Инфотека (Q25460443) у Википодацима

УДК 004.62: [030:004.738.5

Ранка Станковић
ranka.stankovic@rgf.bg.ac.rs
Универзитет у Београду
Рударско-геолошки факултет
Београд, Србија

Лазар Давидовић
lazarmdavidovic@gmail.com
Универзитет у Београду
Београд, Србија

САЖЕТАК: *Википодаци* (енгл. Wikidata) су база знања Задужбине Викимедија која представља заједнички извор различитих врста података које користе не само други Википедијини пројекти, већ све више и бројне апликације семантичког веба. У овом раду ћемо презентовати пример интеграције Википодатака са дигиталним библиотекама и екстерним системима, као и могућност убрзања припреме и уноса података на примеру радова из часописа за дигиталну хуманистику *Инфотека*.

КЉУЧНЕ РЕЧИ: семантички веб, отворени повезани подаци, викподаци, Инфотека, метаподаци часописа

РАД ПРИМЉЕН: 24. јун 2021.
РАД ПРИХВАЋЕН: 16. јули 2021.

1. Увод

Википодаци (енгл. Wikidata)[1] су база знања Задужбине *Викимедија* која представља заједнички извор различитих врста података, како конкретних тако и апстрактних. Похрањене податке могу да користе и други Викимедијини пројекти, као што је *Википедија*, али и шира заједница за различите потребе, чиме доприносе померању границе од машинске читљивости ка машинској разумљивости података на вебу. У овом раду ћемо презентовати пример интеграције Википодатака са дигиталним библиотекама и екстерним системима, као и могућност убрзања припреме и уноса података на примеру радова из часописа за дигиталну хуманистику *Инфотека*.

Семантички веб је проширење постојећег веба где се информацијама даје прецизно дефинисано значење и који омогућава бољу сарадњу рачунара и корисника. Отвореност и делимична структурираност ресурса који се развијају у организацији Викимедије је била основа за

1. Викидата

изградњу бројних машински читљивих ресурса, какав је на пример *DBPedia*,[2] који се ослањају на стандардизоване језике семантичког веба. Концепт семантичког веба и технологије отворених повезаних података, проширују традиционални веб употребом стандардног језика за означавање и сродних алата за обраду, где управо RDF (Resource Description Framework), оквир за описивање ресурса на вебу, има велику улогу и омогућава ефикаснија решења за проналажење информација (Shah et al. 2002). Да би семантички веб функционисао, рачунари треба да имају приступ структурисаним колекцијама информација и да утврде дефинисана правила аутоматизованог управљања. Википодаци се управо уклапају у те трендове развоја информационих технологија, које померају границе од машинске читљивости ка машинској разумљивости података на вебу.

Пројекат *Scholia*[3] (Nielsen, Mietchen, and Willighagen 2017) један од првих свеобухватних подухвата осмишљених да се библиографске информације, научни профили аутора и институције представе користећи Википодатке. Управо резултати овог пројекта и доступност радова *Инфотеке* у различитим дигиталним облицима су били инспирација за „викификацију“ радова часописа *Инфотека*. Видевши конкретно веб стране *scholiaEvent*[4] и *scholiaTopic*,[5] покренута је слична акција, да се на основу постојећих метаподатака о радовима из часописа

2 warnings *Инфотека* креирају повезани (RDF) подаци о ауторима и радовима, а на сајту дигиталне библиотеке *Библиша*[6] (Stanković et al. 2015) уз радове из часописа додају линкови ка Википодацима и прикаже граф коауторства, као интерактивна страница. Имплементација представља студију случаја која се може даље проширити на друге часописе, као и на друге случајеве употребе, на пример, конференције и дигиталне библиотеке. Алат *Scholia* се развија у оквиру веће иницијативе *WikiCite*,[7] која настоји да индексира библиографске метаподатке у Википодацима о ресурсима који се

2. DBPedia
3. Scholia, Scholia у Википодацима
4. Преглед протеклих и предстојећих конференција са подсетником са роковима за слање радова ScholiaEvent
5. Преглед научних и стручних радова, као и њихови аутори и теме које се заједнички јављају, груписани по тематским целинама: Википедија, машинско учење, биологија, храна и сл. scholiaTopic
6. Библиша
7. WikiCite

могу користити за поткрепљивање тврдњи изнетих у Википодацима, Википедији или негде другде. У времену када смо преплављени нетачним информацијама на вебу, одговарајуће поткрепљивање информација релевантним изворима сигурно игра важну улогу. Како смо желели да аутоматизујемо, колико год је могуће, процес припреме и уноса података, истражили смо различита решења од којих смо користили два: *OpenRefine*[8] и *QuickStatements*,[9] о чему ће више речи бити у наредним одељцима.

