

Проналажење информација у дигиталним библиотекама



Универзитет у Београду - Универзитетска библиотека "Светозар Марковић"



Република Србија

Министарство културе
и информисања

др Александра Тртовац

Пројекат "Примена напредних технологија у библиотекама", УБСМ, 22.12.2017.

Садржај



Универзитет у Београду - Универзитетска библиотека "Светозар Марковић"

О проналажењу информација

Метаподаци

Аутоматско индексирање

Лексички ресурси и алати

Искуства УБСМ

Закључак

Проналажење информација



Универзитет у Београду - Универзитетска библиотека "Светозар Марковић"

- Термин „**проналажење информација**“ (енг. *information retrieval* – *IR*) први пут је употребио амерички научник **Калвин Нортруп Муерс** (1919-1994) у својој магистарској тези коју је бранио на Институту за технологију у Масачусетсу (MIT) **1948.** године
- Објашњавајући појам цитатом из Муерсовог рада, „Оксфордски речник енглеског језика“ наводи да је **„проналажење информација трагање за информацијама коришћењем референсних материјала, посебно оних информација које су ускладиштене у рачунарским системима.“**

Проналажење информација у библиотекама



Универзитет у Београду - Универзитетска библиотека "Светозар Марковић"

Традиционална библиотека

Електронска библиотека

Дигитална библиотека

Развој библиотека



Универзитет у Београду - Универзитетска библиотека "Светозар Марковић"

