

Александар Д. Вукајловић
Градска библиотека
„Владислав Петковић Дис”, Чачак
acovukajlovic253@gmail.com

Прегледни рад
UDK 026.06:004.42RESCARTA TOOLKIT
027.022:004.4(497.11)
004.382.745
<https://doi.org/10.18485/bibliotekar.2021.63.1.1>

Милан Д. Јовановић
Народна библиотека Пирот
mixaptt@yahoo.com

Андрија Д. Сагић
Библиотека „Милутин Бојић”, Београд
andrija.sagic@milutinbojic.org.rs

СКУП АЛАТА ЗА КРЕИРАЊЕ ДИГИТАЛНЕ БИБЛИОТЕКЕ – RESCARTA TOOLKIT

Сажетак: У овом раду биће представљен скуп алата неопходних за формирање функционалне дигиталне библиотеке које се користе у четири библиотеке у Републици Србији: Библиотека града Београда, Градска библиотека „Владислав Петковић Дис” у Чачку, Библиотека „Милутин Бојић” у Београду и Народна библиотека Пирот. Функционална дигитална библиотека садржи стандардизоване метаподатке, функцију претраге у тексту, адекватан веб-интерфејс за брзу и интуитивну употребу и могућност филтрирања дигиталних објеката према метаподацима. Скуп алата „ResCarta Toolkit” садржи све неопходне компоненте за брзо и ефикасно формирање функционалне дигиталне библиотеке, почевши од уноса метаподатака до веб-апликације за приказ обрађених дигиталних објеката. Алати су представљени према редоследу употребе.

Кључне речи: дигитална библиотека, дигитални алати.

Увод

За разлику од класичних библиотека, већином заступљених у установама културе у Републици Србији, у којима је могућ само преглед појединачних дигиталних објеката, и то најчешће кроз приказ докумената

у пдф формату или галерије фотографија, функционалне дигиталне библиотеке садрже додатне могућности употребе дигиталних објеката попут претраге у самом тексту дигиталног објекта, филтрирања дигиталних објеката према појединачним метаподацима, дељења метаподатака према агрегаторима као што је Претраживач културног наслеђа.¹ За формирање функционалне дигиталне библиотеке неопходно је обезбедити адекватне алате помоћу којих се дигитализована грађа припрема и обрађује за приказ (унос метаподатака у стандарду за дигитализовану грађу, оптичко препознавање карактера за претрагу у тексту, корекција текста, креирање колекција, дељење метаподатака, веб-приказ). За сваку од наведених компонента постоји одговарајући софтвер који употпуњује процес припреме дигитализоване грађе. ResCarta Toolkit² обједињује све неопходне алате за ефикасну и брзу припрему функционалне дигиталне библиотеке.

О скупу алата ResCarta Toolkit

ResCarta представља пакет софтверских апликација отвореног кода који се користи за стварање репозиторијума стандардизованих дигиталних објеката. Алати су отвореног и модуларног дизајна. Модули за креирање дигиталних објеката чувају метаподатке у форматима Конгресне библиотеке METS³ / MODS⁴ / MIX XML.⁵ Модули за прикупљање и индексирање креирају индексе за брзо преузимање текста у записима. Алати се користе за стварање дигиталних колекција, из различитих аналогних и дигиталних извора, које су јавно доступне у оквиру веб-апликације ResCarta.

Алати посебно могу користити малим јавним библиотекама за функционални приказ дигиталних колекција.

ResCarta садржи алате за креирање дигиталних објеката, конверзију података, креирање колекција, индексирање метаподатака и садржаја, алате за приказ итд. Сви ови алати омогућавају креирање, управљање, манипулацију и проверу података којима се приступа преко апликације ResCartaWeb.

¹ <https://kultura.rs>

² <https://rescarta.org/index.php/sw/the-toolkit>

³ <https://www.loc.gov/standards/mets>

⁴ <https://www.loc.gov/standards/mods>

⁵ <https://www.loc.gov/standards/mix>

Списак скупа алата:

- ResCarta Metadata Creation Tool – креирање метаподатака дигиталног објекта;
- ResCarta Data Conversion Tool – конверзија грађе у формат архивских података ResCarta;
- ResCarta Textual Metadata Editor – корекција текста након OCR-а;
- ResCarta Audio Transcription Editor – корекција или креирање аудио-транскрипта;
- ResCarta Collections Manager – креирање и администрација колекција;
- ResCarta Indexer – индексација;
- Checksum Verification Tool – верификација валидности обрађених фајлова.

Алати укључени у ResCarta Toolkit не морају бити инсталирани на истом рачунару на којем се ResCarta архива чува или на рачунару ResCartaWeb сервера. Међутим, рачунар на којем су алати подешени треба да има мрежни приступ архиви ResCarta.

