (*Hyperlink*), веза или референца ка другом ресурсу на Интернету.

Објекти су скуп:

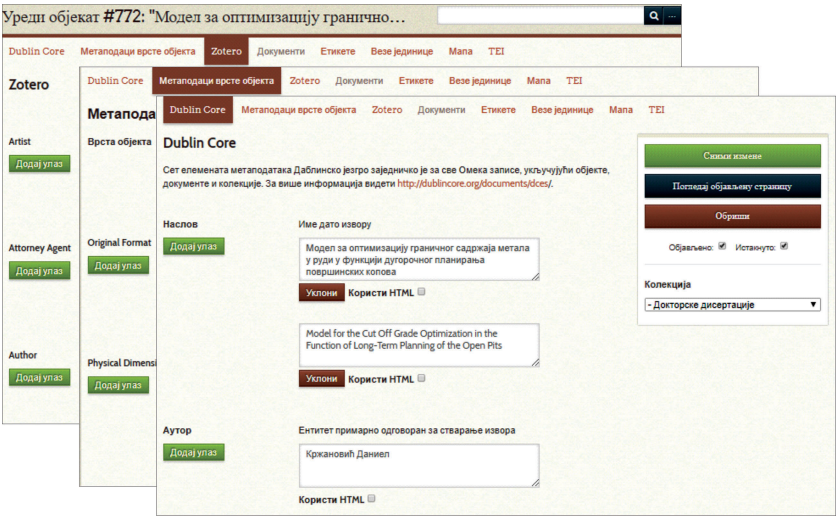
- метаподатака из Даблинског језгра који описују сам дигитални објекат,
- метаподатака који описују врсту објекта (*item type metadata*),
- етикета или ознака (*tag*) и
- докумената (*documents*).

Метаподаци су структурирани тако да опишу, објасне, идентификују, лоцирају или на неки други начин олакшају преузимање, коришћење или управљање изворoм информација (Hodge, 2001). Могу бити (Тртовац, 2016):

- *описни* описују ресурсе за потребе проналажења и идентификације и садрже основне елементе као што су: наслов, аутор, издавач, место, година, језик, јединствени идентификатор, опис, кључне речи, предметне одреднице, апстракт и сл.;
- *структурални* описују структуру сложених ресурса: типове, верзије, везе између дигиталних објеката, везе оригиналног документа и његових верзија укључујући податке о променама. и друге особине;
- *административни* дају информације о употреби и управљању ресурсима везано за интелектуално право и могу бити:
  - метаподаци о правима коришћења (дефинишу управљање правима приступа дигиталном објекту у складу са ауторским правима и заштитом интелектуалне својине);
  - технички метаподаци (доносе податке о датуму креирања, о техничким детаљима извора, величини и врсти датотеке, приступу извору, о свим изменама, о опсегу, о формату приказа);
  - метаподаци о очувању дигиталног објекта;
  - метаподаци о коришћењу (односе се на активно праћење броја корисника који посећују и користе одређени садржај, као и праћење употребе садржаја дигиталног објекта у новом контексту или у новој верзији путем преузимања метаподатака и дигиталног објекта за другу дигиталну библиотеку).

Објекти се најпре описује метаподацима из проширеног Даблинског језгра. Врста објекта се, такође, описује метаподацима, који се разликују у зависности од врсте. На пример, уколико је објекат „усмено предање”, тада су метаподаци који га описују: особа која врши интервју, интервјуисана особа, локација, транскрипција, трајање интервјуа итд.