Електронске
публикације



Еволуција
библиотека

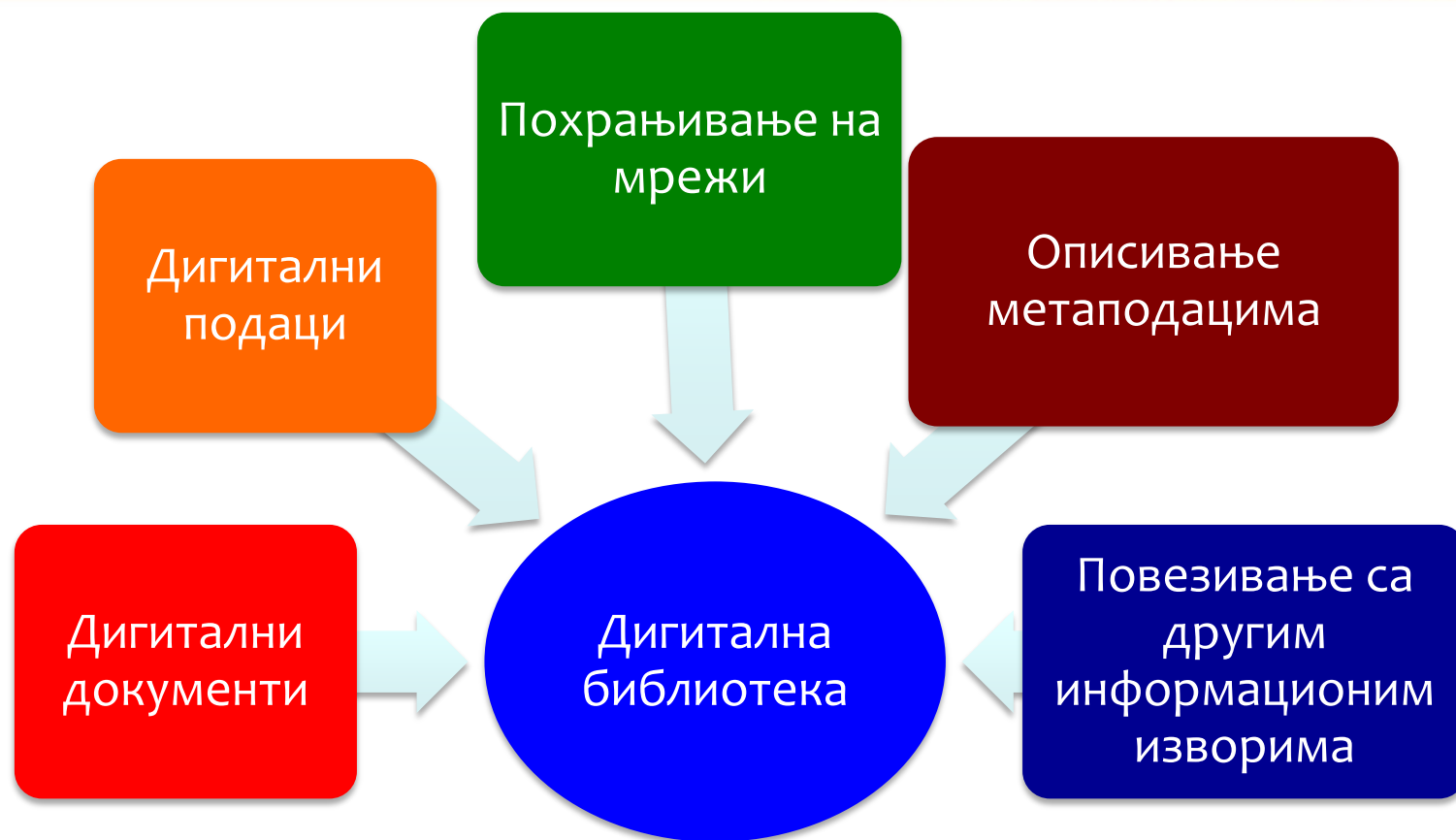


Дигиталне
библиотеке

Дигитална библиотека



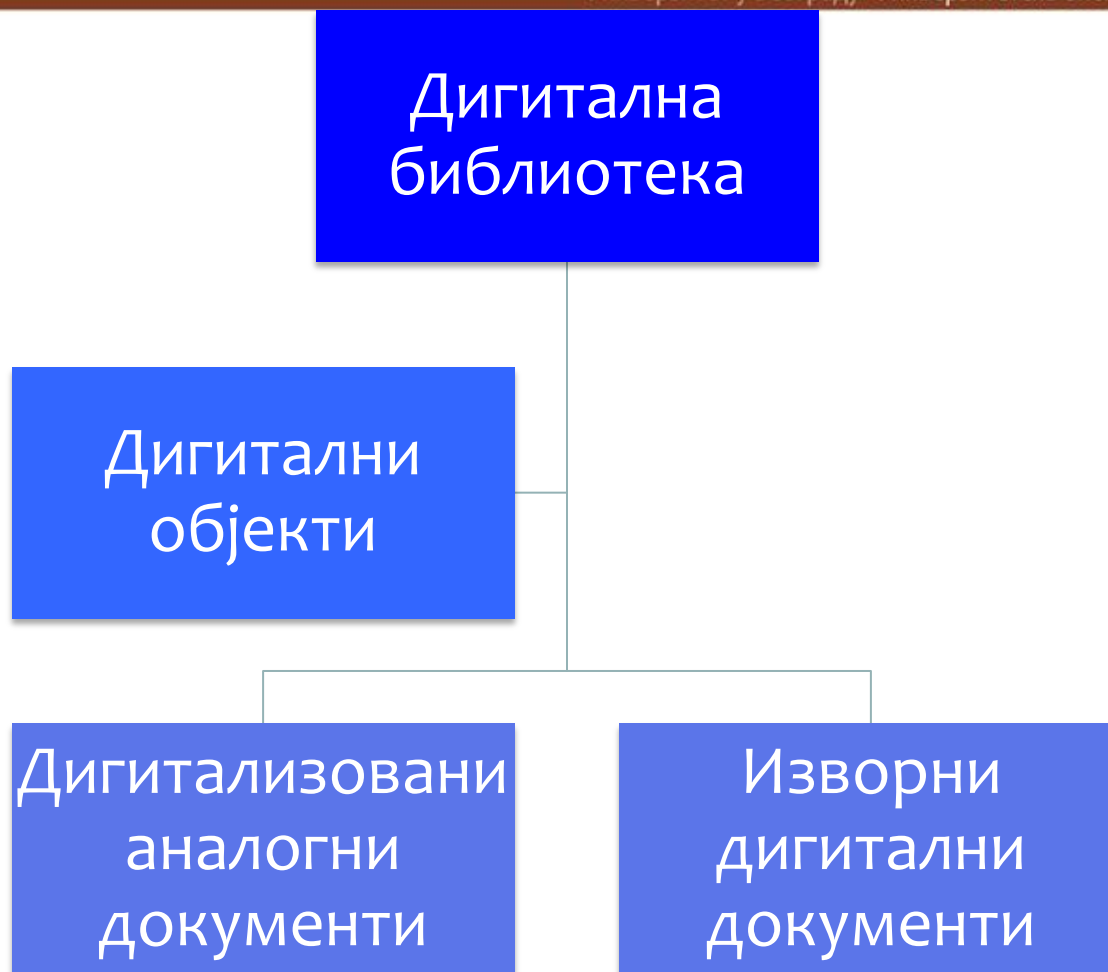
Универзитет у Београду - Универзитетска библиотека "Светозар Марковић"