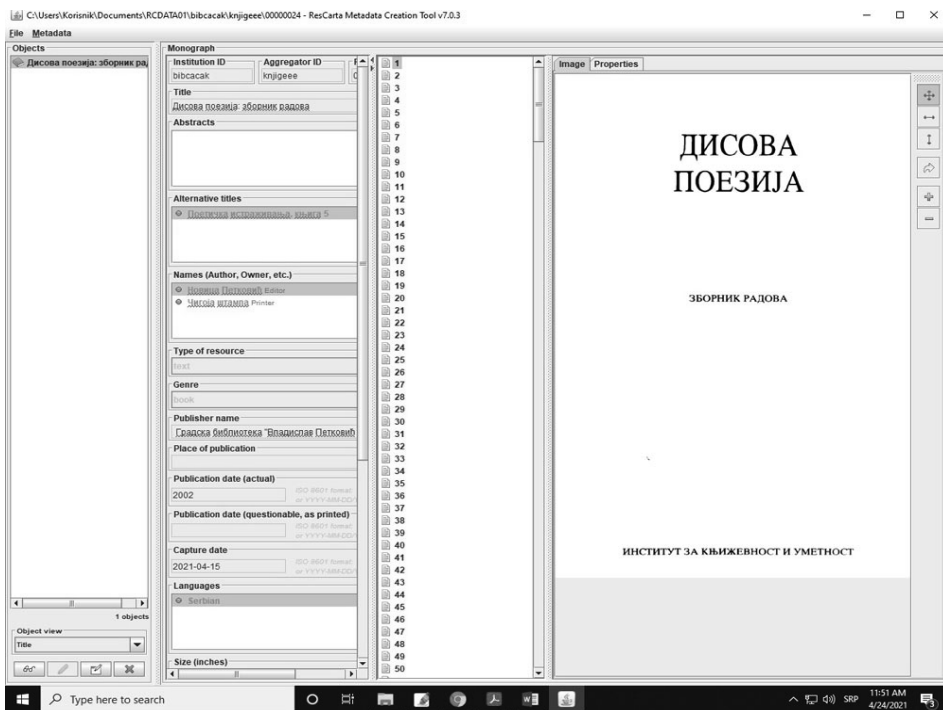
ResCarta Metadata Creation Tool

ResCarta креира метаподатке о објектима („податке о подацима” који се са њима чувају у дигиталном формату). Било којем новом дигитализованом објекту који се направи за уврштавање у дигиталну колекцију потребни су метаподаци објекта да би се организовао и открио.

Метаподаци су структурни подаци који описују запис колекције – наслов, том, аутор, предметна одредница итд.

Алат за стварање метаподатака ResCarta укључује неколико формата за унос метаподатака заснованих на различитим врстама дигиталних објеката, као што су: текстуална, звучна и видео-грађа. Све информације о метаподацима које се унесу, чувају се у стандардним датотекама у формату METS и MODS. Метаподаци се чувају у XML формату. Приказ шема метаподатака у креираном XML фајлу `< xsi:schemaLocation="http://www.loc.gov/METS/ http://www.loc.gov/standards/mets/mets.xsd http://www.loc.gov/mods/v3 http://www.loc.gov/standards/mods/v3/mods-3-4.xsd http://www.loc.gov/mix/ http://www.loc.gov/standards/mix/mix02/mix02.xsd" xmlns:mix="http://www.loc.gov/mix/" xmlns:mods="http://www.loc.gov/mods/v3">`.

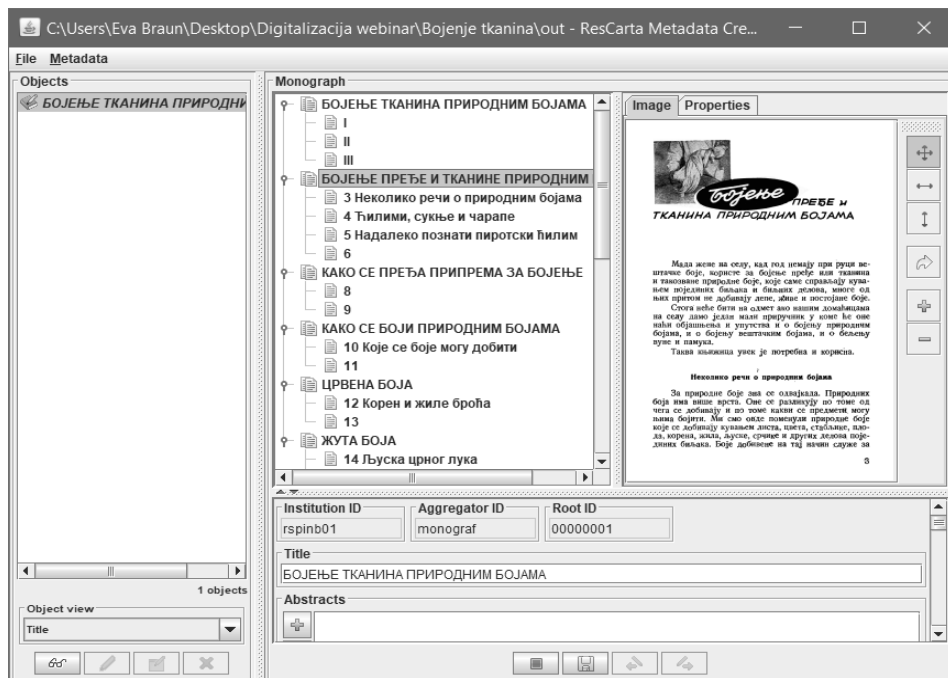
Алат препознаје два типа објеката, објекат је документ (пдф, фотографија) и објекат је директоријум у којем се налази скуп докумената (фајлова), на пример, скуп обрађених скенова једне публикације.



