Digital Library From A Domain Of Criminalistics As A Foundation For A Forensic Text Analysis

Dalibor Vorkapić, Aleksandra Tomašević, Miljana Mladenović, Ranka Stanković, Nikola Vulović

Digital Library From A Domain Of Criminalistics As A Foundation For A Forensic Text Analysis | Dalibor Vorkapić, Aleksandra Tomašević, Miljana Mladenović, Ranka Stanković, Nikola Vulović | International Scientific Conference “Archibald Reiss Days” Thematic Conference Proceedings Of International Significance, Belgrade, 7-9 November 2017 | 2017 | |

http://dr.rgf.bg.ac.rs/s/repo/item/0004992

DIGITAL LIBRARY FROM A DOMAIN OF CRIMINALISTICS AS A FOUNDATION FOR A FORENSIC TEXT ANALYSIS

Dalibor Vorkapić[1]
University of Belgrade – Faculty of Mining and Geology
Aleksandra Tomašević
University of Belgrade – Faculty of Mining and Geology
Miljana Mladenović PhD
eVox Solutions
Ranka Stanković PhD
University of Belgrade – Faculty of Mining and Geology
Nikola Vulović
University of Belgrade – Faculty of Mining and Geology

Abstract: This paper presents a model that provides harvesting, preparation, metadata description, management and exploitation including full text search over documents from a domain of criminalistics written in Serbian. Proposed approach is applied in a web portal that collects various texts derived from journals of The Academy of Criminalistics and Police Studies, Criminal code of Serbia, the "Tara" and "Reiss" conferences, and from some of PhD dissertations related to this field of research. After text processing, a corpus containing over 5500 pages of plain text is created and prepared for publication as an online resource for full text search using Omeka, an open source content management system for on line digital library development. Search capabilities, both full text and metadata search are customized and improved by query expansion via web service relaying on the Serbian morphological dictionary and the Serbian WordNet semantic network for providing morphological and semantic text search expansion. The paper outlines possibilities for further use and analysis on a digital library as a corpus, annotation, tagging, document classification and clustering, as well as sentiment analysis with first results in that direction.

Keywords: Omeka, WordNet, full text search, morphological and semantic text search, query expansion.

INTRODUCTION

A digital library as a special library with a focused collection of digital objects stored as electronic documents can vary in size and scope, and can be maintained by individuals, organizations or institutions. The digital content may be stored locally, or accessed remotely via computer networks. An electronic library is a type of information retrieval system. For this research experiment, the texts from the field of criminology were collected, comprising the articles from journals of The Academy of Criminalistics and Police Studies, Criminal code of Serbia, the "Tara" and "Reiss" conferences, and from several PhD dissertations related to this field of research. The text that is not in Serbian language was removed, as well as tables, figures, references and links, as usual preparation for corpus processing. After this preparation, the text collection contained 5,500 pages of plain text, in A4 format, which was used for further text analysis and processing. For digital objects management Omeka[2], as a web publishing platform and a content management system (CMS) was selected. It is developed by the Centre for History and New Media (CHNM) at George Mason University specially for scholarly content, with an emphasis on digital collections and exhibits. While Omeka may not be as readily customizable as other platforms designed for the widespread use, such as WordPress, Omeka has been used by many academic and cultural institutions, mainly because of its built-in features for cataloguing and presenting digital collections. The content development in Omeka is complemented by an extensive list of descriptive metadata fields that are compliance with Dublin Core, a standard used by libraries, museums and archives. This additional layer helps in establishing proper source attribution, standards for description and organization of digital resources as important aspects of scholarly work in the classroom settings but often overlooked in general blogging platforms.