Сарадња *Викимедије Србије*[10] са Универзитетом у Београду има дугу традицију (Stakić 2009). Универзитетска библиотека „Светозар Марковић“ заједно са Математичким факултетом Универзитета у Београду и Викимедијом Србије је 2015. године почела пројекат *Викибиблиотекар*, са идејом да се постави што више квалитетних садржаја на Википедију (Popović, Ševkušić, and Stakić 2015). Википодатке, као мрежу отворених података је користила Андоновски (2020) за описивање језичких ресурса, конкретно романа, који су саставни део српско-немачког литерарног корпуса (Andonovski, Šandrih, and Kitanović 2019). На Рударско-геолошком факултету се студенти више година обучавају за унос и коришћење Википодатака,[11] а на докторским студијама *Интелигентни системи*, при Универзитету се у оквиру предмета *Представљање знања* и *Семантички веб* истражују апликативне могућности отворених података. У оквиру COST акције CA16204 (2017-2021) „Удаљено читање за европску историју књижевности“[12] се ради на уносу метаподатака о српским романима из корпуса *srpELTeC*[13] (Krstev et al. 2019) и повезивању Википодатака са различитим апликацијама, од којих је једна Аурора.[14] Резултатима који се представљају овим

8. OpenRefine (претходно Google Refine) је алат за рад са неуређеним подацима: чишћење, претварање из једног формата у други, уз допуњавање екстерним подацима путем веб сервиса OpenRefine

9. Алат за уређивање ставки Википодатака: додавање и уклањање изјава, ознака, описа и сл. QuickStatements

10. Vikimedija

11. Унос и коришћење Википодатака

12. Један од најважнијих циљева ове акције је припрема вишејезичног корпуса (названог European Literary Text Collection - ELTeC) који ће, када буде потпуно завршен, садржати по 100 романа први пут објављених у периоду 1840-1920. за више европских језика.

13. srpELTeC

14. Аурора

радом су допринели и чланови Друштва за језичке ресурсе и технологије *JePTex*.[15]

2. Википодаци

Википодаци (енгл. Wikidata) су база знања чија је сврха да буде заједнички извор одређених врста података (нпр. број становника државе, место рођења, датум оснивања), које користе други Викимедијини пројекти као што је Википедија. У том смислу је слична Викимедијиној остави где се складиште медијске датотеке којима се приступа из других Викимедијиних пројеката. Википодаци су оријентисани на документе, усредсређени на ставке, које представљају теме, концепте или објекте. Свакој ставци додељен је јединствени, трајни идентификатор, позитиван цео број са префиксом великог слова Q, познатог као „QID“. Ово омогућава превођење основних информација потребних за препознавање теме коју ставка обухвата, а да се не фаворизује било који језик, са циљем да се обезбеди јединственост значења конкретног појма.

Наведимо неке примере ставки: места (Нови Сад: Q55630, Лондон: Q84, Звездара (Београд): Q12645852), особе (Ђорђе Балашевић: Q342045, Тим Бернерс-Ли: Q80, Хеди Ламар: Q49034), догађаје (Први српски устанак: Q368689, концерт: Q182832, маратон: Q40244), предмете (столица: Q15026, чаша: Q81727, тигањ: Q127666)), појмове (радост: Q935526, страх: Q44619, појам: Q151885), књижевна дела (*Горски вијенац*: Q1192476, *Дон Кихот*: Q480, *Игра престола* (књига): Q1751870), филмови (*Лепота порока*: Q4239792, *Коса* (филм): Q757156), серије (*Игра престола*: Q23572, *Ало, ало!*: Q425628), балет (*Дон Кихот* (балет): Q1239463)… Концепти који стоје иза ставки треба да буду јединствени, али се дешава да постоје две ставке под истим називом, Никола Тесла (Q9036) представља славног научника, а *Никола Тесла* (Q2732597) представља насеље (Q486972) у Нишкој Бањи (Q954986) које је по њему добило име. Препорука је да се код вишезначних ентитета у загради дâ додатно појашњење, као што је претходно наведени *Дон Кихот* (балет) или *Игра престола* (књига). Дакле, ставка је повезана са јединственим идентификатором (QID), идентификатор је повезан са паром: насловом и описом, како би се уклонила било каква двосмисленост.