Дигитална библиотека



Универзитет у Београду - Универзитетска библиотека "Светозар Марковић"



Дескриптори



Универзитет у Београду - Универзитетска библиотека "Светозар Марковић"

- Дескриптори, као **идентификатори** како аналогних тако и дигиталних објеката омогућавају претрагу преко:
 - **метаподатака** или
 - **садржаја**.
- Додела **метаподатака** дигиталним објектима у веома је слична **каталогизацији** у традиционалним библиотекама.
- Код доделе **дескриптора садржаја** нужно је вршити **индексирање документа**, а оно може бити ручно или аутоматско.

Метаподаци



Универзитет у Београду - Универзитетска библиотека "Светозар Марковић"

Метаподаци (структуриране информације)

Библиографски

Административни

Структурални



Проналажење дигиталних објеката

Идентификација
извора

Организовање дигиталних
објеката

Окупљање
извора о
сличним
темама

Дистинкција
различитих
извора

Елиминација
истих извора

Формати за унос метаподатака



Универзитет у Београду - Универзитетска библиотека "Светозар Марковић"

XML - Прошириви језик за обележавање

- најзначајнији и најшире распрострањен формат за пренос података;
- лако разумљива и проширива хијерархијска структура.

RDF - Оквир за опис извора

- напреднија синтакса која наглашава везу између елемената података (не само хијерархијску);
- та синтакса је саставни део развоја **семантичког веба**;
- оригинално креиран као **модел за метаподатке**, па је наишао на значајну **употребу у библиотекама**.

Формати за унос метаподатака



Универзитет у Београду - Универзитетска библиотека "Светозар Марковић"

DC

MODS

METS

LOM

TEI...

- **Оријентисани су ка корисницима**, структура је лака за разумевање и у многоме су **једноставнији и функционалнији** од формата за библиографски опис публикација
- Потребно је пажљиво дефинисање и одабир **обавезних поља** (енг. *mandatory field*) метаподатака

Обавезни метаподаци



Универзитет у Београду - Универзитетска библиотека "Светозар Марковић"

Мали број
метаподатака

- Отежано проналажење информација

Превелики број
метаподатака

- Смањена мотивација корисника

Дескриптори садржаја



Универзитет у Београду - Универзитетска библиотека "Светозар Марковић"

- **Омогућавају претрагу комплетног садржаја** документа.
- **Садржина документа** успешно се може претражити само ако је тај документ индексиран.
- **Додела идентификатора садржаја** корисницима омогућава да преко неког конкретног упита добију одговор у виду одговарајућег документа.

Индексирање



Универзитет у Београду - Универзитетска библиотека "Светозар Марковић"

Штампани или
дигитални
документ

Анализа
садржаја

Придруживање
индексних
термина -
идентификатора

Индексирање



Универзитет у Београду - Универзитетска библиотека "Светозар Марковић"



Аутоматско индексирање



Универзитет у Београду - Универзитетска библиотека "Светозар Марковић"

Аутоматско индексирање

- **фреквенција** појављивања термина (речи) у тексту
- **идентификовање** индивидуалних речи у тексту
- **лексичка анализа**, уклањање функционалних речи и високофреквентних термина јер су недовољно специфични применом искључне листе
- свођење преосталих речи на њихов **корен** или **речнички облик** – лему и сл.

