
Од (слике) текста ка графу знања: дигитализација и семантичко обогаћивање енциклопедије Свезнање

Јована Рађеновић^{1,2}[0000-0002-2707-3870],
Биљана Рујевић¹[0000-0002-9103-3902],
Александра Томашевић¹[0000-0003-1583-7447],
Оливера Китановић¹[0000-0002-7571-2729],
Ранка Станковић¹[0000-0001-5123-6273]

¹ Универзитет у Београду, Рударско-геолошки факултет, Србија,
{jovana.radjenovic|biljana.rujevic|aleksandra.tomasevic|olivera.kitanovic
|ranka.stankovic}@rgf.bg.ac.rs

² Универзитет у Београду, Математички факултет, Србија

Научни рад

DOI: 10.18485/judig.2026.2.ch5

Апстракт

Рад представља поступак дигитализације, структурног обогаћивања и изградње графа знања заснованог на историјској српској енциклопедији „Свезнање“ из 1937. године. Процес је обухватио скенирање, оптичко препознавање текста, ручну корекцију грешака, структурно уређивање и конверзију у формат TEI-XML. Посебна пажња посвећена је очувању изворног правописа и структуре текста, као и дефинисању прецизних упутстава за сараднике који су учествовали у корекцији и анотацији материјала. У раду су описани најчешћи проблеми дигитализације, укључујући мешање ћирилице и латинице, погрешно препознавање карактера и структурне грешке у тексту.

Након корекције текста примењене су методе препознавања именованих ентитета (NER), повезивања ентитета са базом знања Википодаци (NEL), као и екстракције релација између ентитета. Енциклопедијски чланци су додатно класификовани према домену и семантичком типу, чиме је омогућена изградња семантички обогаћеног графа знања. Добијени ресурс представља значајан допринос дигиталној хуманистици и развоју језичких технологија за српски језик, омогућавајући напредну семантичку претрагу, анализу културног наслеђа и интеграцију са екосистемом повезаних отворених података.

Додатно, применом великих језичких модела над препричаним и сажетим верзијама одредница генерисан је и скуп парова питања и

одговора заснован на садржају енциклопедије, намењен даљим применама у системима за одговарање на питања и приступима заснованим на генерисању потпомогнутом претрагом (RAG).

Кључне речи: српски језик, дигитализација, графови знања, обрада природног језика, препознавање именованих ентитета, повезивање именованих ентитета, генерисање питања и одговора

1. Увод

Дигитализација историјских енциклопедија представља важан корак у очувању и доступности структурираног културног знања, али су многи постојећи пројекти и даље ограничени на OCR (Optical character recognition) конверзију текста. Andersson et al. (2026) представили су поступак за реконструкцију структуре енциклопедија кроз издвајање одредница, идентификацију чланака, класификацију ентитета, повезивање различитих издања и повезивање са базом знања Википодаци (Wikidata)¹. Примењен на више издања шведске енциклопедије „Nordisk familjebok“, овај приступ је постигао високу тачност у издвајању одредница и класификацији, показујући могућност аутоматске обраде историјског енциклопедијског знања и његове интеграције у окружења повезаних података. Wu et al. (2020) предложили су оквир за изградњу графова знања из више онлајн енциклопедија, са циљем превазилажења ограничења оскудних неенглеских података на Википедији. Овај оквир комбинује модуле за екстракцију и повезивање знања, омогућавајући периодично издвајање садржаја енциклопедија и обраду нових чланака у реалном времену. Повезивање ентитета и својстава реализовано је хеуристичким и полунадгледаним методама, које су показале боље резултате од постојећих приступа и омогућиле креирање великих вишејезичких графова знања.

Циљ овог рада је креирање семантички обогаћеног графа знања заснованог на историјској српској енциклопедији „Свезнање“² применом различитих метода и техника дигитализације и обраде природног језика - следећи смернице TEI (Text Encoding Initiative) кодирања³, препознавања именованих ентитета (NER, Named Entity Recognition), екстракције релација и повезивања са

¹ <https://www.wikidata.org/>

² <https://sr.wikisource.org/wiki/>

³ <https://tei-c.org/>

базом знања (Entity Linking). Граф знања, познат и као семантичка мрежа, у најопштијем смислу представља структурирану базу података која повезује ентитете из стварног света, као што су објекти, догађаји, ситуације или концепти и илуструје везу између њих. Рад представља поступак изградње графа знања на основу енциклопедије „Свезнање“, опште енциклопедије на српском језику која је први пут објављена 1937. године у Београду (Petrović, 1937). Енциклопедија обухвата широк спектар тема из области науке, историје, географије, културе и технике, са циљем да пружи свеобухватно знање широј читалачкој публици. Објављена је у два тома на укупно 1500 страница (2720 колона). Текст је праћен илустрацијама појединих енциклопедијских чланака и штампан у две колоне са пагинираним колонама. У оквиру овог истраживања илустрације нису коришћене. Енциклопедија представља значајан културни и образовни ресурс у историји српског издаваштва и данас има велики значај за дигиталну хуманистику и истраживања графова знања. Значајан део овог енциклопедијског ресурса доступан је у дигиталном облику преко Википедије, док је додатно дигитализоване примерке обезбедила Универзитетска библиотека „Светозар Марковић“.

Поред изградње графа знања, значајан правац примене дигитализованог енциклопедијског садржаја представља и генерисање скупова питања и одговора (енгл. Question Answering — QA) намењених тренирању и евалуацији језичких модела. За српски језик и даље је доступан ограничен број оваквих ресурса. Недавно објављени резултати (Radenović et al., 2026) указују на то да комбинација великих језичких модела и људске валидације може представљати ефикасан приступ за развој QA ресурса за српски језик. У раду се приказује поступак генерисања QA скупа на основу енциклопедије „Свезнање“, чиме се исти дигитализовани ресурс ставља у функцију два комплементарна задатка обраде природног језика.