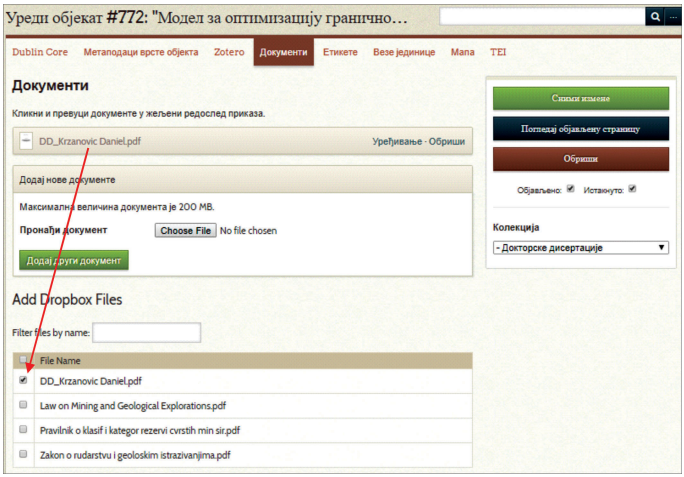
Уколико је објекат „веб страница”, тада се уноси само URL, а уколико је објекат „особа”, одговарајући метаподаци су: датум рођења, место рођења, датум смрти, занимање, биографија, библиографија итд. Објекат је могуће описати и метаподацима који ће га учинити видљивим у оквиру платформе Зотеро (слика 2).




**Слика 2.** Описивање објеката метаподацима

Документи који се придружују објектима могу се увозити појединачно и масовно, било помоћу додатка *Drop Box* (слика 3) било увозом из датотеке у формату CSV. Документи могу бити различитих формата, а неки од најчешће коришћених су:

– за текст: txt, css, csv, rtf, rtx, doc, docx, pdf, pps, ppt, pptx;
– за табеле: xls, xlsx;
– за базе података: mdb;
– за слике: bmp, gif, jpeg, jpg, tiff, png;
– за видео записе: avi, divx, mpeg, mov, mp4;
– за аудио записе: mp3, mid, midi, wav, wma;
– за извршне датотеке: exe, zip.




**Слика 3.** Панел за увоз докумената

Сваком објекту могу се доделити етикете или ознаке (*tags*)[17] хијерархијски неструктуриране кључне речи или фразе које класификују садржај тако да се он касније може лако проналазити. Слика 4 приказује панеле за унос геопросторних одредница, додељивање ознака и успостављања релација између објеката.

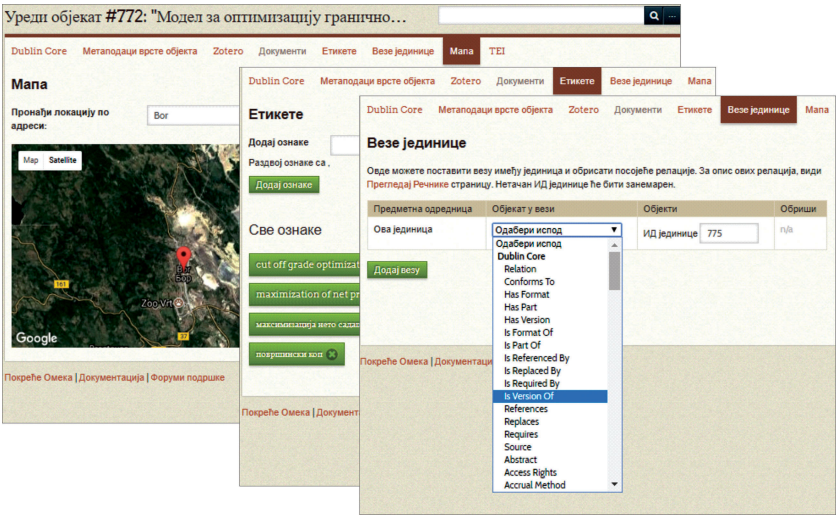
На панелу *Мапа* додају се информације о локацији која је у вези са објектом, уз могућност претраживања објеката према локацији.

Између два или више објекта могуће је успоставити релације. На пример, Закон о изменама и допунама Закона о безбедности и здрављу на раду је у релацији *isPartOf* са Законом о безбедности и здрављу на раду.