Технологије за пречишћавање текста



Универзитет у Београду - Универзитетска библиотека "Светозар Марковић"

OCR

- **Оптичко препознавање карактера** (Optical Character Recognition)

OLR

- **Оптичка сегментација чланака и класе** **странице** (Optical Layout Recognition)

NER

- **Препознавање именованих ентитета** (Named Entity Recognition)

- Наведене технологије доприносе **квалитетнијој** **аутоматској екстракцији** информација из дигиталних докумената.

Укратко о OLR



Универзитет у Београду - Универзитетска библиотека "Светозар Марковић"

docWORKS[e] - [3 / ### / Newspaper / L'Écho de Paris] - ReadOnly

Document View Detail System Help

Start PageFrames LayoutElements PageNumbers Hierarchy Exported

doc

- pages
 - [1] : TitlePage
 - [2] : imagePage
 - [3] : imagePage
 - [4] : imagePage
 - [5] : imagePage
 - [6] : imagePage
- docContent : Hierarchy
 - [1] : Issue
 - Content
 - [1] : Article (1)
 - Heading
 - Content
 - [2] : Paragraph
 - [3] : Paragraph
 - [4] : Paragraph
 - [5] : Paragraph
 - [6] : Paragraph
 - [7] : Paragraph
 - [8] : Paragraph
 - [9] : Paragraph
 - [10] : Paragraph
 - [11] : Paragraph
 - Footnotes
 - Illustrations
 - Tables
 - [2] : Article (1)
 - [3] : Article (1)
 - [4] : Article (1)
 - [5] : Article (1)
 - [6] : Article (1)
 - [7] : Article (1)
 - [8] : Article (1)
 - [9] : Article (1)
 - [10] : Article (1)

SpecialPages

1
2
3
4
5 5
6

L'ÉCHO de Paris

21.095. 24.06.1937

RÉFLEXIONS

LA POLITIQUE
L'Allemagne et l'Italie
UNE MANIFESTATION
ON GARDE LES MÊMES
LES CHANCES
LES ÉLÉGANCES
A l'écoute...
Il y a cent ans
CE QUE L'ON DIT...

L'ÉCHO DE PARIS

Édition de 5 heures

RÉFLEXIONS sur la guerre d'Espagne

L'Allemagne et l'Italie se retirent définitivement du système de contrôle

UNE MANIFESTATION des anciens combattants belges contre l'annexion aux Italiens

LES CHANCES DU CABINET CHAUTEMPS

LES ÉLÉGANCES A LA GRANDE COURSE DE HAIES GAGNÉE PAR "EVOHE II"

A l'écoute... Il y a cent ans le Roi Bourgeois jouait le Polka du Roi Soleil

ON GARDE LES MÊMES

Textblock

Active: 0:01:32, Idle: 0:06:17 110 GB NUM

Лексички ресурси и алати



Универзитет у Београду - Универзитетска библиотека "Светозар Марковић"

- Примена побројаних технологија је уско повезана са развојем лексичких ресурса за српски језик.

Корпус савременог српског језика (СрпКор)

Морфолошки речници српског језика

Српски wordnet

Терминолошки речници

Корпуси поравнатих текстова на више језика

Лексички алат Библиша

Улога библиотекара



Универзитет у Београду - Универзитетска библиотека "Светозар Марковић"



... до

Од...



Како побољшати проналажење информација



Универзитет у Београду - Универзитетска библиотека "Светозар Марковић"

- Стручњаци који раде на **додели метаподатка** треба да буду изузето добро обучени да би могли да претпоставе на које све начине корисник може поставити упит.
- Корпус докумената који се претражују преко комплетног текста мора бити **опремљен квалитетним лексичким ресурсима и морфолошким речницима** да би проналажење информација дало релевантне резултате.
- Дигитална библиотека која у проналажењу информација примењује **лексичке ресурсе и морфолошке речнике** нуди неупоредиво **ефикасније и прецизније проналажење информација** од оне у којој се користе само дескриптори метаподатака.