2. Дигитализација енциклопедије „Свезнање”

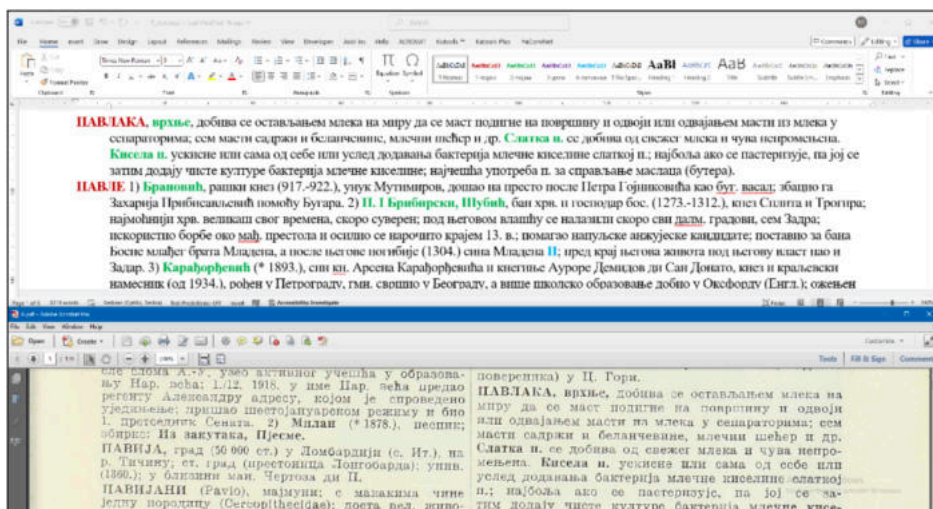
Садржај енциклопедије, који је доступан на вебу, путем Викизворника, конвертован је у структурирани TEI-XML формат. Треба напоменути да су одреднице Викизворника коришћене такве какве јесу, без поређења са оригиналним издањем Енциклопедије. Њихов квалитет није додатно провераван, имајући у виду

да је процес дигитализације који ћемо описати захтеван и дуготрајан посао који траје више од три године и још увек није окончан. Новодигитализовани материјал је прошао кроз процес оптичког препознавања карактера (OCR), које је праћено опсежним ручним исправкама које су радили сарадници на пројекту, екстерни сарадници и студенти. Полазни материјал чинило је скенирано издање енциклопедије из 1937. године, скенирано у Универзитетској библиотеци „Светозар Марковић“, након чега је извршено оптичко препознавање карактера (OCR). Иако савремени OCR системи омогућавају релативно успешно рашчитавање старих штампаних текстова, уочен је велики број грешака условљених квалитетом штампе, специфичностима предратне ортографије и мешањем писама. Осим тога проблеми су се јављали и због старог фонта, двоколонског распореда текста, оштећења папира. Како би се превазишли ови изазови осмишљен је и спроведен детаљан поступак ручне корекције и структурног уређивања текста.

Сарадницима који су учествовали у корекцији текста дата су прецизна упутства са циљем очувања аутентичности оригиналног издања и обезбеђивања конзистентности у обради. Посебно је наглашено да циљ није модернизација језика нити исправљање историјског правописа, већ креирање верне дигиталне репрезентације оригинално публикованог текста. На пример, облици као што су „отсечена“, „председник“ или „хемиски“ нису исправљани у савремене облике („одсечена“, „председник“, „хемијски“), јер представљају легитимне облике у време настанка енциклопедије. Кориговане су искључиво грешке настале током OCR процеса.

Процес корекције организован је тако да сарадници истовремено прате PDF факсимил оригиналне стране енциклопедије и Ворд документ са OCR текстом. Ради лакшег поређења и уочавања грешака, препоручено је коришћење подељеног екрана и укључивање приказа скривених карактера како би се учили вишкови размака, нежељени контролни знакови и проблеми у формирању које је потребно кориговати (Слика 1). Посебна пажња посвећена је интерпункцији, где је захтевано доследно постављање размака након интерпункцијских знакова и избегавање размака испред њих.

Један од најважнијих аспеката обраде било је структурно издвајање енциклопедијских одредница. Свака главна одредница и њен опис морали су бити смештени у засебан пасус, у скла-



Слика 1: Кориговање и обележавање текста стиливима у Ворду.

ду са структуром оригиналног издања. Додатно су дефинисани кориснички стилови за обележавање различитих типова елемената, као што су главни појмови, потпојмови и наслови књижевних или уметничких дела. Припремљени су стилови у окружењу Ворд који су јасно визуелно приказивали дефинисане сегменте текста. Главне одреднице, које су у оригиналу најчешће писане великим словима, означаване су стилем „1-појам“, док су подпојмови и секундарне одреднице добијали стил „2-појам“. Наслови књига, часописа и уметничких дела означавани су посебним стилем „3-дело“. Овакво структурно обележавање омогућило је каснију аутоматску трансформацију текста у TEI XML формат и даљу семантичку обраду.

Посебан изазов представљала је диграфска природа српског језика и мешање ћирилице и латинице током OCR процеса. Како поједина слова имају исти облик у оба писма (нпр. А, О, Р, М, Ј, К, Е, Т, Х), често је долазило до погрешног препознавања писма, што је касније представљало озбиљан проблем за претрагу и аутоматску анализу текста. Због тога су латинични сегменти текста визуелно обележавани посебним стилем дефинисаним светлоплавом бојом како би сарадници лакше учили погрешно препознате делове. На пример, у случајевима где су речи „РАХ“ или „рах“ биле погрешно рашчитане ћирилицом као „РАХ“ или

„рах“, било је неопходно извршити конверзију у исправно латинично писмо.

Поред корекције текста, сарадници су били задужени и за очување важних семантичких елемената оригинала. То је укључивало унос специјалних симбола као што су стрелице ($\rightarrow$), које означавају упућивање на сродне или синонимне појмове, као и крстић ($\dagger$), који означава годину смрти личности. Сарадницима су дати и детаљни примери најчешћих OCR грешака. Уочено је често мешање визуелно сличних карактера као што су „н, и, п“; „с, е“; „О, 0“; „З, 3“; као и проблеми у препознавању латиничних и ћириличних слова. Пошто енциклопедија садржи и латиничне стране речи, научне називе и цитате, један од честих проблема било је мешање писама у оквиру исте речи. Сарадници су добили упутства да такве случајеве исправљају и да воде рачуна о исправном писму, јер ове грешке касније значајно утичу на претраживост текста, препознавање ентитета и семантичку обраду.

Посебна пажња посвећена је и случајевима непотпуно препознатог текста. Уколико је недостајала читава одредница, сарадници су били у обавези да означе место посебним коментаром, док су мањи недостајући сегменти ручно прекуцавани из оригиналног документа. У случајевима када су недостајали мањи делови текста, они су ручно прекуцавани, док су потпуно недостајуће одреднице обележаване посебним коментаром који почиње са три звездице (***) , како би касније могле бити додатно проверене и реконструисане. На овај начин успостављен је контролисан и стандардизован поступак корекције који је омогућио добијање поузданог дигиталног текста погодног за даље фазе структурног и семантичког обогаћивања.

Овако дефинисан поступак ручне корекције и структурног обележавања обезбедио је висок степен конзистентности и квалитета дигитализованог корпуса. Прецизна упутства дата сарадницима омогућила су да се OCR текст трансформише у поуздан, структурно организован и машински читљив ресурс, погодан за даљу TEI анотацију, препознавање именованих ентитета, повезивање са Wikidata базом и изградњу графа знања.