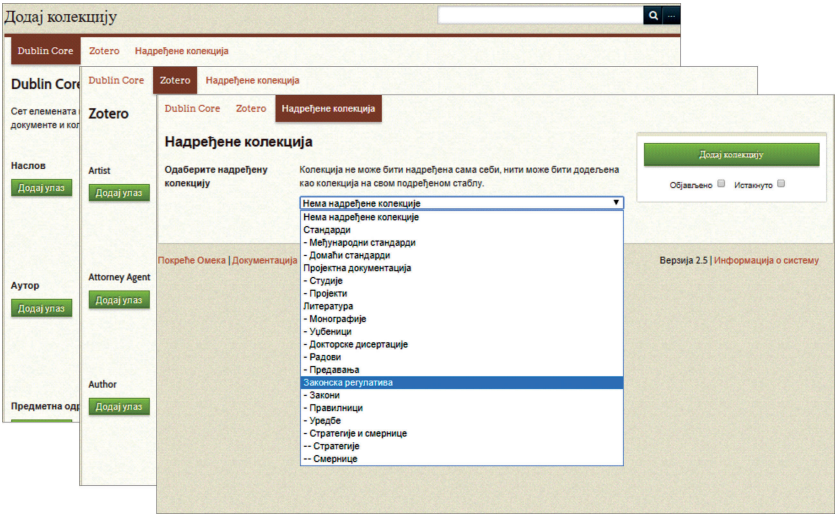
### 3.2 Колекције

Колекције представљају скупове објеката груписаних тако да се могу лакше претраживати. Детаљно се описује метаподацима из проширеног Даблинског језгра и са платформе Зотеро (слика 5). Могу имати хијерархијску структуру, што значи да могу имати дефинисане надређене и подређене колекције. Једна колекција не мора имати

[17] https://omeka.org/codex/Managing_Tags_2.0




**Слика 4.** Унос геопросторних одредница, додељивање ознака и успостављања релација између објеката



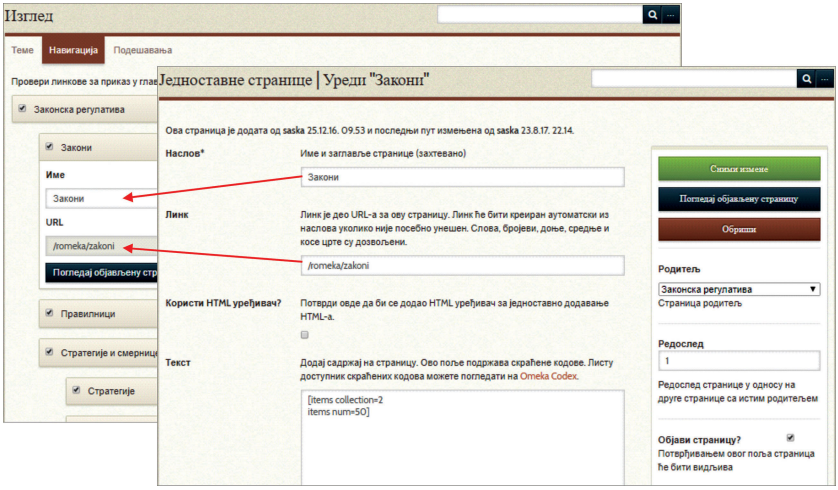

**Слика 5.** Панели за креирање колекција

подређене колекције или их може имати неограничен број, али може имати само једну надређену колекцију. За сваку колекцију могуће је дефинисати видљивост на веб страници.

### 3.3 Веб стране

Како би објекти и колекције били видљиви на вебу, неопходно је креирати веб странице. Осим назива веб стране неопходно је дефинисати и релативну путању до ње, као део URL-адресе те веб стране (слика 6). У делу *Текст* дозвољен је унос кодова са листе скраћених кодова, којима се прилагођава изглед саме стране. У примеру (слика 6) дат је кôд за приказ колекције која се идентификује са ID 2 (*Закони*), при чему се број приказаних објеката на једној страни ограничава на 50:

```
[items collection=2
 items num=50]
```




**Слика 6.** Панели за креирање колекција

После креирања веб стране, на панелу за уређивање изгледа сајта, у делу *Навигација* уносе се основни подаци (назив и URL) за сваку

појединачну веб страну, а њихов распоред уређује се једноставним превлачењем поља на жељену позицију. На овај начин веб стране се хијерархијски структурирају и на сајту се приказују у виду падајућих менија.