Употреба лексичких ресурса српског језика



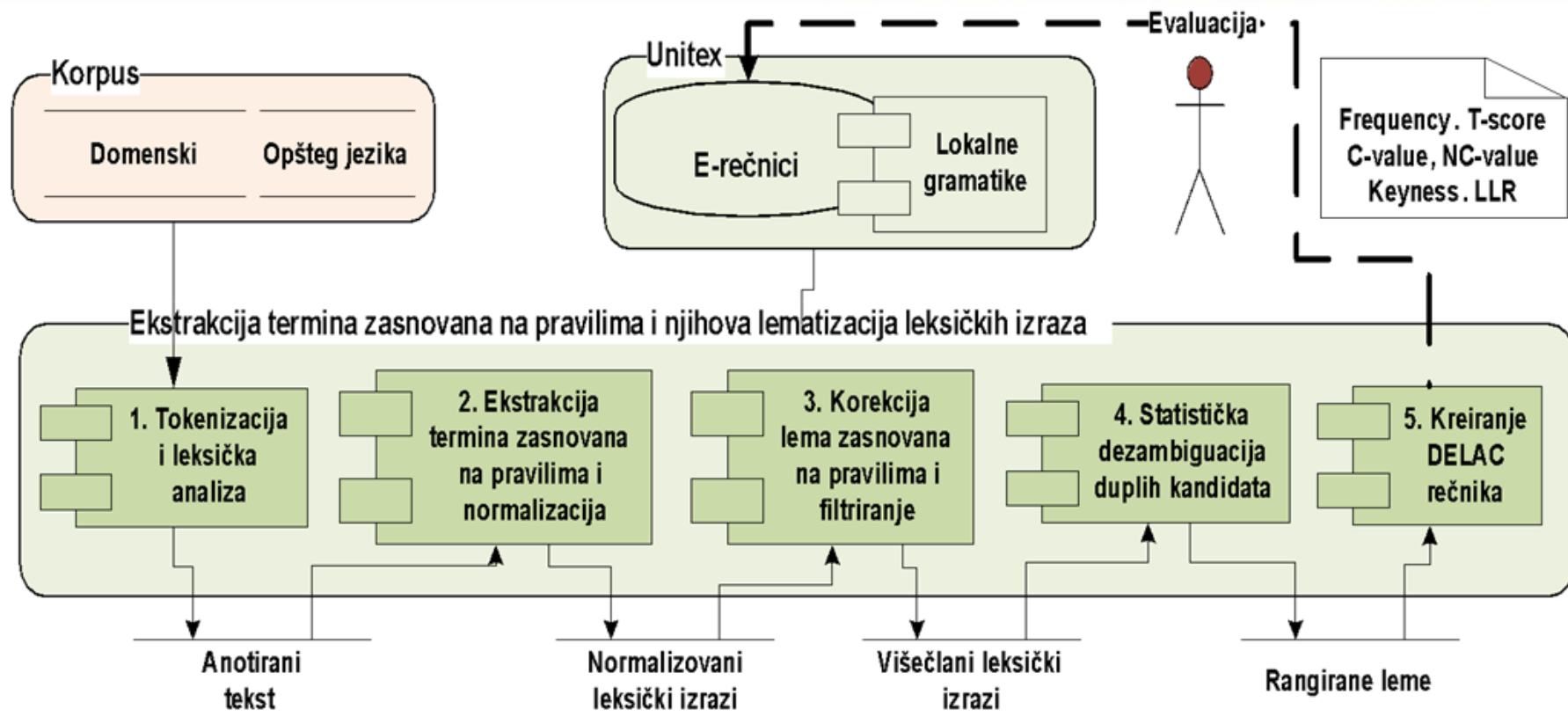
Универзитет у Београду - Универзитетска библиотека "Светозар Марковић"