2.1. Конвертовање из Ворда

Слика 2 представља пример некоригованог изворног OCR текста који су сарадници добијали у почетној фази. Примећује се

да енциклопедијски чланци нису правилно разграничени, имајући у виду да је очигледно да су одреднице приметне у оквиру текста чланка. Приметне су и грешке у OCR-у текста које су поменуте у претходном сегменту рада.

ПАВИЈА, град (50 000 ст.) у Ломбардији (с. Ит.), на р. Тичину; ст. град (престоница Донгобарда); унцв. (1360.); у близини ман. Чертоза ди П. ПАВИЈАНИ (*Pavio*), мајмуни; с макакима чине Једну породицу (*Cercopithecidae*); доста вел. животиње *песче* њушке; код извесних врста кожа на њушци и жуљевима за *седње* обојена *јасно* црвено. ПАВИЉОН (фрц.), шатор; кућица, продавница; хладњак, сеница у парку; споредна болничка приземна зграда; изложбена зграда неке државе; истакнут део неке зграде, купола.

ПАВИТ (*Clematis vitalba*, фам. *Ranunculaceae*), по зната шумска *повијуша* у с. и ј. *Евр.*; расте у влажним жупним шумама веома бујно, тако да отежава пролаз као какав *тропски дијан*; стабло дрвенасто, понекад знатно дебело; цвета касно у лето *беличастим* цветовима; у *Евр.* има више врста; већином повијуше (^ сл.).

ПАВИЋЕВИЋ Иван (1869.-1925.), *политичар*; *учествовао* у слободоумним покретима у Србији; после *пављанског* атентата осуђен на 20 год. робије; 1905. био мин. нар. *привр.*, затим мин. *унутр.* дела; 1920. и 1921. заузимао положај наместника (владиног повереника) у Ц. Гори.

ПАВЛАКА, *врхње*, добива *се* остављањем млека на миру да се маст подигне на површину и одвоји или одвајањем масти из млека у сепараторима; сем *масти* садржи и беланчевине, млечни шећер и др. Слатка п. се добива од свежег млека и чува непромењена. *Кисела п.* ускине или *сама* од себе или услед додавања *бактерија* млечне *киселине* *слаткој* п.; *најбоља* ако се *пастеризује*, па *јој* се затим *додају* *чисте* културе *бактерија* млечне *киселине*; најчешћа употреба п. за *справљање* маслаца (бутера).

Слика 2: Некоригован и несегментиран текст у Ворду.

Слика 3 илуструје изглед дела текста након корекције текста и обележавања предефинисаним стиловима од стране сарадника. Грешке у OCR-у су кориговане. Види се да су енциклопедијски чланци разграничени, одреднице означене црвеном бојом, потпојмови и секундарне одреднице означене зеленом бојом, називи дела љубичастом а латинични део текста плавом бојом.

ПАВИЈА, град (50 000 ст.) у Ломбардији (с. Ит.), на р. Тичину; ст. град (престоница Донгобарда); унцв. (1360.); у близини ман. Чертоза ди П.

ПАВИЈАНИ (*Pavio*), мајмуни; с макакима чине једну породицу (*Cercopithecidae*); доста вел. животиње *песче* њушке; код извесних врста кожа на њушци и жуљевима за *седње* обојена *јасно* црвено.

ПАВИЉОН (фрц.), шатор; кућица, продавница; хладњак, сеница у парку; споредна болничка приземна зграда; изложбена зграда неке државе; истакнут део неке зграде, купола.

ПАВИТ (*Clematis vitalba*, фам. *Ranunculaceae*), позната шумска повијуша у с. и ј. *Евр.*; расте у влажним жупним шумама веома бујно, тако да отежава пролаз као какав *тропски дијан*; стабло дрвенасто, понекад знатно дебело; цвета касно у лето *беличастим* цветовима; у *Евр.* има више врста; већином повијуше (→ сл.).

ПАВИЋЕВИЋ Иван (1869.-1925.), политичар; учествовао у слободоумним покретима у Србији; после иванџанског атентата осуђен на 20 год. робије; 1905. био мин. нар. *привр.*, затим мин. *унутр.* дела; 1920. и 1921. заузимао положај наместника (владиног повереника) у Ц. Гори.

ПАВЛАКА, *врхње*, добива се остављањем млека на миру да се маст подигне на површину и одвоји или одвајањем масти из млека у сепараторима; сем *масти* садржи и беланчевине, млечни шећер и др. *Слатка п.* се добива од свежег млека и чува непромењена. *Кисела п.* ускине или *сама* од себе или услед додавања *бактерија* млечне *киселине* *слаткој* п.; *најбоља* ако се *пастеризује*, па *јој* се затим *додају* *чисте* културе *бактерија* млечне *киселине*; најчешћа употреба п. за *справљање* маслаца (бутера).

Слика 3: Коригован и стиловима обележен текст у Ворду.

Новодигитализовани и ручно кориговани део тренутно обухвата око 6.000 одредница. Очекује се да ће преостали део бити завршен до краја године.

2.2. Конвертовање из Викизворника

Део преузет са Викизворника тренутно садржи 10.162 одреднице. За даљу обраду, енциклопедијски чланци подељени су на мање целине на основу пасуса и одредница, како би се очувала контекстуална повезаност и унапредила каснија семантичка обрада. Слика 4 представља изглед приказа енциклопедијских чланака у Викизворнику. Примећује се да су у тексту означене одреднице, секундарне одреднице и наслови дела и да је целокупан текст представљен у латиничном писму. Ово је опција која се бира на Викизворнику – могући су ћирилични, латинични и мешовити приказ писма.

ДИЈАБАЗ (грч.), површинска магматска стена офитске структуре, састављена од плагиокласа и пироксена, поред којих се јављају покашто и хорнбленда, магнетит, апатит и др.; распадањем добива зеленкасту боју; стога се убраја у групу тзв. зелених стена: код нас доста распрострањен у друштву серпентина и др. зелених стена, нарочито у Ј. Србији и Босни.