## 4. Дигитална библиотека ROmeka@RGF

Дигитална библиотека ROmeka@RGF[18] садржи 209 текстова првенствено из области рударства, али и из области безбедности и заштите на раду и процене ризика, као блиско повезаних области. Селекција докумената дигиталне библиотеке се базирала на доступним дигиталним ресурсима и за потребе ове дигиталне библиотеке није рађено додатно скенирање докумената расположивих искључиво у папирном облику.

Објекти су сврстани у 4 надређене колекције и 15 подколекција. Њихова хијерархијска структура приказана је у табели 1.

| **Колекција** | **Подколекције** | |
|---|---|---|
| Законска регулатива | Закони | |
| | Правилници | |
| | Уредбе | |
| | Стратегије и смернице | Стратегије |
| | | Смернице |
| Пројектна документација | Студије | |
| | Пројекти | |
| Литература | Монографије | |
| | Уџбеници | |
| | Докторске дисертације | |
| | Радови | |
| | Предавања | |
| Стандарди | Међународни стандарди | |
| | Домаћи стандарди | |

**Табела 1.** Структура колекција у дигиталној библиотеци

[18] Назив ROmeka@RGF је настао комбиновањем скраћеног назива Рударског одсека Рударско-геолошког факултета из Београда и веб платформе Омека

Сви објекти детаљно су описани свим врстама метаподатака који су наведени у одељку 3.1, изузев метаподатака о коришћењу чија се употреба планира у наредном периоду.

Видљивост на вебу омогућена је за већину дигиталних објеката, али ограничења су уведена за све објекте из колекција *Пројектна документација* и *Стандарди*, и неколико објеката из колекције *Литература*. Разлог за ово ограничење везано је за питање права публиковања, било због поверљивости података било због ауторских права.

Сви текстови који су ускладиштени у дигиталну библиотеку биће даље коришћени за терминолошка истраживања. У ту сврху из текстова су уклоњени делови на страном језику, табеле, слике, референце и линкови. Спајањем свих текстова, ради заједничке обраде, формирана је једна текстуална датотека величине 39 MB, са 6.200 страна текста формата A4. Након обраде текста добијено је: 150.365 реченица, 2.719.086 (100.414 различитих) једночланих лексичких јединица. Око 1900 једночланих термина, специфичних за области рударства, безбедности и заштите на раду и процене ризика, су у припреми, након чега ће уследити и екстракција вишечланих термина према методологији описаној у раду (Stanković et al., 2012). У плану је интегрисање претраге корпуса рударских текстова са Корпусом српског језика SrpKor.

## 5. Претраживање текстуалних ресурса

При потрази (која подразумева скуп метода и техника) за информацијама у текстуалним ресурсима, претражују се сами ресурси или траже метаподаци који описују те ресурсе (Baeza-Yates and Ribeiro-Neto, 1999). Системи за претраживање информација се базирају на два концепта: упит (енг. *query*) и објекат (енг. *object*). Упити су формални захтеви за потребним информацијама које корисник уноси у систем за претраживање информација. Објекти су ентитети који садрже тражене информације. Упити корисника се сравњују са објектима који су најчешће ускладиштени у базама података. Примери објектних података (енг. *object data*) су документи и веб стране.

Једноставно претраживање информација у текстуалним ресурсима се имплементира као сравњивање ниске карактера, (енг. *string matching*), не узимајући у обзир синтаксичка или семантичка својства тражене

речи. Ови упити се састоје од једне или више речи, које су евентуално повезане логичким операторима „и/или”.

Када је у питању претраживање дигиталне библиотеке, формулисање сложенијих упита је омогућено преко проширених регуларних израза. У зависности од ресурса који се претражује, одговори могу бити, на пример, документи, метаподаци или листа веб страница.

При претраживању се мора обратити пажња на три основне мере: одзив (енг. *recall*), прецизност (енг. *precision*) и рангирање (енг. *ranking*). Одзив изражава меру потпуности одговора који су добијени на постављени упит и представља се као количник укупног броја релевантних пронађених докумената и укупног броја релевантних докумената. Прецизност изражава меру коректности одговора који су добијени на постављени упит и представља количник укупног броја релевантних пронађених докумената и укупног броја пронађених одговора. Одзив указује на то колико је систем свеобухватан током претраживања релевантних информација.