- Могући модел за прецизније проналажење информација у дигиталним библиотекама, архивским и музејским колекцијама.
- Подразумева коришћење лексичких ресурса, алата и технологија за пречишћавање текста
- Модел тестиран на допуни терминолошког речника из библиотекарства и информатике на основу базе „Библиотекарског терминолошког речника” ауторки Љ. Ковачевић, В. Ињац и Д. Бегенишић (НБС, 2004)
- Валидација модела је извршена на претраживању корпуса дигиталних текстова из библиотекарства и информатике

Архитектура система за аутоматску екстракцију



Универзитет у Београду - Универзитетска библиотека "Светозар Марковић"



Ranka Stanković, Cvetana Krstev, Ivan Obradović, Biljana Lazić and Aleksandra Trtovac. „Rule based automatic multi-word term extraction and lemmatization.” In: 10th edition of the Language Resources and Evaluation Conference (LREC), 23-28 May 2016, Portorož (Slovenia)



BIBLIŠA: ALIGNED COLLECTION SEARCH TOOL

[Home](#) [Metadata browse](#) [Metadata search](#) [Mongo search](#) [Manage data](#) [Help](#) [Tutorial](#) [About](#)

WELCOME TO ALIGNED TEXT COLLECTION SEARCH TOOL!

Keyword Text collection

Synonyms	en	sr
☒ WordNet...		
☒ Dictionary of Librarianship ...	digital library,d-lib,electronic library	digitalna biblioteka
☒ Biblimir ...		
☐ GeolISSTerm...		
☐ RudOnto ...		
☒ Termi ...		digitalna biblioteka

☐ Include Hypernyms ☐ Include Hyponyms

Database: ☐ MarkLogic ☒ MongoDB Match query: ☒ EN&SR ☒ EN ☒ SR

☒ Morphological query expansion

Библиша - пример



Nedeljkov, 2016, vol. XVI:1/2, ID: 1.2016.1/2.5 metadata	n67 Today a great number of Serbian institutions of culture and education have their own digital library or digital repository online and provide access to books in electronic form (IFLA/UNESCO Digital Library Manifesto 2016).	n67 Danas veliki broj ustanova kulture i obrazovanja u Srbiji ima svoju digitalnu biblioteku ili digitalni repozitorijum na internetu, koji omogućava pristup knjigama u elektronskom formatu (IFLA/UNESCO Manifest za digitalne biblioteke, 2016).
Gardašević, 2013, vol. XIV:1, ID: 1.2013.1.4 metadata	Stanislava Gardašević, stanislava.gardasevic@nb.rs, National Library of Serbia, Department for Development of Digital Library and Micrography, Belgrade, Serbia	Stanislava Gardašević, stanislava.gardasevic@nb.rs, Narodna biblioteka Srbije, Odeljenje za razvoj digitalne biblioteke i mikrografiju, Beograd, Srbija
Gardašević, 2013, vol. XIV:1, ID: 1.2013.1.4 metadata	Representative example of such ontology usage is Europeana Data Model (EDM), developed for the purposes of Europeana, European digital library, museum and archive.	Primer uspešnog korišćenja ontologije kao šeme maetapodataka jeste Europeana Data Model (EDM), razvijen za potrebe Europeane4, evropske digitalne biblioteke, muzeja i arhiva.
Lazić et al., 2014, vol. XV:1, ID: 1.2014.1.7 metadata	She presented method of work and functions of digital library based on TMX documents and browser that provides morphological query expansion.	Predstavila je način rada i funkcije digitalne biblioteke zasnovane na TMX dokumentima i pretraživaču koji omogućava morfološko proširenje upita.)
Krstev et al., 2010, vol. XI:2, ID: 1.2010.2.5 metadata	International two-year Master programme in Digital Library Learning (DILL) organized jointly by three partner institutions: Oslo University College, Tallinn University, and Parma Univer-sity offers to its students two interesting courses: "Research Methods and Theory of Science" that presents "... the research process from formulating a problem, via data collection and analysis, to result reporting." and "Users and Usage of Digital Libraries: Quantitative and Qualitative Evaluation" in which students are taught to "how digital libraries are valued by their users, and explore ways of permitting the allocation of resources to areas of user-identified needs.	Dvogodišnji međunarodni diplomski program iz oblasti digitalnih biblioteka (DILL) organizovan u partnerstvu tri institucije: Univerzitetskog koledža u Oslu, Univerziteta u Talinu i Univerziteta u Parmi, nudi studentima dva zanimljiva kursa: "Istraživačke metode i teorija nauke" koji prikazuje "...istraživački proces od formulisanja problema, preko prikupljana podataka, do analize rezultata..." i kursa "Korisnici i korišćenje digitalnih biblioteka: Kvantitativna i kvalitativna evaluacija" u okviru koga studenti uče kako "digitalne biblioteke ocenjuju njihovi korisnici, i na koji način se mogu resursi rasporediti tamo gde su korisnicima najpotrebniji."