ДИЈАБЕЛИ АНТОНИО (Diabelli, 1781.-1858.), композитор, проф. клавира и члан издавачке радње у Бечу; компоновао: клавирску и црк. муз., песме, оперету,

ДИЈАБЕТЕС 1) *diabetes insipidus*, обољење проузроковано поремећајем унутарње секреције; знаци: вел. лучење мокраће, некад по неколико л. дневно; мокраћа не мора, али може имати шећера, беланчевине, или фосфата, оксалата и уреје у већој количини; лечи се екстрактима жлезда с унутарњом секрецијом. 2) *d. melitus*, шећерна болест, обољење нарочито карактеристично повећаном количином шећера у крви и излученим шећером (кроз бубреге) у мокраћи; изазвано поремећајем и недостатком лучења инсулина (хормон панкреаса) услед обољења панкреаса (специјално Лангерхосовог острва); инсулин регулише промет шећера у организму и спречава његово лучење у мокраћу; д. м. је конституционално обољење, најчешће наследно; знаци: осећај велике жеђи и глади, излучивање вел. количине мокраће, која садржи шећер (гликозурија), повећање шећера у крви (хипергликемија), свраб коже, мршављење и општа малаксалост; компликација → кома дијабетична.

DIJABAZ (грч.), površinska magmatska stena ofitske strukture, sastavljena od plagioklasa i piroksena, pored kojih se javljaju pokašto i hornblenda, magnetit, apatit i dr.; raspadanjem dobiva zelenkastu boju; stoga se ubraja u grupu tzv. zelenih stena: kod nas dosta rasprostranjen u društvu serpentina i dr. zelenih stena, naročito u J. Srbiji i Bosni.

DIJABELI ANTONIO (Diabelli, 1781.-1858.), kompozitor, prof. klavira i član izdavačke radnje u Beču; komponovao: klavirsku i crk. muz., pesme, operetu,

DIJABETES 1) *diabetes insipidus*, oboljenje prouzrokovano poremećajem unutarne sekrecije; znaci: vel. lučenje mokraće, nekad po nekoliko l. dnevno; mokraća ne mora, ali može imati šećera, belančevine, ili fosfata, oksalata i uree u većoj količini; leči se ekstraktima žlezda s unutarjom sekrecijom. 2) *d. melitus*, šećerna bolest, oboljenje naročito karakteristično povećanom količinom šećera u krvi i izlučenim šećerom (kroz bubrege) u mokraći; izazvano poremećajem i nedostatkom lučenja insulina (hormon pankreasa) usled oboljenja pankreasa (specijalno Langerhosovog ostrva); insulin reguliše promet šećera u organizmu i sprečava njegovo lučenje u mokraću; d. m. je konstitucionalno oboljenje, najčešće nasledno; znaci: osećaj velike žeđi i gladi, izlučivanje vel. količine mokraće, koja sadrži šećer (glikozurija), povećanje šećera u krvi (hiperglikemija), svrab kože, mršavljenje i opšta malaksalost; komplikacija → кома дијабетична.

Слика 4: Пример приказа енциклопедијских чланака у Викизворнику.

Као пример изгледа кода енциклопедијских чланака у Викизворнику може се навести следећи исечак кода:

ДИЈАБАЗ (грч.), површинска магматска стена офитске структуре, састављена од плагиокласа и пироксена, поред којих се јављају покашито и хорнбленда, магнетит, апатит и др.; распаѓањем добива зеленкасту боју; стога се убраја у групу тзв. зелених стена: код нас доста распрострањен у друштву серпентина и др. зелених стена, нарочито у Ј. Србији и Босни.

ДИЈАБЕЛИ АНТОНИО (Diabelli, 1781.-1858.), композитор, проф. клавира и члан издавачке радње у Бечу; компоновао: клавирску и црк. муз., песме, оперету,

ДИЈАБЕТЕС 1) **diabetes insipidus**, обољење проузроковано поремећајем унутарње секреције; знаци: вел. лучење мокраће, некад по неколико л. дневно; мокраћа не мора, али може имати шећера, беланчевине, или фосфата, оксалата и урее у већој количини; лечи се екстрактима жлезда с унутарњом секрецијом. 2) **d. mellitus**, шећерна болест, обољење нарочито карактеристично повећаном количином шећера у крви и излученим шећером (кроз бубреге) у мокраћи; изазвано поремећајем и недостатком лучења инсулина (хормон панкреаса) услед обољења панкреаса (специјално Лангерхансовог острва); инсулин регулише промет шећера у организму и спречава

Садржај добијен на овај начин је затим структуриран применом XML анотације инспирисане TEI стандардом, укључујући ознаке за одреднице, наслове и важне семантичке јединице. Овако структурирана репрезентација омогућила је стандардизовану претходну обраду и олакшала касније NLP задатке као што су издвајање ентитета, семантичка класификација и изградња графа знања.

3. Обогаћивање текста семантичким информацијама

Осим верне дигиталне копије оригиналног документа енциклопедије „Свезнање“, делимично је креиран и облик који је структурно обогаћен и нормализован. Процес обраде обухватио је идентификацију и анотацију енциклопедијских одредница, разрешење скраћеница, структурирање метаподатака и семантичко обогаћивање, чиме је сирови дигитализовани текст трансформиран у машински читљив и интероперабилан ресурс погодан за окружење повезаних података (linked data) и изградњу графова знања. Радни ток укључивао је структурну анотацију енциклопедијских чланака, делимично аутоматско разрешење скраћеница, препознавање именованих ентитета (NER), као и повезивање ентитета са базом Википодаци. За семантичку анотацију посебно важан корак представља NER, који се ослања на моделе развије-

не у оквиру пројекта TESLA (Text Embeddings – Serbian Language Applications), способне да препознају седам типова ентитета: личности, организације, локације, професије, догађаје, демониме и уметничка дела (Ikonić Nešić et al., 2026). Препознати ентитети су затим повезани са спољним базама знања, пре свега са Википодацима, чиме је омогућена семантичка нормализација и успостављање односа између ентитета. Током претходне обраде уклоњени су сувишни и нерелевантни делови текста, док су дубликати ентитета обједињени ради смањења недоследности и побољшања квалитета података.

4. Екстракција релација

Посебно важан део представља екстракција релација, која прати приступ успостављен у раду (Stanković et al., 2026), где је екстракција релација примењена над скупом TeslaNER+, великим корпусом од преко 150.000 реченица аотираних именованим ентитетима и повезаних са базом Википодаци. Деветнаест семантичких релација из више домена, укључујући и књижевне текстове, издвојено је применом правила заснованих на коначним аутоматима (FST, finite state transducers) и приступима заснованим на великим језичким моделима (Large language model prompting).

Скуп релација је унапред дефинисан, а могуће комбинације субјекат–објекат ограничене су типовима ентитета. У овом истраживању коришћене су следеће релације: између личности (colleagueOf, managerOf, parentOf, relativeOf, siblingOf, marriedTo, spouseOf); између личности и организација (educatedAt, managerOfOrg, affiliatedWith); између личности и локација (residesIn, bornAt, diedAt); између личности и дела (authorOf); између организација (partOfOrg); између организација и локација (basedIn); између локација (partOfLoc); као и релације које повезују догађаје са личностима (eventPers) и локацијама (eventLoc).