Слика 1. ResCarta Metadata Creation Tool

Поред наслова, који је обавезан податак код монографских публикација да би се сачувао дигитални објекат, Metadata Creation Tool омогућава унос и следећих информација: сажетак, алтернативне наслове, имена (аутора, власника итд.), тип ресурса, жанр, назив издавача, место издавања, датум објављивања, датум скенирања грађе, језик (овај податак је битан уколико се ради OCR текста), димензије грађе, напомене везане за грађу, предмети, ограничења везана за приступ и репродукцију, алтернативне идентификаторе (УДК, ISBN, ISSN...). Препорука је да се унесе што више података за сваки објекат у колекцији.

Структурни метаподаци (пагинација, одељци, поглавља итд.) могу се унети тако што се десним тастером миша кликне на страницу чије податке желимо да изменимо (Слика 2), па селекцијом више страна одједном може да се бира опсег страна и приказ арапских или римских бројева.



Слика 2. Структурни метаподаци

Унос ових метаподатака биће директно видљив кориснику путем веб-апликације у делу за навигацију.

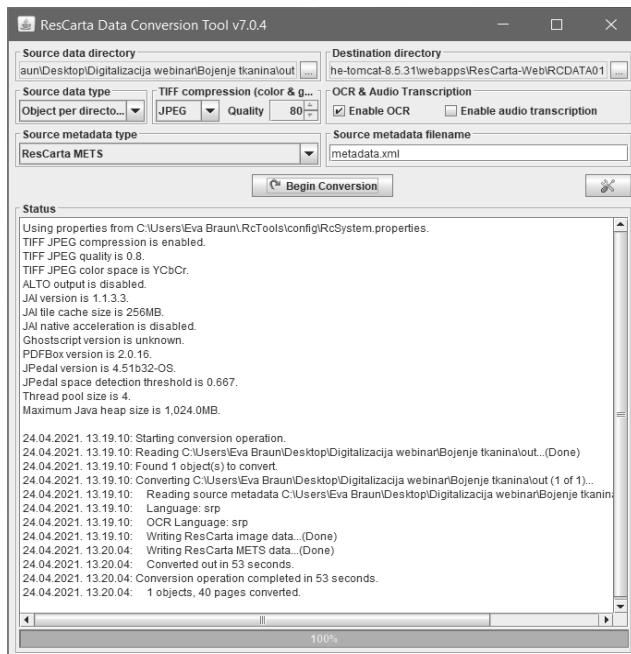
ResCarta алат за конверзију података

Овај алат служи за конверзију TIFF, JPEG, PDF, WAV или MP4 датотеке у формат архивских података у програму ResCarta. Конвертовање је неопходно да би се наставио рад на дигиталним објектима са наредним алатима.

На самом почетку рада овим алатом, корисник треба да изабере улазни директоријум у којем се налазе датотеке за конверзију, као и излазни директоријум где ће се наћи сви конвертовани објекти. Уколико излазни директоријум не постоји у моменту када започне процес конверзије, он ће аутоматски бити креиран и обавезно мора носити ознаку **RCDATA01**.

Као и код претходног алата за унос метаподатака, и алат за конверзију нуди две могућности приликом одабира организације улазних фајлова: „објекат по директоријуму” и „објекат по фајлу”.

Избором опције „објекат по директоријуму” добија се избор у подешавањима излазног TIFF формата са JPEG компресијом или без ње, као и то да ли корисник жели да алат изврши OCR или по потреби аудио-транскрипцију код WAV фајлова.




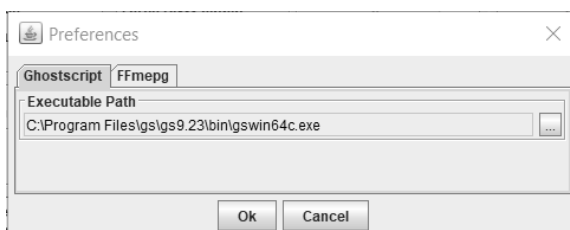
Слика 4. Конверзија података

OCR конверзија врши се преко Tesseract Open Source OCR Engine⁶ и ова опција се активира обележавањем поља у прозору поред текста **Enable OCR**. Препознавање карактера директно зависи од избора језика код уноса метаподатака у претходном алату. Основни језик за OCR је енглески, док је за подршку за српски или неки други језик потребно преузети одговарајући пакет са линка <https://github.com/tesseract-ocr/tessdata/tree/3.04.00>. Преузети пакет треба сместити на локацију C:\Program Files\RcTools-7.0.4\tessdata, а затим унети и измену за додати језик у фајлу RcTessLangMap.properties који се налази на локацији C:\Users\{UserName}\RcTools\config.