15. JePTex

Универзитет у Београду је државни универзитет и члан Европске асоцијације универзитета	
Универзитет у Београду је основао Доситеј Обрадовић 1808.	
Q240631 P31 Q875538. Q240631 P463 Q868940.	Q240631 P31 Q875538; P463 Q868940; P112 Q347659; P571 „1808“.
Q240631 P112 Q347659. Q240631 P571 „1808“.	
Доситеј Обрадовић је рођен 17. фебруара 1742 у Чакову, а умро је 7. априла 1811. Био је лингвиста, песник, писац и филозоф.	
Q347659 P569 „17. фебруар 1742“. Q347659 P19 Q325736. Q347659 P570 „7. април 1811“. Q347659 P106 Q14467526. Q347659 P106 Q49757. Q347659 P106 Q36180. Q347659 P106 Q4964182.	Q347659 P569 „17. фебруар 1742“; P19 Q325736; P570 „7. април 1811“; P106 Q14467526; P106 Q49757; P106 Q36180; P106 Q36180.

Табела 1. Примери ставки Википодатака

Идентификатор ставке (QID), осим што је повезан са насловом и описом, може имати више псеудонима и одређени број изјава (тврђења, израза) којима се представљају њена својства и вредности. Изјава је уређена тројка: (ставка, својство, вредност), где је ставка (Q) – било која тема (особа, предмет, место, концепт), својство (P). Релација[16] или карактеристика релевантна за ставку може бити на пример: боја косе (P1884) за људе, издавач (P123) за публикована дела, оснивање (P571) за организације и сл. Вредност ставке може бити сам „литерал“ односно ниска карактера (на пример: дужина Дунава је 2860 км) или референца на неку другу ставку (на пример главни град Србије је Београд). Ставка може бити описана низом изјава од којих свака даје једну чињеницу или податак о ставки. У табели 1 је дато неколико примера реченица на природном језику и записивање истих информација на Википедији, у

16. У математици ако је уређени пар (x, y) у релацији ρ тада кажемо да је елемент x у релацији са елементом y и пишемо као тројку: $x \rho\, y$. Слично, ставке у Википодацима релације изражавају као тројке, тако да релација Тесла-начин живота-вегетаријанство се бележи као Q9036 P1576 Q83364.

виду тројки субјекат, предикат и објекат (лева колона) и у скраћеној нотацији (десна колона).

У претходној табели, иза примера на српском језику су у првој колони дате тројке, односно реченице облика субјекат-предикат-објекат. Ако желимо да се прецизније изразимо, рећи ћемо да су то RDF тројке, где је RDF скраћеница од Resource Description Framework (Q54872), односно оквир за описивање ресурса на вебу. Реченице се завршавају тачком. У другој колони је приказана скраћена нотација, којом се избегава понављање субјекта, тако да знак „;“ упућује да се следећи предикат односи на исти субјекат.

Важна карактеристика Википодатака је да имају два лица: једно намењено људима и друго намењено машинама, што омогућава бројне примене у обради природних језика. Поменимо неке: класификација текста, индексирање, анализа текста, генерисање текста, сумаризација, нормализација, повезивање и сл. Друга важна карактеристика је вишејезичност, која омогућава да се свака ставка може повезати са називом на било ком језику који је регистрован у Викимедијиним ресурсима, што отвара пут бројним применама, почев од аутоматског превођења и класификације вишејезичних докумената, до анализе садржаја на вебу и друштвеним мрежама.