Експериментални резултати на 23.000 реченица показују да FST приступ постиже већу прецизност (0.77), али мањи одзив (0.39), док LLM приступ обезбеђује веома висок одзив (0.97) по цену ниже прецизности (0.65), већих рачунарских трошкова и спорије обраде.

5. Граф знања

Предложени граф знања обједињује текстуалне, структурне и семантичке информације издвојене из енциклопедијских чланака. Поред повезивања ентитета, чланци су аутоматски класификовани према домену и семантичком типу, укључујући категорије као што су биографије, историјски догађаји, локације, организације, појмови и културни артефакти. Добијени граф омогућава напредну семантичку претрагу, истраживање односа између ентитета и интероперабилност са спољним базама знања. Добијени ресурс представља значајан допринос српској дигиталној хуманистици и језичким технологијама, обезбеђујући поново употребљиву инфраструктуру за семантичку претрагу, анализу корпуса и апликације засноване на знању.

5.1. Класификација енциклопедијских одредница

У циљу семантичког обогаћивања енциклопедије и изградње графа знања, енциклопедијске одреднице класификоване су према свом домену и семантичком типу. Оваква класификација омогућава лакшу организацију знања, ефикасније повезивање ентитета са спољним базама знања, као и примену различитих метода аутоматске обраде природног језика, укључујући препознавање именованих ентитета (NER), повезивање ентитета (NEL), екстракцију релација и семантичку претрагу.

Класификација је организована хијерархијски и обухвата следеће основне категорије:

- **Личности** – историјске личности, научници, уметници, писци, политичари, владари, спортисти, религијске личности и др.
- **Локације** – државе, градови, регије, планине, реке, мора, археолошки локалитети и други географски појмови.
- **Организације** – институције, универзитети, библиотеке, музеји, компаније, удружења, политичке и верске организације.
- **Догађаји** – историјски догађаји, ратови, битке, револуције, конференције, културне манифестације и природне катастрофе.
- **Дела** – књижевна, уметничка и музичка дела, часописи, новине и друге публикације.

- **Научни и стручни појмови** – појмови из области математике, физике, хемије, биологије, медицине, права, филозофије и лингвистике.
- **Биолошки ентитети** – биљке, животиње, микроорганизми и анатомски појмови.
- **Хемијски ентитети** – хемијски елементи, једињења, минерали и хемијске реакције.
- **Технички и материјални објекти** – машине, уређаји, возила, оружје, архитектонски и индустријски објекти.
- **Религијски и митолошки појмови** – божанства, светитељи, митолошка бића, религијски ритуали и појмови.
- **Језички и књижевни појмови** – језици, писма, граматички појмови, стилске фигуре и књижевни правци.
- **Друштвени појмови** – занимања, титуле, идеологије, етничке и друштвене групе.

Овако дефинисан скуп категорија омогућава семантичко структурирање енциклопедијског садржаја и представља основу за изградњу графа знања и повезивање са ресурсима као што су Википодаци и екосистем повезаних отворених података (Linked Open Data). Класификација је такође погодна за примену у задацима аутоматске класификације текстова, препознавања ентитета, семантичке претраге и генерисања питања и одговора.

5.2. Разрешење скраћеница и њихова употреба у класификацији

Један од важних корака у обради енциклопедије „Свезнање“ представљало је разрешење великог броја скраћеница карактеристичних за енциклопедијски стил писања прве половине XX века.⁴ Скраћенице су коришћене за означавање језика, географских појмова, професија, научних дисциплина, титула, институција, историјских периода и других категорија. Њихово аутоматско и полуаутоматско разрешење било је неопходно ради побољшања читљивости текста, унапређења препознавања именованих ентитета и тачније семантичке класификације енциклопедијских одредница.

У процесу нормализације коришћен је унапред дефинисан речник скраћеница, који је обухватао најчешће облике присутне

⁴ <https://sr.wikisource.org/wiki/Sveznanje/Skraćenja>

у енциклопедији. Посебно значајну групу чиниле су језичке скраћенице, као што су: *ит.* (италијански), *фрц.* (француски), *енгл.* (енглески), *нем.* (немачки), *грч.* (грчки), *лат.* (латински), *тур.* (турски) и друге. Ове скраћенице коришћене су као важан индикатор приликом класификације појмова који се односе на стране језике, етимологију или културно-историјски контекст.

Посебну категорију чиниле су скраћенице за географске и административне појмове: *Евр.* (Европа), *Ит.* (Италија), *Рус.* (Русија), *Ц. Гора* (Црна Гора), *Б. П.* (Балканско полуострво), *Јадр. М.* (Јадранско море), као и ознаке страна света: *с.* (северни), *ј.* (јужни), *и.* (источни), *з.* (западни). Оне су коришћене као додатни сигнал за препознавање и класификацију географских ентитета и локација.

Важну улогу имале су и скраћенице које означавају професије, титуле и институције: *д-р* (доктор), *проф.* (професор), *еп.* (епископ), *архиеп.* (архиепископ), *унив.* (универзитет), *акад.* (академија), *гмн.* (гимназија). Оне су коришћене за идентификацију личности, организација и институција, али и за екстракцију релација између ентитета.

У оквиру научних и стручних појмова разрешење скраћеница било је важно за тематску класификацију чланака: *физ.* (физички), *биол.* (биолошки), *хем.* (хемијски), *теол.* (теолошки), *геогр.* (географски), *архт.* (архитектонски), *књиж.* (књижевни), *екон.* (економски). На основу ових ознака било је могуће аутоматски одредити домен енциклопедијске одреднице.

Посебно су значајне биле симболичке ознаке: * за годину рођења и † за годину смрти, које су коришћене као индикатори биографских чланака и временских информација.

Разрешење скраћеница примењено је током претходне обраде текста, а добијене нормализоване форме коришћене су у даљим задацима семантичке анотације, препознавања именованих ентитета, повезивања са ресурсима Википодатака и изградње графа знања. Овај поступак значајно је допринео побољшању квалитета аутоматске класификације и семантичке интерпретације енциклопедијског садржаја.

6. Генерисање новог садржаја на основу чланака

За генерисање новог садржаја коришћени су претходно дигитализовани и структурирани чланци енциклопедије, припре-

мљени поступком описаним у претходном делу рада. Чланци су били организовани у фајлове у Ворд формату, у којима су коришћена четири прилагођена стила за означавање сегмената текста унутар сваке одреднице. Развијена процедура их чита директно из Ворда, конвертује у XML и пресликава их у семантичке категорије. Стили и семантичке категорије приказани су у табели 1.