Проблеми у претраживању текстуелних ресурса се могу класификовати у две категорије:

– општи, који не зависе од језика и
– проблеми који су специфични за поједини језик или групу језика.

Проблем при претраживању текстова на српском језику представљају различите кодне шеме као и постојање два алфабета (ћириличног и латиничног). То је случај са дигиталном библиотеком ROmeka@RGF с обзиром да су документа увежена о оригиналној форми, на оба писма. Претраживање садржаја само по једном писму је могуће у случају претраживања без проширења упита, док проширење упита аутоматски подразумева оба писма. Претраживање докумената на српском језику је сложен процес због веома богатог морфолошког система. Упитима се најчешће траже речи у свом канонском облику (номинатив једнине за именице, инфинитив за глагол) (Lazić et al., 2016). Међутим, документа могу садржати било који флективни облик променљиве речи. Овај проблем постаје сложенији ако се у обзир узму и сложене речи и синоними (Stanković, 2009).

Елементи дигиталне библиотеке које је могуће укључити у процес претраживања су: метаподаци, документи, етикете, извештаји, изложбе, веб странице. Изворно, претраживања се врше преко кључних речи, булових оператора и потпуним подударањем. Дигитална библиотека ROmeka@RGF унапређена је

имплементирањем проширених упита. Коришћени су веб сервиси (Stanković et al., 2012) и морфолошки електронски речници за српски језик (Krstev et al., 2008; Stanković et al., 2016):

– за морфолошко проширење упита:
  http://hlt.rgf.bg.ac.rs/vebran/api/delafs/ključna_reč
– за семантичко и морфолошко проширење упита:
  http://hlt.rgf.bg.ac.rs/vebran/api/sinonimi/ključna_reč

Претраживање без и са морфолошким и семантичко-морфолошким проширењима упита биће приказан на примеру претраживања лексеме *хомогенизација*.

Упитом без проширења тражи се само облик номинатива једнине:

*хомогенизација*

Упитом са морфолошким проширењем, у претрагу су укључени сви флективни облици тражене лексеме, на оба писма (латиници и ћирилици), тако да се у овом случају траже облици:

*homogenizacija, homogenizacijama, homogenizacije, homogenizaciji, homogenizacijo, homogenizacijom, homogenizaciju, хомогенизација, хомогенизацијама, хомогенизације, хомогенизацији, хомогенизацијо, хомогенизацијом, хомогенизацију.*

Упитом са семантичким и морфолошким проширењем, у претраживање су укључени и синоними и лексичке метонимије, као и њихови флективни облици:

*homogenizacija, homogenizacijama, homogenizacije, homogenizaciji, homogenizacijo, homogenizacijom, homogenizaciju, ujednačavanje kvaliteta uglja, homogenizacija kvaliteta uglja, homogenizacijama kvaliteta uglja, homogenizacije kvaliteta uglja, homogenizaciji kvaliteta uglja, homogenizacijo kvaliteta uglja, homogenizacijom kvaliteta uglja, homogenizaciju kvaliteta uglja, homogenizacija uglja, homogenizacijama uglja, homogenizacije uglja, homogenizaciji uglja, homogenizacijo uglja, homogenizacijom uglja, homogenizaciju uglja, upravljanja kvalitetom uglja, upravljanje kvalitetom uglja, upravljanjem kvalitetom uglja, upravljanjima kvalitetom uglja, upravljanju kvalitetom uglja, upravljanja kvalitetom, upravljanje kvalitetom, upravljanjem kvalitetom, upravljanjima kvalitetom, upravljanju kvalitetom, хомогенизација, хомогенизацијама, хомогенизације, хомогенизацији,*