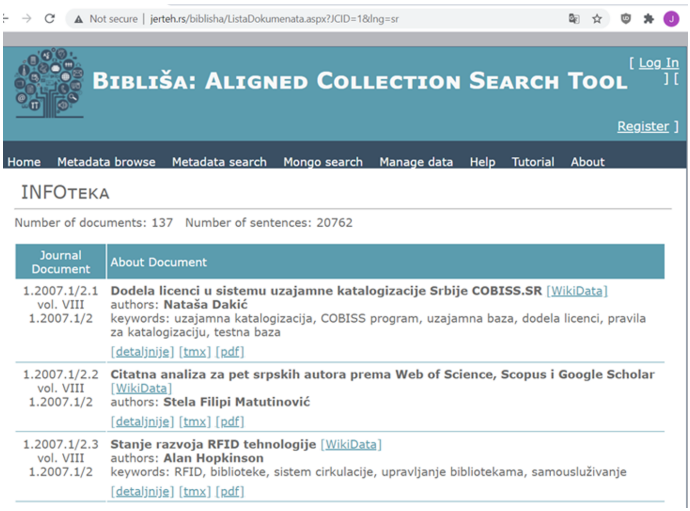
3. Аутоматизација уноса у Википодатке

Појединачни унос података је често временски захтеван посао, али се може убрзати када већ постоје подаци похрањени у различитим дигиталним форматима. Уз одговарајућу припрему, они се могу полуаутоматски увести у Википодатке. Дакле, основна идеја је да се убрза унос података, о резултатима истраживања у области дигиталне хуманистике у Србији и старих српских романа, како би се повећала видљивост уједно и српског језика, нашег културног наслеђа и резултата истраживања у Србији, а свакако отворио пут и за многе друге скупове података.

Да би аутоматизација уноса била могућа, први корак који је требало урадити јесте прикупљање и припрема података. Други корак се односи на избор назива ставки у подацима које ће се користити за идентификацију предиката и креирање шеме уноса. Шемом дефинишемо повезивање вредности са ставком, односно, субјектом помоћу предиката као посредника.

Иако је крајњи циљ унос података о радовима, унос њихових аутора је био неизоставан корак и предуслов даљем раду. По завршеном уносу креирани су SPARQL[17] упити за различите приказе, користећи и интегрисане технологије у Википодацима за визуализацију резултата.

Сваки рад часописа Инфотека из двојезичне дигиталне библиотеке Библиша је повезан са припадајућим Википодацима, тако да може директно да се дође до конкретног, појединачног приказа у Википодацима, али и да се интегришу неки од корисних визуалних приказа у оквиру саме апликације. Слика 1 приказује пример панела у Библиши за приказана прва три рада у колекцији, где се види да се иза наслова сваког рада налази линк ка путањи ресурса у Википодацима.



Слика 1. Панел дигиталне библиотеке Библиша са прегледом метаподатака о Инфотеци

Први рад са листе би се превео на језик Википодатака на следећи начин:

17. SPARQL је протокол и упитни језик за RDF базе података који омогућава извлачење вредности, истраживање структуре података, спајања података који долазе из различитих база, трансформацију RDF података из једног речника или графа у други. Спецификација упитног језика SPARQL.

„Додела лиценци у систему узајамне каталогизације Србије COBISS.SR“ (Q98785010)
instance of (P21) academic journal article (Q18918145);
author P(50) Наташа Дакић (Q99281474).

Може се видети да је рад представљен идентификатором Q98785010, да је примерак (P21) класе академских чланака (Q18918145) и да има аутора Q99281474.

Већ поменути алат *OpenRefine*, који је иницијално развио Google, и алат *QuickStatements* који је развио члан тима Википодатака, Магнус Манске се често користе заједно и може се рећи да се међусобно допуњују. *QuickStatements* користи текстуални TSV или CSV формат који се ефикасно генерише алатом *OpenRefine*. Разлика између ова два алата се огледа и у грануларности трансакција, јер *OpenRefine* уноси измене у једном кораку, па разрешавање грешака насталих при уносу може довести до дуплирања података, док *QuickStatements* појединачно уноси сваку ставку и омогућава боље праћење целокупног процеса. Примери добре праксе упућују да се *OpenRefine* користи за припрему уноса у базу Википодатака, док се физички унос RDF тројки ради коришћењем алата *QuickStatements*.

Први корак је свакако припрема података у виду *CSV* датотеке, потом креирање *OpenRefine* пројекта и учитавање припремљених података. Даље следи препознавање постојећих ставки у Википодацима – неопходан корак који треба да омогући повезивање садржаја датотеке са идентификаторима (*QID*) постојећих ставки и унос нових, уколико оне већ не постоје. У овој фази је ручна провера и евентуална промена неопходна. Креирање схеме скупа за унос дефинише предикате који ће повезивати субјекте и објекте у RDF тројкама и врло је важан корак. Наводимо карактеристичне примере са коментарима:

- наслов (P1476), на енглеском и српском (оба писма, ћирилицом и латиницом због претраге);
- главна тема стваралачког дела (P921), кључне речи, при чему су оне које су већ постојале повезане, а нове су додате као примерци са називима ставки на српском и енглеском;
- издавач (P123);
- језик дела или имена (P407);
- датум издавања (P577), представљен само годином;
- број страна (P1104), том (P478), публикација (P433);
- објављено у (P1433) Инфотека (Q25460443);

– лиценца (P275);
– пун текст доступан на (P953).