Опција конверзије „објекат по фајлу” користи се за сваки документ у оквиру директоријума и представља засебан дигитални објекат, нпр. група

⁶ <https://tesseract-ocr.github.io/tessdoc/OldVersionDocs.html>

разгледница, фотографија или публикација које се налазе у PDF формату. Пре почетка конверзије PDF докумената неопходно је да се у подешавањима преко иконице  изабере путања до инсталације програма GhostScript који је претходно инсталиран на рачунару, а чија је намена аутоматска обрада PDF датотека.



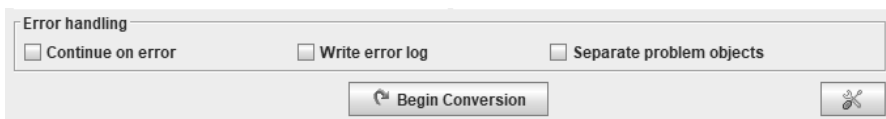
Слика 5. Путања за апликацију GhostScript

Додатне опције које нуди алат приликом конверзије PDF докумената су:

- Читање **налова** PDF документа – где ће алат извући наслов из документа уколико он постоји, у супротном ће за назив документа бити узет за име фајла;
- Читање **аутора** PDF документа – где ће алат, као и у претходном случају, извући име аутора документа уколико оно постоји, уколико ту информацију не пронађе користиће податке унете у пољу *default metadata*;
- **Force B&W output** – креирани излазни фајлови биће црно-бели укључујући и фотографије у колору или палети сивих боја;
- **Archive Source PDF** – подразумева да алат уз конвертоване излазне TIFF слике складишти и оригинални PDF документ.

Управљање грешкама може се контролисати преко следећих опција:

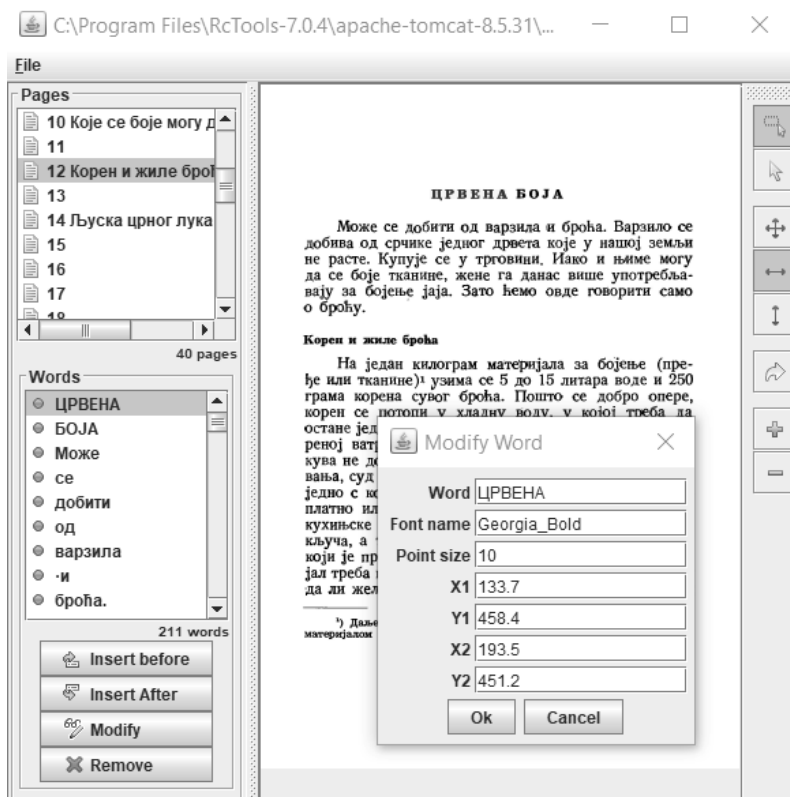
- **Continue on error** – програм ће наставити да ради уколико наиђе на грешку;
- **Write error log** – програм ће креирати евиденцију свих проблема и грешака током конверзије;
- **Separate problem objects** – програм ће креирати посебан директоријум у оквиру постојећег радног директоријума где ће складиштити све датотеке код којих постоји нека грешка.