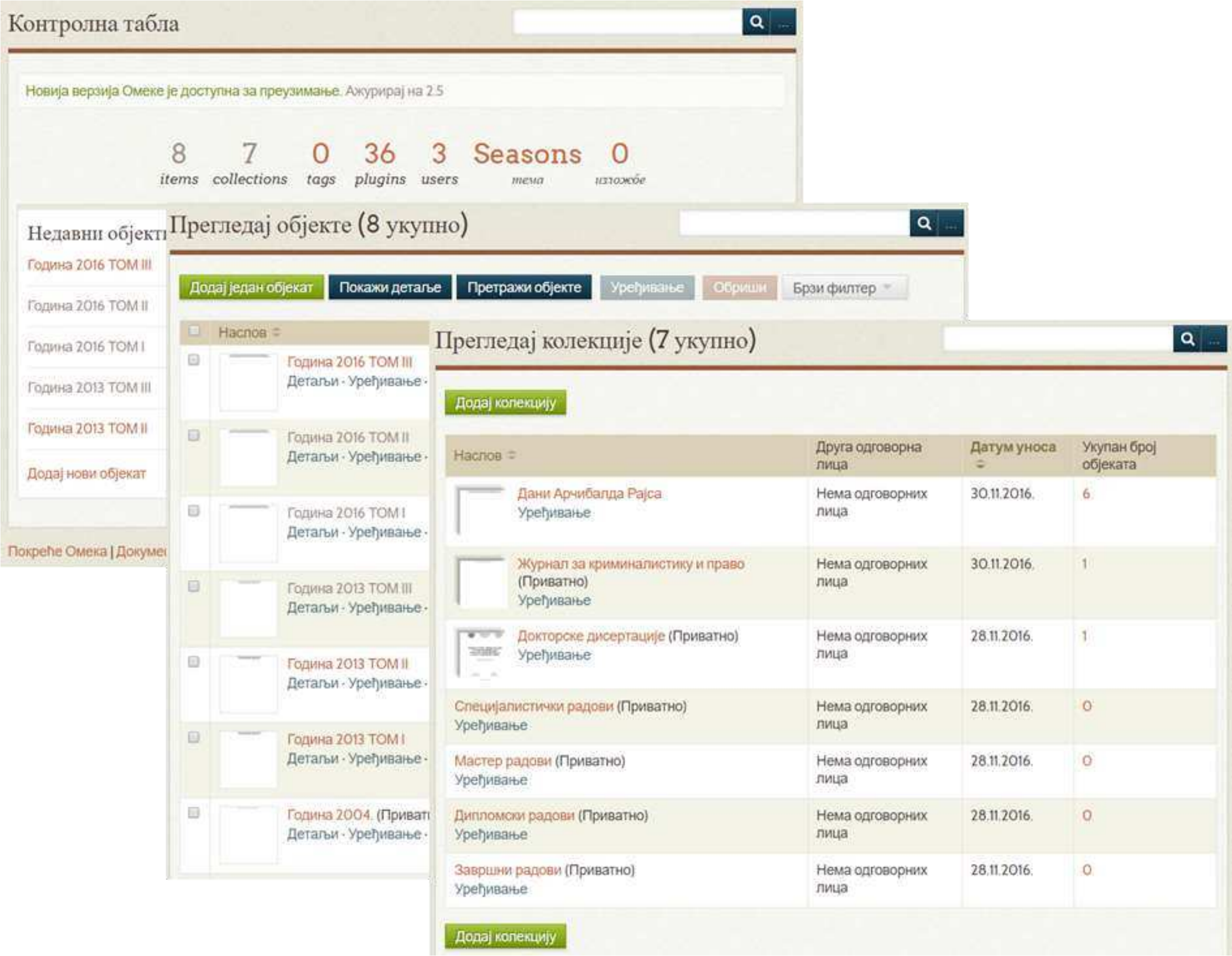
[1] dalibor.vorkapic@rgf.bg.ac.rs
[2] https://omeka.org/

For the digital library presented in this paper, Omeka is installed on operating system Ubuntu 15.10 in a virtual machine. This virtual machine uses 8GB of RAM and 127GB of memory which is dynamically allocated on the storage and one virtual Xeon processor that runs at 2.6GHz. There are several preconditions for installing Omeka:

- Operating systems on which Omeka can work are: Fedora, OpenSUSE and Ubuntu
- It requires an HTTP server, but the recommended is Apache
- Database management system is MySQL server, version 5 or later
- It requires PHP version 5.3.2 or later

The Omeka platform is published under GNU licence[3] (General Public), while basic installation, documentation, plug-ins and best-practice examples are freely available at https://omeka.org. The customisation is user friendly and enable parameters adjustments, including: database name, authentication details, interface language. For Serbian, the localisation is available and easy to implement with changing in */application/config* just *locale* = "" to *locale* = "*sr_RS*".