Шта радимо у УБСМ?

Универзитет у Београду - Универзитетска библиотека "Светозар Марковић"

- **Фразно претраживање** садржаја, односно претрага преко задатог упита састављеног од две или више речи које се у одређеном редоследу појављују у тексту
- Израда **METS/ALTO** фајлова за сваки документ у оквиру дигиталне библиотеке (**позиција речи** у документу)
- Алат базиран на отвореном софтверу **bnlviewer** комбинује добре особине METS и ALTO стандарда, даје **добре резултате** приликом претраге на задати упит



Претраживе дигитализоване историјске новине

Претрага

Напредна претрага

Преглед

"универзитетска библиотека"

10 резултата по страници

Поређај по скору

Све новине...

Од:

До:

Нађи

Резултати 1-10 од 95 пронађених. Претрага трајала 14 милисекунди.

[Следећи >>](#)

Просветни гласник, страна 106

1. 9. 1901.

. и 2 ириправника). Оксфорд **универзитетска библиотека** има 14 чиновника (1 библиотекар са платом од ... руковаоца (амануензиса), 6 практиканата, и 1 приправника). Штрасбург **универзитетска** и земаљска **библиотека** ... библиотекара, 4 библиотекара, 2 помоћника, и 1 лриправника). Бон. **Универзитетска** библ. има 9 чиновннка. (1 ... чиновника. Тургш, **библиотека** академије наука има 9 чиновника. Штутгарт, краљевска јавна **библиотека** има 10 чиновника. Дрезда краљевска јавна **библиотека** има 12 чиновника (1 директора, 2 виша библиотекара, 5

Просветни гласник, страна 78

1. 8. 1901.

1010 ПРОСВЕТНИ ГЛАСНИК Ерланген, **Универзитетска библиотека**, 185.000 св. без дисертација. Буџет 25.437 марака. Есен, Круиова бибдиотека има 94.305 св. а Круиова техничка **библиотека** има 30.000 св. 700 ... **библиотека** 1900 год. имала је 111.809 св. од тога 7553 карата. Буџет 1.500 марака. Еил, **Универзитетска** Еенигсберг, **Универзитетска библиотека**, 230.000 св. и 1.100 рукоп., 1899 г. читано у библиотеци 5.934 св. а ... 80.000 св. 600 рукописа. Буџет 2.500 марака. Читано 1899 г. 3.641 св. Лајициг, **Универзитетска библиотека**

Просветни гласник, страна 79

1. 8. 1901.

. **Универзитетска библиотека**, 400.000 св. (2.900 штамп. првина), 2000 рукописа, 700 карата, 3.600 слика, и 3.200 ... штамп. првина). Буџет 1.500 марака. Тосток. **Универзитетска библиотека** од 318.000 св. Буџет само за ... рукописа, 1.000 повеља. Буџет 1.100 марака. Тибинген. **Универзитетска библиотека** 340.000 св. 3.000 рукоииса ... је 4.192 лица. Буџет 2.500 марака. Марбург. **Универзитетска** биб.шотека 150.000 св. Пуџет 25.754 м. Метен (Баварска). Бенедиктинска **библиотека** са 80.000 свез. (400 штамп. првина). Буџет 1.200 марака. Мец



Манхајм, Јавна библиотека. 1900 год. имала је 23.175 св. 1899 г. посећивало је 4.192 лица. Буџет 2.500 марака.