Слика 6. Опције за управљање грешкама

ResCarta Textual Metadata Editor

Уређивач текстуалних метаподатака додаје, уређује или брише текстуалне метаподатке (речи које се могу претраживати и које се чувају са сликама докумената) у документе у формату ResCarta. Документи који се скенирају помоћу OCR (*Optical Character Recognition*) софтвера у пдф формат и који се конвертују могу захтевати употребу ResCarta Editor текстуалних метаподатака за исправљање грешака у OCR излазу.



Слика 7. Корекција текста

Овај алат се такође може користити за додавање појмова за претрагу фотографијама (означавање) или другим нетекстуалним објектима. Ови појмови за претрагу могу се наћи помоћу траке за претрагу пуног текста у ResCarta веб-апликацији.

Унос текстуалних метаподатака могућ је искључиво на датотекама које су већ конвертоване помоћу **Data Conversion Tool**, а не и на „сировим“ неформатираним датотекама.

Прозор за рад са текстуалним метаподацима можемо посматрати кроз три сегмента:

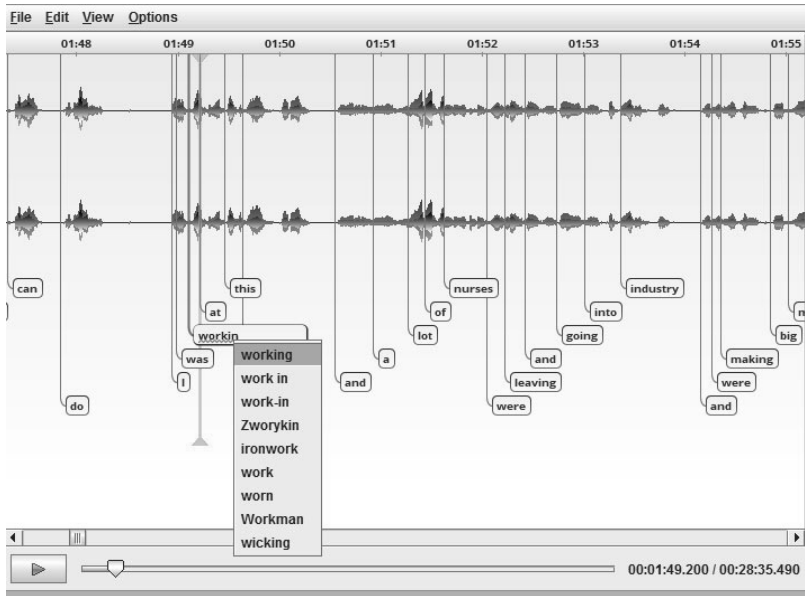
- Део са странама – где су приказане све стране у оквиру једног дигиталног објекта;
- Део са речима – где се налазе све речи на изабраној страни и испод којих се налазе алати за додавање, измену или брисање текста;
- Део за преглед – где ће бити приказана изабрана страна и на њој жутом маркацијом означена реч коју уређујемо, као и сет алата за селекцију текста и подешавање приказа стране.

За сваку реч у тексту постоји информација о фонту и величини фонта, као и о њеној позицији на страни дефинисаној кроз X,Y координате (X1,Y1 представљају координате за доњи леви угао, а X2,Y2 за горњи десни угао).

Audio Transcription Editor (ATE)

Постоји много програма за оптичко препознавање знакова за стварање текста из слика које садрже текст, али мало је извора софтвера за израду транскрипција аудио-датотека које садрже изговорени текст. Алат за претварање података један је од ретких програма за аутоматску транскрипцију звука (Automatic Audio Transcription, ААТ). Као и рани OCR програми, квалитет аутоматски произведене транскрипције зависиће од квалитета оригиналног записа. Дакле, биће потребан алат за препознавање аудио транскрипција.

АТЕ се може користити и као алат за ручну транскрипцију. Коришћење алата за претварање података са укљученим ААТ-ом избациће локације речи за оно што ААТ препознаје. Дакле, препознате енглеске речи могу се заменити одговарајућим језичким изразом. Овај алат кодира текст користећи Unicode шему, тако да је подржана већина светских језика.