The Digital library that will be presented in this paper is available at http://master-kpa.rgf.rs/ for search and browse public use and editing management authorized use. The digital library document collection is accessible through user friendly application, that is organised in several categories: Journal for criminalistics and right, Archibald Reiss, Doctoral Dissertation, and other (final process). Further classification is possible, so categories can contain subcategories, to achieve better organisation of digital objects. For each category or subcategory, it is possible to define a specific collection that will display the entire content of the collection. Apart from navigation and browsing of content, simple and advanced search are available. The administrator panel enables installing and customizing the appearance of add-ins that are essential for full system functioning, so the platform could adequately respond to the requests. There are four user roles: *super-user* that can do all tasks in Omeka digital library, *admin* for users administration tasks, but without access to the settings panel, the *contributor* role for editor and *researcher* users for authorized accessing to digital objects.



Контролна табла

Новија верзија Омеке је доступна за преузимање. Ажурирај на 2.5

8 items 7 collections 0 tags 36 plugins 3 users Seasons тема 0 изложбе

Недавни објекти

Година 2016 ТОМ III

Година 2016 ТОМ II

Година 2016 ТОМ I

Година 2013 ТОМ III

Година 2013 ТОМ II

Додај нови објекат

Покреће Омека | Докуме

Прегледај објекте (8 укупно)

□	Наслов
□	Година 2016 ТОМ III Детаљи · Уређивање ·
□	Година 2016 ТОМ II Детаљи · Уређивање ·
□	Година 2016 ТОМ I Детаљи · Уређивање ·
□	Година 2013 ТОМ III Детаљи · Уређивање ·
□	Година 2013 ТОМ II Детаљи · Уређивање ·
□	Година 2013 ТОМ I Детаљи · Уређивање ·
□	Година 2004. (Приват Детаљи · Уређивање ·

Прегледај колекције (7 укупно)

Наслов	Друга одговорна лица	Датум уноса	Укупан број објеката
Дани Арчибалда Рајса Уређивање	Нема одговорних лица	30.11.2016.	6
Журнал за криминалистику и право (Приватно) Уређивање	Нема одговорних лица	30.11.2016.	1
Докторске дисертације (Приватно) Уређивање	Нема одговорних лица	28.11.2016.	1
Специјалистички радови (Приватно) Уређивање	Нема одговорних лица	28.11.2016.	0
Мастер радови (Приватно) Уређивање	Нема одговорних лица	28.11.2016.	0
Дипломски радови (Приватно) Уређивање	Нема одговорних лица	28.11.2016.	0
Завршни радови (Приватно) Уређивање	Нема одговорних лица	28.11.2016.	0