Марбург. Универзитетска библиотека 150.000 св. Буџет 25.754 м.

Метен (Баварска). Бенедиктинска библиотека са 80.000 свез. (400 штамп. првина). Буџет 1.200 марака.

Мец, Градска библиотека, 80.000 св. (600 штамп. првина), и 1300 рукописа. Буџет 3.500 марака.

Минхен. Универзитетска библиотека, 400.000 св. (2.900 штамп. првина), 2000 рукописа, 700 карата, 3.600 слика, и 3.200 новина. Буџет за набавку књига 15.000 марака. Читалаца је имала 1899 год. 77.413. За читање су две собе једна за професоре а друга за ђаке. — Државна библиотека има 990.000 св. (13.000 штамп. првина) и 40.000 рукописа. 1898 г. читано је 60.000 св. ван библиотеке 80.000 св. на страну послато 12.000 св. Буџет за набавку књига 70.000 марака. — Библиотека музеја има 4.500 св. и графичку збирку од 40.000 листова.

И шта још...



Универзитет у Београду - Универзитетска библиотека "Светозар Марковић"

- Оптичко препознавање руком писаног текста
 - алат [Transkribus](#) развијен на Универзитету у Инсбруку
 - УБСМ га користи као део заједничког рада у оквиру COST акције "E-READ" (Horizont 2020)
- Дигитализовани лисни каталог до 1963.
- Руком писана и литографисана предавања професора Велике школе
- Дневник и писма Бранимира Ћосића...
- Сарадња са библиотекама, музејима, архивима

И још...



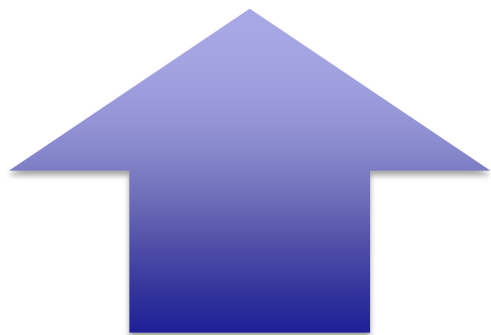
Универзитет у Београду - Универзитетска библиотека "Светозар Марковић"

- Народни граматички речник српског језика УБСМ у сарадњи са Викимедијом Србија...
- Прва фаза **унос именица** у свим падежима
- **Волонтерски рад** великог броја институција и појединаца
- Речник у **отвореном приступу**
- Резултати рада допринеће квалитетнијем проналажењу информација преко свих **морфолошких облика** именица

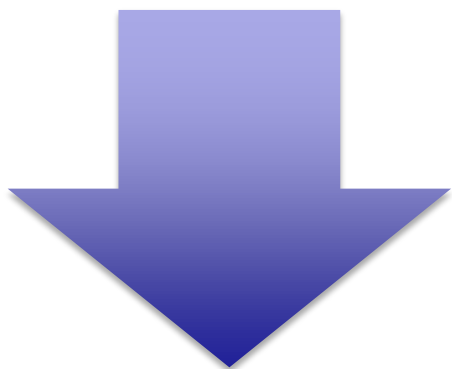
Закључак (1)



Универзитет у Београду - Универзитетска библиотека "Светозар Марковић"



Дескриптори садржаја и
претрага преко целокупног
текста дигиталног
документа



Дескриптори метаподатака
и претрага преко података
о дигиталном документу

Закључак (2)



Универзитет у Београду - Универзитетска библиотека "Светозар Марковић"

Аутоматско
индексирање

Технологије за
пречишћавање
текста

Морфолошки
речници

Лексички ресурси и
алати

Проналажење
информација на
српском језику

Видљивост српског
језика у
дигиталном
окружењу



ХВАЛА НА ПАЖЊИ!

aleksandra@unilib.bg.ac.rs