Стил	Семантичка улога
1- <i>pojam</i>	Одредница — главни термин
2- <i>pojam</i>	Појам — варијанта или поткатегорија
3- <i>delo</i>	Дело — наслов књиге, слике итд.
4- <i>latinica</i>	Латиница — стране и лат. речи

Табела 1: Прилагођени стилови и њихове семантичке улоге.

Први Gemini промт налаже моделу да преформулише одредницу у потпуне течне реченице развијајући све скраћенице. У први промт је уграђена референтна таблица са приближно 70 скраћеница (нпр. *с.* → *северни*, *ист.* → *историјски*, * *1840.* → *рођен 1840. године*, *Евр.* → *Европа*, *полит.* → *политички*).

Треба напоменути да разрешавање није сасвим једнозначно јер се неретко на саму одредницу реферише скраћеницом која се састоји од почетног слова и тачке. На слици 5 се може видети да се у тексту уместо *Стара Павлица* помиње *Стара П.*.

Други промт налаже моделу да напише сажетак отприлике упола краћи од оригинала, задржавајући најважније информације — идентитет, датуме, место и кључна дела или доприносе. За кратке одреднице сажетак се своди на једну реченицу; за дугачке одреднице са вишеструким поткатегоријама модел даје структуриран извод са најзначајнијим поткатегоријама.

ПАВЛИЦА, **Стара и Нова**, ман. на Ибру код Рашке. **Стара П.**, грч. црква из 11. в. с 1 кубетом, у рушевинама. **Нова П.**, задужбина Стевана и Вука, синова челника Мусе, из доба кн. Лазара; крстообразна црква с певничким конхама; била ж. манастир; постоји и данас; веома важан њен живопис.

ПИЗА, град (78 000 ст.) у Ит. на р. Арну недалеко (11 км) од њена ушћа у Сред. М.; у П. развијена инд. памука, има унив.; познати криви торањ у П. шисок 54 м (→ сл.).

Слика 5: Пример учитаних и обележених енциклопедијских чланака.

Слика 6 приказује пример аутоматске трансформације историјског енциклопедијског чланка из сажетог лексикограф-

ског облика у више семантички обогачених текстуалних репрезентација. Полазни енциклопедијски запис садржи велики број скраћеница, згуснуту синтаксу и карактеристичан стил старих енциклопедија, што може представљати изазов како за савремене читаоце, тако и за системе за аутоматску обраду природног језика.

ПАВЛИЦА
<p>ORIGINAL</p> <p>ПАВЛИЦА, Стара и Нова, ман. на Ибру код Рашке. Стара П., грч. црква из 11. в. с 1 кубетом, у рушевинама. Нова П., задужбина Стевана и Вука, синова челника Мусе, из доба кн. Лазара; крстообразна црква с певничким конхама; била ж. манастир; постоји и данас; веома важан њен живопис.</p>
<p>ПРЕПРИЧАВАЊЕ (разрешене скраћенице)</p> <p>Павлица је име за два манастира, Стару и Нову Павлицу, који се налазе на реци Ибар, близу града Рашке. Стара Павлица је грчка црква која потиче из једанаестог века. Имала је једно кубе, а данас су од ње остале само рушевине. Нова Павлица је задужбина Стевана и Вука, који су били синови челника Мусе. Изграђена је у време кнеза Лазара. То је црква крстообразног облика, са певничким конхама, а некада је била женски манастир. Овај манастир постоји и данас, а његов живопис, односно фреске, веома су важне.</p>
<p>САЖЕТАК</p> <p>Павлица, манастир на Ибру код Рашке, чине Стара (грч. црква из 11. в., у рушевинама) и Нова (задужбина Мусиних синова из доба кн. Лазара, ж. манастир са важним живописом).</p>
ПИЗА
<p>ORIGINAL</p> <p>ПИЗА, град (78 000 ст.) у Ит. на р. Арно недалеко (11 км) од њена ушћа у Сред. М.; у П. развијена инд. памука, има унив.; познати криви торањ у П. висок 54 м (→ сл.).</p>
<p>ПРЕПРИЧАВАЊЕ (разрешене скраћенице)</p> <p>ПИЗА је град у Италији који броји 78 000 становника. Налази се на реци Арно, на само 11 километара од њеног ушћа у Средоземно море. У Пизи је развијена индустрија памука, а град такође има и универзитет. Познати Криви торањ у Пизи висок је 54 метра.</p>
<p>САЖЕТАК</p> <p>ПИЗА, град у Италији, познат по свом универзитету и развијеној индустрији памука. Чувени Криви торањ у Пизи висок је 54 метра.</p>

Слика 6: Пример разрешавања скраћеница и сажимање енциклопедијских чланака.

У првој фази систем врши разрешење скраћеница и контекстуално препричавање текста, чиме се оригинални чланак претвара у експлицитнији и читљивији наративни облик уз очување основних фактографских информација. У другој фази генерише се сажета модернизована верзија текста, која представља компактну репрезентацију кључних биографских и библиографских података. Приказани примери показују да систем успешно

идентификује и реконструише важне ентитете, односе и контекстуалне информације, као што су географски положај, историјски период, врста локалитета, његов културно-историјски значај и друге релевантне карактеристике. На тај начин енциклопедијски текстови постају погоднији за даљу семантичку анализу, претрагу, изградњу графова знања и генерисање питања и одговора у оквиру система заснованих на великим језичким моделима и RAG (Retrieval-Augmented Generation) архитектурама.

На слици су приказани примери чланака *Павлица* и *Пица* који описују културно-историјске локалитете, при чему у процесу препричавања систем успешно врши разрешење скраћеница, реконструкцију синтаксичких односа и експлицитно увођење контекста, чиме текст постаје приступачнији и погоднији за аутоматску семантичку анализу.

7. Генерисање питања и одговора

Широк тематски обухват енциклопедије „Свезнање“, која доноси сажете информације о појмовима, догађајима, личностима и природним феноменима из готово свих области знања тога времена, чини је погодним извором и за генерисање QA скупова података намењених задацима општег знања.

Обрађена енциклопедија представља добру основу за развој скупа података TESLA-Sveznanje-QA⁵. Скуп је генерисан применом техника промпт инжењеринга, при чему велики језички модел (LLM) добија енциклопедијски чланак као контекст и на основу њега генерише парове питања и одговора (Stanković et al., 2025). Поред прва два промпта, описана у претходном одељку, који се користе за разрешавање скраћеница и препричавање, односно сажимање садржаја одреднице, развијен је и трећи Gemini промпт намењен генерисању питања и одговора на основу тако обрађене верзије чланка. За сваку одредницу од модела се захтева да генерише тачно три питања и одговарајуће одговоре, а резултат врати у формату JSON. Додатним упутством се захтева да враћена JSON структура буде валидна и да не садржи пратећи текст или ознаке кода, чиме је омогућено непосредно аутоматско рашчлањивање резултата применом функције `json.loads()`.