Figure 1. *Control panel, on which administrator regulates the Digital Library*

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

FORENSIC LINGUISTICS

The linguistic study of forensic texts is a part of the field of Natural Language Processing, which includes text types classification and syntax and semantic analysis of texts written in a natural language. Various texts are subject of the study: Acts of Parliament (or other law-making body), private wills, court judgements and summonses and the statutes of the bodies such as States and government departments, cross-examination, evidence presentation, judge's direction, police cautions, police testimonies in court, summing up to a jury, interview techniques, the questioning process in court and police interviews, etc. Generally speaking, any text or item of spoken language can potentially be a forensic text when it is used in a legal or criminal context.[4]
An important part of the forensic texts study is a threat communication. Threat is an important feature in a ransom demand. Ransom demands are examined to identify between genuine and false threats. An example of a ransom note analysis can be seen in the case of the Lindbergh kidnapping, where the first ransom note. From the sentence, the kidnapper makes the claim that the child is in good hands but to make such a claim, the note would have to be written before the perpetrator enters the premises. Therefore, the claim is false (at the time of writing) since the kidnapper had not even encountered the child when he wrote the note. [5] Kidnappers may write statements that later end up being true, such as "your child is being held in a private location" being written ahead of time.
Here are some facts about suicide letters: A suicide note is typically brief, concise and highly propositional with a degree of evasiveness. A credible suicide letter must be making a definite unequivocal proposition in a situational context. The proposition of genuine suicide is thematic, directed to the addressee (or addressees) and relevant to the relationship between them. Suicide notes generally have sentences alluding to the act of killing oneself, or the method of suicide that was undertaken. [6] The contents of a suicide note could be intended to make the addressee suffer or feel guilt. Genuine suicide letters are short, typically less than 300 words in length. Extraneous or irrelevant material is often excluded from the text.
Next forensic text type is death row statements. They either admit the crime, leaving the witness with an impression of honesty and forthrightness; or deny the crime, leaving the witness with an impression of innocence. They may also denounce witnesses as dishonest, critique law enforcement as corrupt to portray innocence or seek an element of revenge in their last moments. Death row statements are within the heavily institutionalized setting of death row prisons. The Forensic Linguistics Institute holds a corpus of these documents and is conducting research on them. And the last but not the least important is social media. Social media statements are often context specific, and their interpretation can be highly subjective. Forensic application of a selection of stylistic and stylometric techniques in a simulated authorship attribution case involving texts has been done in relation to Facebook.[7] Analysis of social media postings can reveal whether they are illegal (e.g. sexist) or unethical (e.g. intended to harm) or whether they are not (e.g. simply provocative). [8]

SOFTWARE SOLUTIONS MODEL

The human language processing group (HLT group) at the University of Belgrade is engaged for many years now in a task of producing various language resources[9], both corpora and lexicons. Given the fact that these resources have been developed for many years, they have naturally been conceived within different frameworks and the technological point of view. Although the HLT group made every reasonable effort to keep the resources as coherent and standardized as possible, a certain level of heterogeneity was inevitable. Hence, due to the growth of the volume of resources as well as their different usage, there was a need for developing a set of tools that would facilitate the maintenance and exploitation in different domains and scenarios. Embarking on this task, the HLT group has produced a workstation for language resources, labelled LeXimir and set of web services Vebran, which greatly enhances the potentials of manipulating each particular resource as well as several resources simultaneously[10].

[4] John Olsson (2008).. Forensic Linguistics, Second Edition. London: Continuum ISBN 978-0-8264-6109-4

[5] Falzini, Mark W. (9 September 2008). "The Ransom Notes: An Analysis of Their Content & "Signature""

[6] John Olsson (2004). An Introduction to Language Crime and the Law. London: Continuum International Publishing Group

[7] C.S. Michell (2013). Investigating the use of forensic stylistic and stylometric techniques in the analysis of authorship on a publicly accessible social networking site (Facebook) (MA in Linguistics thesis). University of South Africa

[8] C. Hardaker (2015). The ethics of online aggression: Where does “virtual” end, and “reality” begin? BAAL Conference on The Ethics of Online Research Methods. Cardiff

[9] Cvetana Krstev, Duško Vitas, “Corpus and Lexicon - Mutual Incompletness ”, in Proceedings of the Corpus Linguistics Conference, 14-17 July 2005, Birmingham, eds. Pernilla Danielsson and Martijn Wagenmakers, ISSN 1747-9398, http://www.corpus.bham.ac.uk/PCLC/, 2005

[10] Cvetana Krstev, Ranka Stanković, Duško Vitas, Ivan Obradović, “The Usage of Various Lexical Resources and Tools to Improve the Performance of Web Search Engines”, in Proceedings of the Sixth Interantional Conference on Language

To keep development and use of the applications and resources at the same time, without frequent conversions, the strategy for the development was to support original formats used in another software tools for language resources processing (Unitex, WorNet, LeXimir, Vebran). Another important decision was to try re-using the available software, avoid developing existing functions, but focus on the necessary functions that did not exist and then integrate them with other tools.