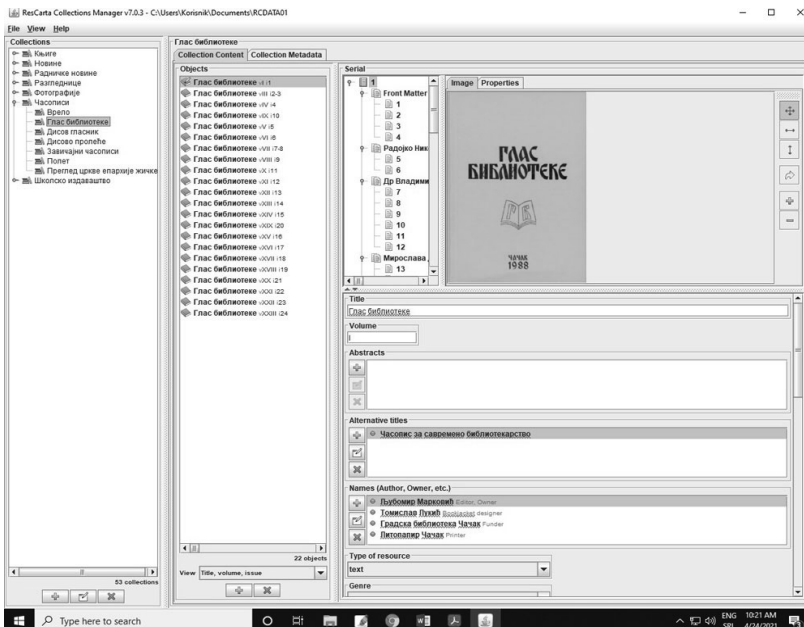
Слика 8. Корекција аудио-транскрипта

ResCarta Collections Manager

Креирани дигитални објекти распоређени су у структуру директоријума и имају повезане метаподатке уграђене у датотеке заједно са спољном стандардном XML датотеком метаподатака. Следећи корак је прикупљање ових предмета у дигиталне колекције.

ResCarta Collections Manager користи се за стварање колекција усклађених са софтвером ResCarta и додавање или брисање из ResCarta колекције дигиталних објеката. ResCarta Collections Manager испишује METS датотеку која се чува на највишем нивоу ResCarta архиве података. ResCarta Collections Manager такође се може користити за претварање робусних метаподатака METS-а у једноставне Dublin Core XML датотеке за дељење метаподатака у другим системима преко OAI-PMH⁷ протокола.

⁷ <https://www.openarchives.org/pmh>



Слика 9. Администрација колекција

ResCarta Indexer

Једном када су организовани дигитални објекти и дефинисане колекције, потребно је поставити их на веб-локацију да би их корисници могли видети. Овај програм ће створити пуни индекс за сваку реч у архиви. На овај начин су индексирани елементи метаподатака у датотекама након оптичког препознавања карактера (OCR) и аутоматске аудио-транскрипције (AAT).

ResCarta Indexer креира индекс ResCarta објеката у корисничким колекцијама. То је библиотека претраживача текстова високих перформанси, у потпуности написана у програмском језику Јава; то је технологија погодна за готово сваку апликацију која не зависи од одређеног оперативног система, инсталација алата је доступна за Linux, Windows и MacOS оперативне системе. ResCarta Indexer користи једноставан унос локација података за стварање овог напредног индекса који омогућава брзи приступ метаподацима и подацима речи који се налазе у ResCarta колекцијама (базама података).

Алат ResCarta Indexer служи да индексира ResCarta колекције како би оне биле претраживе. Свака промена унесена у алату ResCarta Collections Manager мора се индексирати како би била претражива кроз веб-апликацију.



Слика 10. Индексација података

Rebuild Index – Ово поље углавном није означено, а када се означи, алат **ResCarta Indexer** **обрисаће постојећи индекс** и направиће нови, чак и у случају када нема промена у подацима. На пример, уколико је индекс оштећен мора се користити опција Rebuild Index како би се креирао нови и потпуни индекс.

Verbose – Ова опција служи за потпуно информисање корисника о тренутном статусу индексирања, те приказује напредак сваког објекта који је индексан. Када је ово поље искључено, биће видљив само укупни радни процес (приказан у процентима) на дну панела алата. Колекције које имају више хиљада објеката могу се брже индексирати када је ова опција искључена.

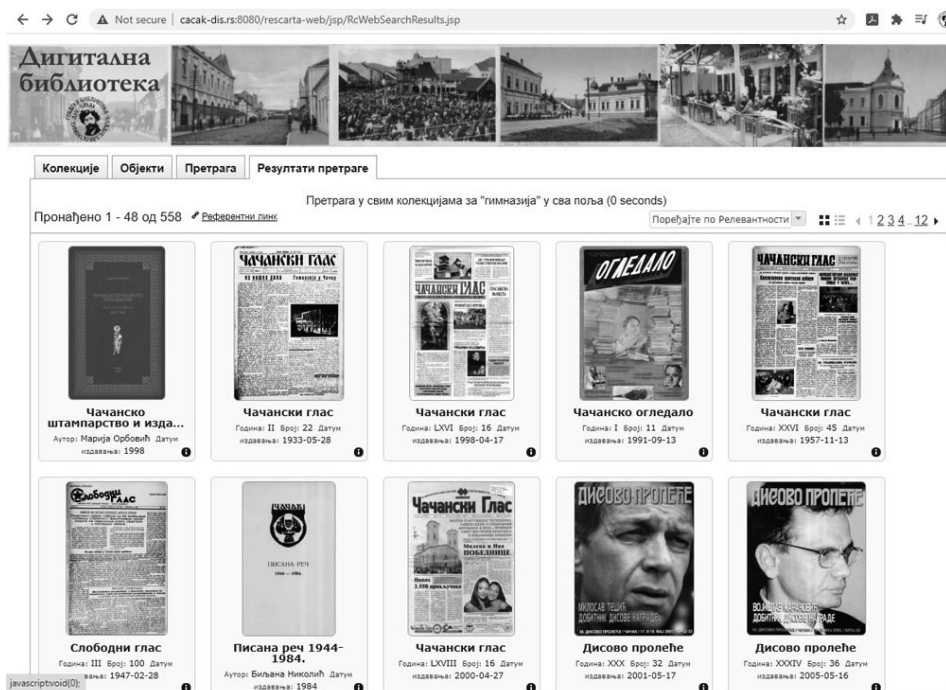
Omit "All" Field – Ова опција из индексације искључује поље **All**, редукујући величину индекса и вероватно може убрзати перформансе претраге веома великих колекција.