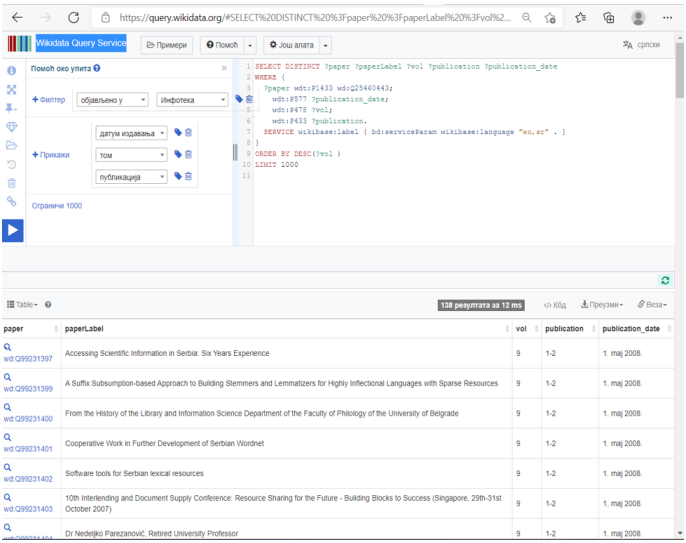
Имајући у виду да се својства ређе додају, препорука је да се истраже слична својства и својства сличних ресурса, пре него се донесе одлука за њихово додавање. Уз то, посебну пажњу треба посветити и ограничењима својстава која се могу видети у сугестијама при уносу. Имајући у виду здружени рад дистрибуираних корисника, неминовно је да се понекад могу наћи дупликати Википодатака, а решење за те ситуације је опција спајања дупликата или елиминисања, о чему се може више видети на одговарајућој страници.

Након иницијалног уноса података, даљи унос је настављен после сваког публикованог броја *Инфотеке*, тако да се сада може пронаћи 138 радова из *Инфотеке*. Креиран је HTML који интегрише сервис за постављање упита *Wikidata Query Service* са *Библишом*. Написани су упити који добављају табеле последњих објављених радова, заступљености кључних речи у радовима, слике аутора, табелу профила аутора, граф коауторства, дистрибуције аутора по полу, и сл. Подаци о ауторима обухватају елементарне податке које свакако у наредном периоду треба допунити новим садржајима: институцијом у којој раде, областима истраживања, референцама ка референтним истраживачким базама и сл.

Као пример наводимо једноставан упит који приказује списак радова из Инфотеке, број, волумен и датум публиковања.

```
SELECT ?paper ?paperLabel ?vol ?publication ?publication_date
WHERE {
?paper wdt:P1433 wd:Q25460443;
       wdt:P577 ?publication_date;
       wdt:P478 ?vol;
SERVICE wikibase:label {bd:serviceParam
                        wikibase:language "en,sr".}
}
ORDER BY DESC(?vol )
```

Илустрација претходног упита у радном окружењу *Wikidata Query Service* се може видети на слици 2, где се горњи део панела користи за креирање упита, а у доњем се приказују резултати, при чему се начин приказа може бирати у зависности од типа упита (табела, графички приказ, мрежа, приказ на временској линији, мапа и сл.).



Слика 2. Окружење *Wikidata Query Service* за рад са упитима

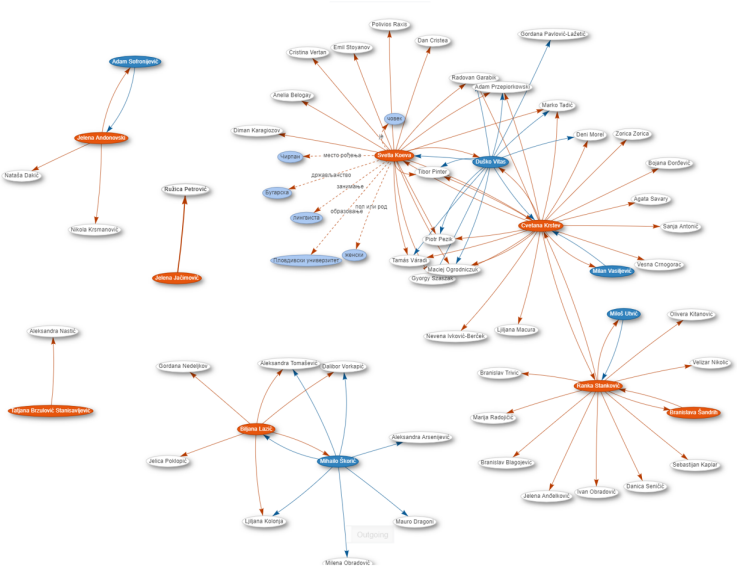
На слици 3 се може видети део графа коуторства[18] који прибавља Википодатке путем сервиса за претрагу *Wikidata Query Service*.