There are diverse ways to integrate existing and new systems into a hybrid tool. Every method has good and bad characteristics and because of that it can't be used in every situation. Problems exist and best way to solve them is to use different methods. The motivation for creating hybrid systems can be to improve methods, to solve problems complexity with multiple tasks and resolve multifunctionality. Improving the method can be achieved by integrating different methods to overcome specific limitations and disadvantages, combining the method with poor specifications with other method with different specifications. For problems with multiple tasks, or subtasks, which can't be solved with one method, hybrid systems are being created to solve all subtasks with appropriate method. Realisation of multifunctionality is motivated by need to create hybrid systems, within a single architecture, for solving problems in different ways. These systems functionally emulate a variety of methods.
The RESTfull Web services based on Unitex routines are used for the implementation of morphological analysis and output generation relaying on electronic dictionaries. For query expansion are combined morphological and semantic vocabularies, because synonymous terms are taken from WordNet[11] and terminological databases. The hybrid system is replacing of a simple query, based on a keyword, with the expanded one.

METADATA

The metadata are a core for the development of digital libraries, as the structured information that describes, explains, locates, or otherwise makes it easier to retrieve, use or manage an information resource. These data are the key to ensuring that resources will be secured and continue to be accessible into the future. Metadata is often called data about data or information about information.[12] There are several different types of metadata: descriptive – that describes a resource purpose, such as discovery and identification of objects that includes basic elements: title, author, publisher, location, date, language, a unique identifier, description, keywords, subject headings, abstract etc.; structural – describe types, versions and links between digital objects (e.g. connect the original document and all its versions, whereby include information about versions and the information about latest change in them etc.); administrative – contain information on the rights of access to the digital object in accordance with copyright and intellectual property protection, the source, size and type of files on access to the source, the size and display format, and using them to actively monitor the number of users who visit and use certain content.[13] Web platform Omeka use Dublin Core as a standard for displaying metadata. The Dublin Core includes a set of elements to describe a wide range of sources in the network and aims to: simplicity in the creation and maintenance so that each user could make a set of descriptive statements understandable semantics to facilitate searches across the global network to all who are in need of information; localization: originally implemented in English, but there are versions that are written for many other languages (Serbian, Russian, Chinese, Finnish, Norwegian, Japanese...)[14] Dublin Core consists of 15 basic elements: title, subject, description, type, source, relation, coverage, creator, publisher, contributor, rights, date, format, identifier and language. Most often these elements are sufficient to describe the digital object. Web platform Omeka has an extension (or plugin) Dublin Core Extended, that represents the extended list of the basic set of elements. It includes the following elements: Abstract, Access Rights, Accrual Method, Accrual Periodicity, Accrual Policy, Alternative Title, Audience, Date Available, Bibliographic Citation, Conforms To, Date Created, Date Accepted, Date Copyrighted, Date Submitted, Audience Education Level, Extent, Has Format, Has Part, Has Version, Instructional Method, Is Format Of, Is Part Of, Is Referenced By, Is Replaced By, Is Required By, Date Issued, Is Version Of, License, Mediator, Medium, Date Modified, Provenance, References, Replaces, Requires, Rights Holder, Spatial Coverage, Table Of Contents, Temporal Coverage, Date Valid, The dc-rdf Output Format.