N-grams – Маркирање овог поља искључује опцију Lucene N-grams и редукује величину индекса. Ова опција омогућава секвенцијалну претрагу речи и употребу симбола * који замењује остатак речи, на пример унета претрага срп* даће резултате пронађених речи српски, српског итд.

ResCartaWeb

ResCartaWeb је апликација заснована на веб-прегледачу за приступ дигиталним материјалима у познатом окружењу веб-прегледача. До сада су приказани алати који служе за припрему и обраду дигиталних објеката, формирање колекција и индексацију. ResCartaWeb представља засебну апликацију за приказ дигиталних колекција и објеката на јавно доступној веб-адреси. Комплетан садржај приказан у веб-апликацији налази се у бази RCDATA01 која је формирана употребом претходно објашњених алата.

Доступне су странице са приказом колекција, приказом појединачних објеката са функцијом селекције према метаподацима и појединачних колекција и потколекција, једноставно претраживање, приказ резултата претраге и преглед дигиталног објекта.



Слика 11. Веб-апликација – страна Резултати претраге

Једноставно претраживање омогућава претрагу целог текста у свим доступним дигиталним објектима за унете појмове за претрагу. Објекти који садрже један погодак или више њих наведени су у резултатима

претраживања и ранжирани према релевантности у складу са терминима за претрагу.

Могуће је користити и логичке упите приликом претраге, односно претрагу на основу близине термина, користећи једну од опција: „AND”, „OR”, „NO” и „NEAR”


- „AND” захтева да су сви упитни термини присутни и запису;
- „OR” проналази објекте који садрже најмање један упитни термин;
- „NOT” проналази објекте који истовремено укључују први и искључују други упитни термин;
- „NEAR” проналази записе у којима су упитни термини удаљени за највише пет речи.

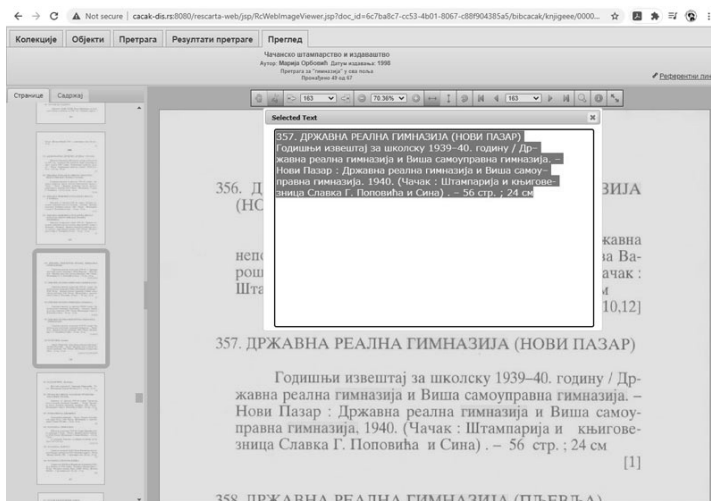
Резултати претраживања приказују се редом према укупном броју погодака / укупном броју појмова на страници у сваком документу (у поређењу са осталим објектима у збирци), а прво се приказује објект са највећим бројем погодака. За референцу, укупан број резултата претраге приказан је на врху добијених резултата, заједно са терминима за претрагу.

Страница **преглед** приказује изабране дигиталне материјале, а помоћу навигације по страници и алата за преглед доступне су опције за навигацију и преглед објеката.



Слика 12. Алати за преглед дигиталних објеката

Ако слика садржи текст, помоћу  алатке за исецање текста **Select**, изабрани део текста се копира у текстуалном формату. Квалитет текста зависи од употребе и уређивања оптичког препознавања знакова или првобитно претвореног дигиталног документа. Изабрани текст појавиће се у дијалогском оквиру за текст, који се даље може копирати помоћу команди за копирање.