⁵ https://huggingface.co/datasets/te-sla/QuestionAnswering/blob/main/SerbianQA-Gen/Sveznanje/train-00000-of-00001.parquet?sql_row=54

Процес генерисања QA парова аутоматизован је тако да се обрађене енциклопедијске одреднице читају директно из Ворд документа, при чему се формат .docx обрађује применом библиотеке python-docx. Као и у претходној фази обраде, оригинални енциклопедијски чланак се најпре трансформише тако да се разреши бројне скраћенице присутне у тексту. Претходна анализа је показала да су питања и одговори генерисани на основу тако обрађених чланака квалитетнији од оних генерисаних на основу оригиналних чланака.

У процесу генерисања, садржај се припрема примарно у латиничном писму, при чему се током обраде врши аутоматско пресловљавање из ћирилице у латиницу. По потреби, добијени скуп се може једноставно и брзо пресловити у ћирилично писмо.

Ако погледамо одредницу ПЕЋИ, чији оригинални текст у документу гласи:

ПЕЋИ, служе за произвођење топлоте, нпр. за грејање (**собне п.**), за кување (**штедњаци**), за топљење руда и метала (**металуршке п.**), за обраду метала (**п. за каљење**), за печење робе (**цигларске п.**); има их с наизменичним и непрекидним радом, према томе да ли се пре поновне употребе мора чекати да се охладе или то није потребно (као нпр. код прстенастих цигларских и. тзв. »рингофен«). **П. са муфнама**, пламена п. са 1 или више → муфни. (На сл.: загревање п.)

на основу препричане верзије без скраћеница модел генерише репрезентативна питања и одговоре:

1. **За шта све могу да служе пећи према њиховој основној намени?** — *Пећи служе за произвођење топлоте која се користи за грејање, кување, топљење и обраду метала, као и за печење различите робе попут цигле.*
2. **Како се пећи деле с обзиром на начин њиховог рада?** — *Пећи се деле на оне са наизменичним радом, код којих се мора чекати да се охладе пре поновне употребе, и оне са непрекидним радом где то није потребно.*
3. **Шта је карактеристично за прстенасте цигларске пећи познате као »рингофен«?** — *Оне спадају у пећи са непрекидним радом, што значи да се не мора чекати да се охладе пре поновне употребе.*

Одреднице се обрађују редом, уз паузу од једне секунде између узастопних позива API-ју ради избегавања ограничења учесталости захтева. За сваку одредницу резултати обраде које чине оригинални текст, препричана верзија, сажетак и генерисана питања и одговори, истовремено се уписују у структурирани `.json` фајл, погодан за даљу машинску обраду, и у читљив `.txt` фајл са јасним раздвајањем одредница. На овај начин добијен је скуп TESLA-Sveznanje-QA.

Слика 7 приказује крајњи резултат процеса, односно пример енциклопедијског чланка са препричаном и сажетом верзијом, за којима следе питања и одговори које је модел генерисао.

На приказаном примеру се може видети на који начин је систем извршио разрешавање скраћеница и контекстуално препричавање садржаја, као и да се при томе из кратког енциклопедијског записа генерише експлицитнији и семантички јаснији текст уз очување изворних информација. Види се и додатно сажет и модернизован приказ чланка, чиме се демонстрира могућност креирања више нивоа текстуалне репрезентације над истим изворним садржајем. На крају, у доњем делу слике приказана су аутоматски генерисана питања и одговори изведени из проширеног текста. Ови парови питања и одговора издвајају кључне информације везане за саму одредницу, показујући како се енциклопедијски садржај може трансформисати у структурирано знање погодно за семантичку претрагу, образовне системе, дигиталну хуманистику и RAG системе.

Овако формиран скуп података представља вредан ресурс не само за развој и евалуацију система за одговарање на питања, већ и за фино подешавање (енгл. *fine-tuning*) мањих, ефикаснијих језичких модела специјализованих за српски језик. Овакви модели могли би да се покрећу локално, без ослањања на екстерне комерцијалне сервисе, чиме би се смањили трошкови и повећала независност система за обраду природног језика заснованих на овом ресурсу.

8. Закључак

У раду је приказан процес трансформације ресурса из међуратне енциклопедије у граф знања. Први део рада посвећен је опису процеса ретроцифрелизације енциклопедије „Свезнање“, у смислу корекције текста у погледу грешака насталих у процесу

ПЕЋИ
<p>ORIGINAL</p> <p>ПЕЋИ, служе за производње топлоте, нпр. за грејање (собне п.), за кување (штедњаки), за топљење руда и метала (металуршке п.), за обраду метала (п. за каљење), за печење робе (цигларске п.); има их с наизменичним и непрекидним радом, према томе да ли се пре поновне употребе мора чекати да се охладе или то није потребно (као нпр. код прстенастих цигларских и. тзв. »рингофен«). П. са муфнама, пламена п. са 1 или више → муфни. (На сл.: загревање п.)</p>
<p>ПРЕПРИЧАВАЊЕ (разрешене скраћенице)</p> <p>Пећи су уређаји који служе за производњу топлоте. Користе се за различите намене, на пример за грејање, као што су собне пећи, за кување, попут штедњака, за топљење руда и метала, што су металуршке пећи, за обраду метала, као што су пећи за каљење, затим за печење робе, попут цигларских пећи. Према начину рада, пећи могу имати наизменични или непрекидни рад, у зависности од тога да ли се пре поновне употребе мора чекати да се охладе или то није потребно, као што је случај код прстенастих цигларских пећи које се још називају и такозвани рингофен. Постоје и пећи са муфнама, а то су пламене пећи које имају једну или више муфни. На слици је приказано загревање пећи.</p>
<p>САЖЕТАК</p> <p>Пећи су уређаји за производњу топлоте који се користе за грејање, кување, топљење и обраду метала, као и за печење материјала попут цигле. Деле се према начину рада на оне са наизменичним и непрекидним радом, а могу бити и специјализоване, попут пламених пећи са муфнама.</p>
<p>ПИТАЊА И ОДГОВОРИ</p> <p>R1: За шта све могу да служе пећи према њиховој основној намени? О: Пећи служе за производње топлоте која се користи за грејање, кување, топљење и обраду метала, као и за печење различите робе попут цигле.</p> <p>R2: Како се пећи деле с обзиром на начин њиховог рада? О: Пећи се деле на оне са наизменичним радом, код којих се мора чекати да се охладе пре поновне употребе, и оне са непрекидним радом где то није потребно.</p> <p>R3: Шта је карактеристично за прстенасте цигларске пећи познате као »рингофен«? О: Оне спадају у пећи са непрекидним радом, што значи да се не мора чекати да се охладе пре поновне употребе.</p>

Слика 7: Пример енциклопедијског чланка са препричаном и сажетом верзијом, за којима следе питања и одговори.