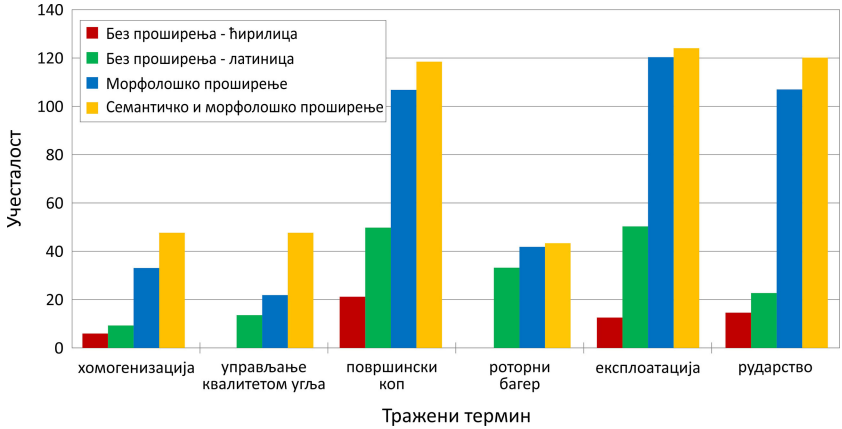
*хомогенизацијо, хомогенизацијом, хомогенизацију, уједначавање квалитета угља, хомогенизација квалитета угља, хомогенизацијама квалитета угља, хомогенизације квалитета угља, хомогенизацији квалитета угља, хомогенизацијо квалитета угља, хомогенизацијом квалитета угља, хомогенизацију квалитета угља, хомогенизација угља, хомогенизацијама угља, хомогенизације угља, хомогенизацији угља, хомогенизацијо угља, хомогенизацијом угља, хомогенизацију угља, управљања квалитетом угља, управљање квалитетом угља, управљањем квалитетом угља, управљањима квалитетом угља, управљању квалитетом угља, управљања квалитетом, управљање квалитетом, управљањем квалитетом, управљањима квалитетом, управљању квалитетом.*

Резултати претраживања дигиталне библиотеке ROmeka@RGF једночланим и вишечланим терминима (*хомогенизација*, *управљање квалитетом угља*, *површински коп*, *роторни багер*, *експлоатација* и *рударство*) приказани су у табели 2.

| **Врста упита** | **хомогенизација** | **управљање квалитетом угља** | **површински коп** | **роторни багер** | **експлоатација** | **рударство** |
|---|---|---|---|---|---|---|
| Без проширења ћирилица | 6 | 0 | 21 | 0 | 13 | 15 |
| Без проширења латиница | 9 | 14 | 49 | 33 | 50 | 23 |
| Морфолошко проширење | 33 | 22 | 106 | 42 | 120 | 106 |
| Семантичко и морфолошко проширење | 47 | 47 | 118 | 43 | 123 | 120 |

**Табела 2.** Резултати претраживања термина у дигиталној библиотеци

На дијаграму приказаном на слици 7 илустровани су резултати претраживања једночланих и вишечланих термина у дигиталној библиотеци ROmeka@RGF, а на основу података датих у табели 2. Уочљиво је да упити са семантичким и морфолошким проширењем дају знатно већи одзив у односу на упите без проширења.




**Слика 7.** Резултати претраживања једночланих и вишечланих термина у дигиталној библиотеци ROmeka@RGF

## 6. TEI (Text Encoding Initiative)

Са циљем екстракције информација из дигиталних објеката који су део дигиталне библиотеке ROmeka@RGF, одлучили смо се да објекте који су у текстуалном облику похрањујемо и у XML формату у складу са смерницама TEI[19]. За овај приступ смо се одлучили јер TEI представља *de facto* стандард за анотацију произвољних типова докумената, укључујући правне текстове и текстове пројектне документације. Анотација у складу са смерницама TEI треба да омогући повезивање делова текста у пројектној документацији који упућују на чланове закона и реферисане законске регулативе.