Слика 13. Селекција текста за копирање

Опширније упутство функција веб-апликације доступно је у видео-формату на адреси https://youtu.be/_Gojjszр8-A

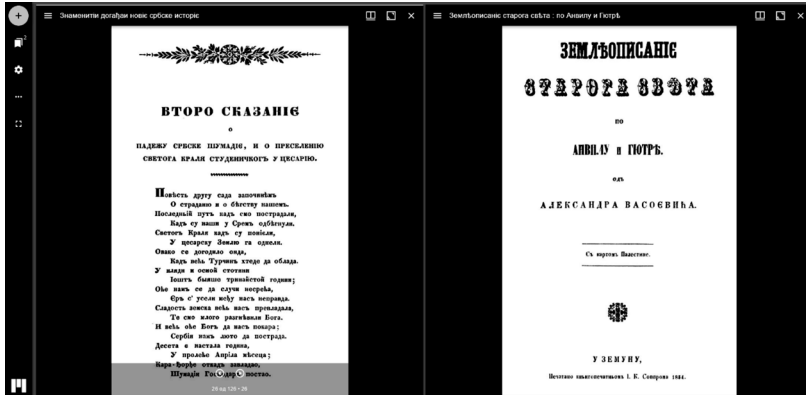
Подршка за стандард IIIF (*International Image Interoperability Framework*)

Последњих година дигитални садржаји водећих светских библиотека приказују се у IIIF⁸ стандарду. Имплементацијом овог стандарда могућ је упоредни преглед више дигиталних објеката из различитих дигиталних колекција на једном прегледачу. На пример, могу се упоредно и на једном месту прегледати дигитални објекти из Британске библиотеке и Националне библиотеке Француске.

У ResCarta веб-апликацију интегрисан је IIIF стандард са додатним прегледачем дигиталних објеката Mirador⁹ који се налази на страници IIIF Преглед.

⁸ <https://iiif.io/community/faq>

⁹ <https://projectmirador.org>



Слика 14. IIIF Преглед – упоредни приказ дигиталних објеката

Закључак

Скуп алата обједињених у једну апликацију ResCarta Toolkit јавним библиотекама пружа ефикасно решење организације и припреме дигиталних садржаја за приказ у функционалној дигиталној библиотеци. Целокупан процес припреме свих типова грађе (текстуалне, сликовне, аудио и видео) обавља се у јединственом систему, на локалном рачунару или на више рачунара у библиотеци. Веб-апликација ефикасно ради и на мањим серверима, али је неопходно обезбедити довољно складишног простора. Инсталација сета алата је једноставна, преко једног инсталационог фајла, на клик. Рад у представљеним алатима је једноставан и може се лако савладати са основним библиотечким знањем. Сви процеси описани у овом раду у складу су са Смерницама за дигитализацију.¹⁰

Izvori:

Documentation for old versions of Tesseract <https://tesseract-ocr.github.io/tessdoc/OldVersionDocs.html>

IIIF Frequently Asked Questions <https://iiif.io/community/faq>

METS – Metadata Encoding & Transmission Standard <https://www.loc.gov/standards/mets>

¹⁰ [https://www.kultura.gov.rs/extfile/sr/5202/smernice%20za%20digitalizaciju%20\(2\).pdf](https://www.kultura.gov.rs/extfile/sr/5202/smernice%20za%20digitalizaciju%20(2).pdf)

Mirador <https://projectmirador.org>

MIX – NISO Metadata for Images in XML Schema <https://www.loc.gov/standards/mix>

MODS – Metadata Object Description Schema <https://www.loc.gov/standards/mods>

Open Archives Initiative Protocol for Metadata Harvesting <https://www.openarchives.org/pmh>

Pretraživač kulturnog nasleđa <https://kultura.rs>

ResCarta <https://rescarta.org/index.php/sw/the-toolkit>

Smernice za digitalizaciju kulturnog nasleđa Republike Srbije [https://www.kultura.gov.rs/extfile/sr/5202/smernice%20za%20digitalizaciju%20\(2\).pdf](https://www.kultura.gov.rs/extfile/sr/5202/smernice%20za%20digitalizaciju%20(2).pdf)

Aleksandar Vukajlović

City Library “Vladislav Petković Dis”, Čačak
acovukajlovic253@gmail.com

Milan Jovanović

Pirot Public Library
mixaptt@yahoo.com

Andrija Sagić

Public Library “Milutin Bojić”, Belgrade
andrija.sagic@milutinbojic.org.rs

SET OF TOOLS USED FOR CREATING A DIGITAL LIBRARY – RESCARTA TOOLKIT

Summary: This paper presents a set of tools necessary for creating a functional digital library. It is used in four libraries in the Republic of Serbia: Belgrade City Library, City Library “Vladislav Petković Dis” in Čačak, Public Library “Milutin Bojić” in Belgrade, and Pirot Public Library. An effective digital library contains standardized metadata, a text search function, a fast and intuitive web interface, and a metadata filter. The ResCarta Toolkit contains all the necessary components to quickly and efficiently create a functional digital library, from metadata input to a web application presenting digital objects. The tools are presented in the order of use.

Keywords: digital library, digital tools.

Примљено: 10. априла 2021.

Исправке: 17. маја 2021.

Прихваћено: 1. јуна 2021.