Resources and Evaluation (LREC'08), Marrakech, Morocco, 28-30 May 2008, European Language Resources Association (ELRA), 2008

[11] Miljana Mladenović, Jelena Mitrović, Cvetana Krstev, "Developing and Maintaining a WordNet: Procedures and Tools", In The Proceedings of Seventh Global WordNet Conference 2014, eds. Heili Orav, Christiane Fellbaume, Piek Vossan, University of Tartu, Tartu, Estonia, January 25-29, 2014, pp. 55-62, 2014, ISBN 978–9949–32–492–7

[12] HODGE, G., 2001. Metadata made simpler, Niso Press

[13] TRTOVAC, A. S., 2016. Deskriptori metapodataka i sadržaja u pronalaženju informacija u digitalnim bibliotekama. Univerzitet u Beogradu-Filološki fakultet.

[14] MILENKOVIĆ, M., 2003. Dublin Core Metadata Initiative (DCMI). Review of the National Center for Digitization, 70-79

USE CASE DIAGRAM

Use case diagram on the left describe search possibilities offered to the user together with different responsibilities for lexicographer, terminologist and for linguists. Terminologist is generally using search on lemma and synonyms, while linguist is more interested in search by linguistic patterns and syntactic graphs. Figure 2 on the right shows a diagram of a use case for corpus preparation that includes: collection of articles, lexical processing resources, describing text with metadata, analysis of unknown words, complement morphological dictionaries, addition to terminology database, transliteration, correction of broken words, correction of optical character recognition errors.

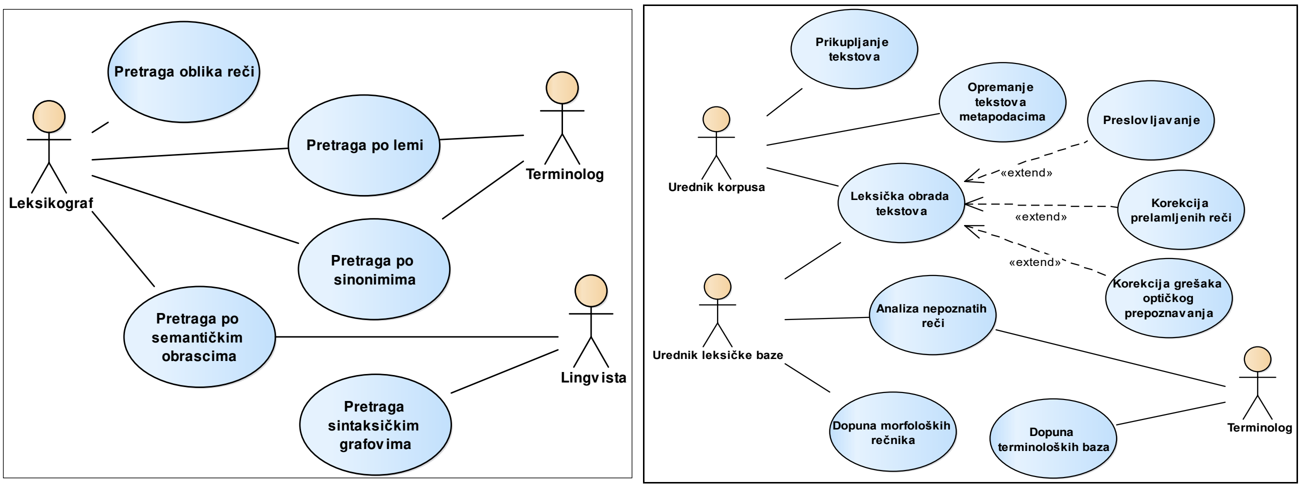


Figure 2. *Use case diagrams: exploitation (left), preparation (right)*

APPLICATIONS FOR LINGUISTIC RESOURCES

The linguistics and lexical resources used for query expansion and text analysis are depicted in Figure 3 on the left, while on the right are main application components of the language support system. Main lexical resources include morphological dictionaries for Serbian language[15], Serbian and English WordNets, terminological databases: Termi, GeolISSTerm, RudOnto and Librarian dictionary. Apart from the grammars in the form finite state automata and transducers, system is using rules for inflection of multiword units. Among textual resources are most important digital libraries, Unitex corpora[16] and CQP web corpora. Linguistic support is implemented via REST web service Vebran that interact from one side with lexical and linguistic resources and from the other with Omeka KPA digital library.