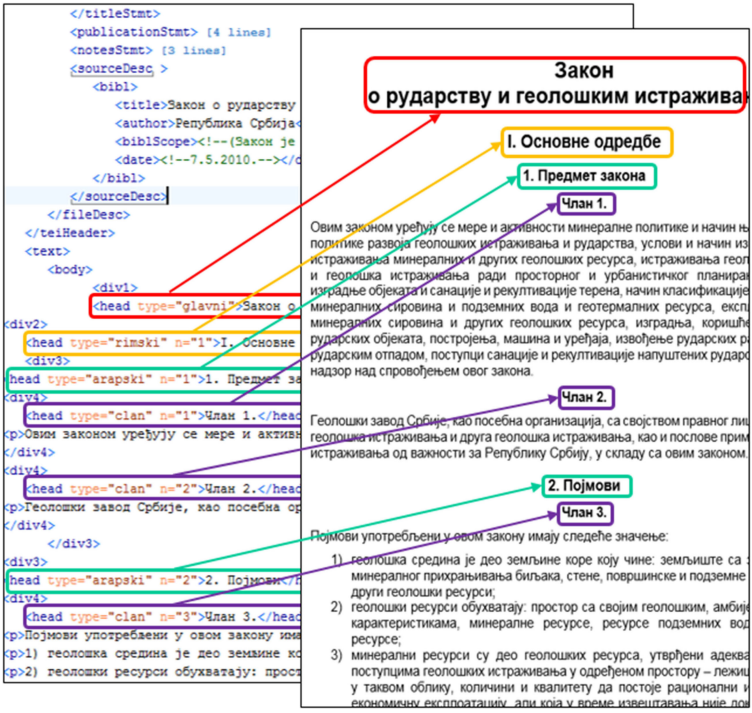
При анотацији у складу са смерницама TEI P5 (TEI, 2017) коришћене су етикете <div1>, <div2>, <div3>, <div4>, у складу са хијерархијским односом на нивоу закона, главе, поглавља и члана (Васиљевић, 2015). Етикета <p> коришћена је као еквивалент тачке, подтачке и алинеје закона.

За означавање наслова и поднаслова коришћене су етикете са различитим вредностима атрибута. У складу са потребама при повезивању, имајући у виду повезивање са члановима закона, користили смо се етикетама <head> за означавање наслова закона, главе, поглавља

[19] http://www.tei-c.org/

и члана. У склопу ове етикете коришћени су атрибути *type* и *n*. Вредности атрибута *type* су кључне за дистинкцију различитих делова закона. Могућа је једна од четири вредности овог атрибута: *glavni*, *glava*, *poglavlje* или *clan*. Вредности атрибута *n* су редни бројеви текућег елемента (дакле, главе, поглавља или члана).

Слика 8 приказује део документа Закон о рударству и геолошким истраживањима означеног у складу са смерницама TEI P5 на претходно описан начин.




**Слика 8.** Део документа означен у складу са смерницама TEI P5

Имајући у виду представљање рударских пројеката, потрудили смо се да, користећи TEI, сачувамо све елементе, а посебно табеле, као битан

```
<table rows="12" cols="2">
   <head>Tabela 1.5.1.</head>
   <row><cell>Parametar</cell> <cell>Vrednost</cell></row>
   <row><cel1>Vlaga, &</cell> <cell>39,22</cell></row>
   <row><cell>Pepeo, %</cell> <cell>17,70</cell></row>
   <row><cell>S ukupni, %</cell> <cell>1,18</cell></row>
   <row><cell>S sagorljiv, %</cell> <cell>0,56</cell></row>
   <row><cell>S u pepelu, %</cell> <cell>0,60</cell></row>
   <row><cell>Koks, %</cell> <cell>34,98</cell></row>
   <row><cell>C-f ix, %</cell> <cell>18, 26</cell></row>
   <row><cell>Isparljivo, %</cell> <cell>26,30</cell></row>
   <row><cell>Sagorljivo, %</cell> <cell>43,14</cell></row>
   <row><cell>Gornja toplota sagorevanja, kD/kg</cell> <cell>11.490</cell></row>
   <row><cel1>Donja toplota sagorevanja, kJ/kg</cell> <cell>10.020</cell></row>
</table>
```

**Слика 9.** TEI P5 репрезентација табеларног приказа

извор информација. Слика 9 илуструје репрезентацију једног табеларног приказа у оквиру TEI верзије пројектне документације.

## 7. Закључак

У раду су проказане предности и мане Омеке као платформе за развој овог типа библиотеке. Уз то, показано је колико примена морфолошких речника утиче на квалитет претраге приликом морфолошког и семантичког проширења упита. Закључак је да комбинација Омеке и морфолошких речника доприноси бољој организацији и претраживости објеката који су похрањени у дигиталну библиотеку ROmeka@RGF.