4. Закључак и даљи планови

Позитивна искуства у раду са Википодацима *Инфотеке* су искоришћена и за унос података о романима и о ауторима романа у Википодатке из вишејезичне колекције ELTEC (European Literary Text Collection) чија ће се једна потколекција састојати од 100 српских романа из периода 1840-1920 коју у оквиру Cost акције: CA16204 – *Distant Reading for European Literary History* развијају чланови Друштва *JePTex* под руководством Цветане Крстев и Ранке Станковић. Припремљен је скуп метаподатака за романе који су до сада дигитализовани и обрађени према захтевима Акције. Рад на Википодацима видимо као континуалну активност, при чему ће се посебна пажња обратити на повезане отворене податке из области лингвистике – LLOD и њихове примене. Морамо

18. Граф коауторства је доступан, може се приступити самом SPARQL упиту или променити опција приказа да буде табеларан, на временској скали, графички, и слично.

свакако бити свесни проблема и ограничења Википодатака и других отворених повезаних података, како бисмо истражили могућности њиховог превазилажења или бар ублажавања.



Слика 3. Граф коаутора на радовима у *Инфотеци*

Википодаци су заиста огромна база знања, која је 1) доступна свима – за читање информација, постављање упита, уређивање и унапређивање; 2) отворена – вишеструка употреба обезбеђена је бесплатном Creative Commons CC0 лиценцом, која даје пуну слободу за коришћење података; 3) вишејезична – ентитети се могу именовати и описивати на било ком природном језику. Ова три кључна својства су главни покретачи за бројне апликације, што верујемо да ће још више инспирисати вики заједницу да се овом ресурсу посвети више пажње. Такозвани мали језици, какав је српски, треба да користе све начине да се изборе за своје место у дигиталном простору, па активности на овом пројекту и сродним пројектима и иницијативама видимо као скроман допринос очувању српског језика у дигиталном добу.

Литература

Andonovski, Jelena, Branislava Šandrih, and Olivera Kitanović. 2019. "Bilingual lexical extraction based on word alignment for improving corpus search." *The Electronic Library.*

Krstev, Cvetana, Jelena Jaćimović, Branislava Šandrih, and Ranka Stanković. 2019. "Analysis of the first Serbian Literature Corpus of the Late 19th and Early 20th century with the TXM platform." In *Book of abstracts of DH_ BUDAPEST_ 2019,* 36–37.

Nielsen, Finn Årup, Daniel Mietchen, and Egon Willighagen. 2017. "Scholia, scientometrics and Wikidata." In *European Semantic Web Conference,* 237–259. Springer.

Popović, Aleksandra, Milica Ševkušić, and Đorđe Stakić. 2015. "Biblioteke i Vikipedija zajedno na vebu: slobodno znanje za sve." *Digitalna humanistika: tematski zbornik u dve knjige, knj. 1,* 151–161.

Shah, Urvi, Tim Finin, Anupam Joshi, R Scott Cost, and James Matfield. 2002. "Information retrieval on the semantic web." In *Proceedings of the eleventh international conference on Information and knowledge management,* 461–468.

Stakić, Đorđe. 2009. "Wiki Technology - Origin - Development and Importance." *INFOtheca-Journal of Informatics & Librarianship* 10 (1-2): 69–78.

Stanković, Ranka, Cvetana Krstev, Biljana Lazić, and Dalibor Vorkapić. 2015. "A bilingual digital library for academic and entrepreneurial knowledge management." In *Proceeding of 10th International Forum on Knowledge Asset Dynamics-IFKAD,* 1764–1777.

Андоновски, Јелена. 2020. "Мрежа отворених података и језички ресурси у процесу изградње српско-немачког литерарног корпуса." PhD diss., Универзитет у Београду, Филолошки факултет, јануар.