OCR-а и обележавања делова у структури енциклопедијског чланка. Након структурне анотације и делимичног разрешавања скраћеница, извршено је препознавања именованих ентитета (NER) и повезивање ентитета са базом знања Википодаци (NEL). Потом је извршена и екстракција релација између ентитета и генерисан је ресурс који представља скуп парова питања и одговора заснованих на енциклопедијским чланцима (TESLA-Sveznanje-QA).

Имајући у виду да је за систематичну припрему материјала неопходно доста људских и временских ресурса, очекује се да

ће до краја године бити завршен рад на свим чланцима енциклопедије „Свезнање“ и да ће граф у том смислу бити комплетан.

Граф знања произведен на приказаној методологији и заснован на енциклопедији објединио је текстуалне, структурне и семантичке информације на основу домена и семантичког типа што омогућава напредну претрагу и истраживање односа између ентитета. Истовремено, ресурс TESLA-Sveznanje-QA представља додатни резултат обраде енциклопедијског садржаја и основу за развој система за одговарање на питања и других језичких технологија. Добијени ресурси представљају значајан допринос српској дигиталној хуманистици и развоју језичких технологија за српски језик, као и могућностима анализе културног наслеђа и интеграције са екосистемом повезаних отворених података.

Рад отвара и неколико праваца за даља истраживања. Мањи узорак генерисаних питања и одговора ручно је прегледан током развоја поступка, при чему су резултати оцењени као задовољавајући. Ипак, за поуздану оцену квалитета на нивоу целог скупа потребно је систематско вредновање генерисаног скупа TESLA-Sveznanje-QA од стране лингвиста, како би се проценила тачност и релевантност аутоматски генерисаних питања и одговора, а не само њихова синтаксичка исправност. Осим тога, паралелно са довршавањем дигитализације и корекције преосталих одредница, планира се проширење поступка генерисања питања и одговора на нове чланке, чиме би скуп TESLA-Sveznanje-QA временом обухватио комплетан садржај енциклопедије Свезнање. Коначно, планира се проширење екстракције релација и на преостале, тренутно необрађене делове енциклопедије, као и повезивање добијеног графа знања са другим постојећим ресурсима за српски језик у оквиру екосистема повезаних отворених података.

Захвалница

Ово истраживање подржао је Фонд за науку Републике Србије, пројекат бр. 7276, Text Embeddings – Serbian Language Applications (TESLA).

Литература

- Albin Andersson, Salam Jonasson, Fredrik Wastring, and Pierre Nugues. 2026. [ATLAS: Article Tracking, Linking, and Analysis of Swedish Encyclopedias](#). In *Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026)*, pages 3566–3576. ELRA.
- Milica Ikonić Nešić, Saša Petalinkar, Olivera Kitanović, Ranka Stanković, and Miloš Utvić. 2026. [CNN-based Named Entity Linking: Serbian Use Case](#). *Poznan Studies in Contemporary Linguistics*.
- Petar Petrović. 1937. *Свезнање: општи енциклопедијски лексикон*. Народно дело, Београд.
- Jovana Rađenović, Olivera Kitanović, Ranka Stanković, and Mihailo Škorić. 2026. [Development of Serbian QA Datasets through Prompt-Based Generation and Human Validation](#). In *Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026)*, pages 3369–3378, Palma, Mallorca, Spain. European Language Resources Association (ELRA).
- Ranka Stanković, Cvetana Krstev, Milica Ikonić Nešić, and Miloš Utvić. 2026. [Relation Extraction from Text of Various Domains: A Serbian Case Study](#). In *The Seventh International Conference Computational Linguistics in Bulgaria (CLIB 2026)*, pages 258–267.
- Ranka Stanković, Jovana Rađenović, Maja Ristić, and Dragan Stankov. 2025. [Creation of a Training Dataset for Question-Answering Models in Serbian](#). In Jasmina Moskovljević Popović and Ranka Stanković, editors, *Proceedings of the International Conference South Slavic Languages in the Digital Environment JuDig : Thematic Collection of Papers*, volume 1, chapter 14, pages 245–263. University of Belgrade — Faculty of Philology, Belgrade. 14.
- Tianxing Wu, Haofen Wang, Cheng Li, Guilin Qi, Xing Niu, Meng Wang, Lin Li, and Chaomin Shi. 2020. [Knowledge Graph Construction from Multiple Online Encyclopedias](#). *World Wide Web*, 23(5):2671–2698.

From Text (Image) to a Knowledge Graph: Digitization and Semantic Enrichment of the Sveznanje Encyclopedia

Jovana Rađenović, Biljana Rujević, Aleksandra Tomašević, Olivera Kitanović, Ranka Stanković

Summary

This paper presents the process of digitization, structural enrichment, and knowledge graph construction based on the historical Serbian encyclopedia „Свезнање“, first published in 1937. The workflow included scanning, OCR-based text recognition, manual error correction, structural editing, and conversion into the TEI-XML format. Special attention was given to preserving the original orthography and document structure, as well as to defining precise annotation and correction guidelines for collaborators involved in the project. The paper discusses the most common OCR-related issues, including script mixing between Cyrillic and Latin alphabets, character recognition errors, and structural inconsistencies in the text.

Following text correction, Named Entity Recognition (NER), Named Entity Linking (NEL) to Wikidata, and relation extraction techniques were applied. Encyclopedia entries were additionally classified according to their domain and semantic type, enabling the construction of a semantically enriched knowledge graph. The resulting resource represents an important contribution to digital humanities and Serbian language technologies, providing support for advanced semantic search, cultural heritage analysis, and integration into the Linked Open Data ecosystem.

In addition, large language models were used to generate a complementary set of question-answer pairs from the encyclopedia entries, intended for further use in question-answering and retrieval-augmented generation (RAG) systems.

Keywords: Serbian language, digitization, knowledge graphs, natural language processing, named entity recognition, named entity linking, question-answer generation

Works are licensed under the terms and conditions of the Creative Commons Attribution 4.0 International License.

<https://creativecommons.org/licenses/by/4.0/>