Имајући у виду одржавање актуелности самих колекција, као и алата за претраживање, у плану је даље праћење развоја нових додатака и технологија који би допринели и повезивању дигиталне библиотеке ROmeka@RGF са другим изворима сродних информација. Такође, је у плану и допуњавање дигиталне библиотеке новим, разноврснијим дигиталним објектима. У плану је и рад на на екстракцији информација из самих дигиталних објеката.

## Захвалност

Истраживање приказано у овом раду је финансијски подржано од стране Министарства просвете, науке и технолошког развоја Републике

Србије, преко пројеката ТР33039 („Унапређење технологије површинске експлоатације лигнита у циљу повећања енергетске ефикасности, сигурности и заштите на раду”).

## Литература

Arms, William Y. *Digital libraries*. Cambridge, Massachusetts, USA: M.I.T. Press, 2000. http://www.cs.cornell.edu/wya/diglib/ms1999/

Baeza-Yates, Ricardo and Berthier Ribeiro-Neto. *Modern information retrieval*, Vol. 463, New York, USA: ACM Press, 1999. http://web.cs.ucla.edu/~miodrag/cs259-security/baeza-yates99modern.pdf

Hodge, Gail. *Metadata made simpler*. Niso Press, 2001.

Krstev, Cvetana, Ranka Stanković, Dusko Vitas and Ivan Obradović. “The Usage of Various Lexical Resources and Tools to Improve the Performance of Web Search Engines”. In *LREC*, Paris, France: European Language Resources Association (ELRA), 219–224, 2008. http://lrec-conf.org/proceedings/lrec2008/pdf/67_paper.pdf

Kucsma, Jason, Kevin Reiss and Angela Sidman. “Using Omeka to build digital collections: The METRO case study”. *D-Lib magazine* Vol. 16, no. 3/4 (2010). http://webdoc.sub.gwdg.de/edoc/aw/d-lib/dlib/march10/kucsma/03kucsma.html

Lazić, Biljana, Danica Seničić, Aleksandra Tomašević and Bojan Zlatić. “Terminological and Lexical Resources Used to Provide Open Multilingual Educational Resources”. Belgrade, Serbia, 2016. http://www.baektel.eu/documents/conferences/eLearning_2016_BL_DS_AT_BZ.pdf

Stanković, Ranka, Cvetana Krstev, Ivan Obradović, Aleksandra Trtovac and Miloš Utvić. “A tool for enhanced search of multilingual digital libraries of e-journals”. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC)*, Istanbul, Turkey, 1710–1717, 2012. http://www.lrec-conf.org/proceedings/lrec2012/pdf/375_Paper.pdf

Stanković, Ranka, Cvetana Krstev, Biljana Lazić Ivan Obradović and Aleksandra Trtovac. “Rule-based Automatic Multi-Word Term Extraction and Lemmatization”. In *Proceedings of the 10th International Conference on Language Resources and Evaluation, LREC*, 507–514, 2016. http://www.lrec-conf.org/proceedings/lrec2016/pdf/1033_Paper.pdf

Stanković, Ranka M. “Modeli ekspanzije upita nad tekstuelnim resursima”. Doktorska disertacija. Univerzitet u Beogradu, Matematički fakultet, 2009.

TEI-Consortium. “TEI P5: Guidelines for Electronic Text Encoding and Interchange. Version 3.2.0.”, 2017. Приступљено 23.8.2017, http://www.tei-c.org/release/doc/tei-p5-doc/en/html/

Васиљевић, Небојша. “Аутоматска обрада правних текстова на српском језику”. Докторска дисертација. Универзитет у Београду, Филолошки факултет, 2015. https://fedorabg.bg.ac.rs/fedora/get/o:10687/bdef:Content/get

Тртовац, Александра С. “Дескриптори метаподатака и дескриптори садржаја у проналажењу информација у дигиталним библиотекама”. Докторска дисертација. Универзитет у Београду, Филолошки факултет, 2016. https://fedorabg.bg.ac.rs/fedora/get/o:12605/bdef:Content/get