[15] Cvetana Krstev. Processing of Serbian – Automata, Text and Electronic Dictionaries, Faculty of philology, Belgrade, 2008

[16] Duško Vitas, Cvetana Krstev, Ivan Obradović, Ljubomir Popović, Gordana Pavlović-Lažetić", An Processing Serbian Written Texts: An Overview of Resources and Basic Tools ", in Workshop on Balkan Language Resources and Tools, 21 Novembar 2003, Thessaloniki, Greece, eds, S. Piperidis and V. Karkaletsis, pp. 97-104, 2003

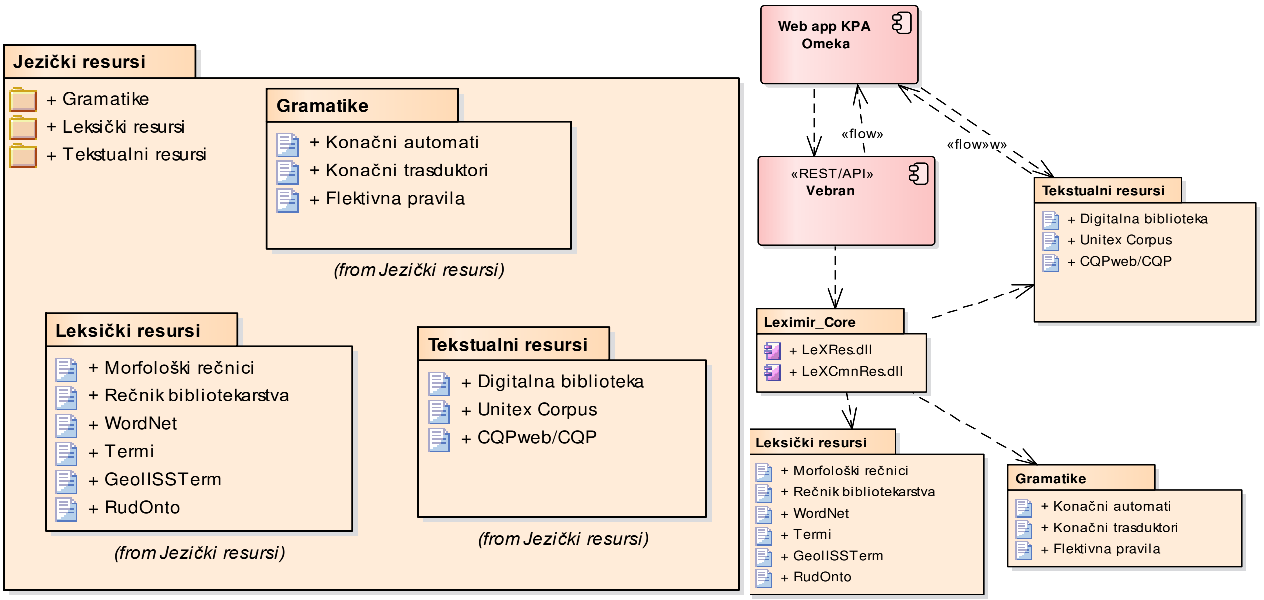


Figure 3. *Linguistic resources applications*

DYNAMIC MODEL QUERY EXPANSIONS

The dynamic aspect of the application is presented in the interaction diagram with a model of query expansion, showing messages sent between objects or class instances, as a series of sequential steps over time. They are used to describe the workflow, message transmission and cooperation of different elements of the system to achieve a result. Figure 3 shows the interaction of the user and system components in case of semantic query expansion[17], which may, but need not include morphological expansion. The class *WNManager* for the WordNet resources management provides a semantic expansion of a query that includes introducing literals from the selected synsets. Expansion with additional relations (hyponymy-hypernym) introduce also literals from the synsets which are connected by the relation of hypernym. Multilingual expansion is implemented using two wordnets and their inter-lingual index. The method differs from the previous only in the possibility of extending the query literals in the other language which can be reached via a synchronized synsets.

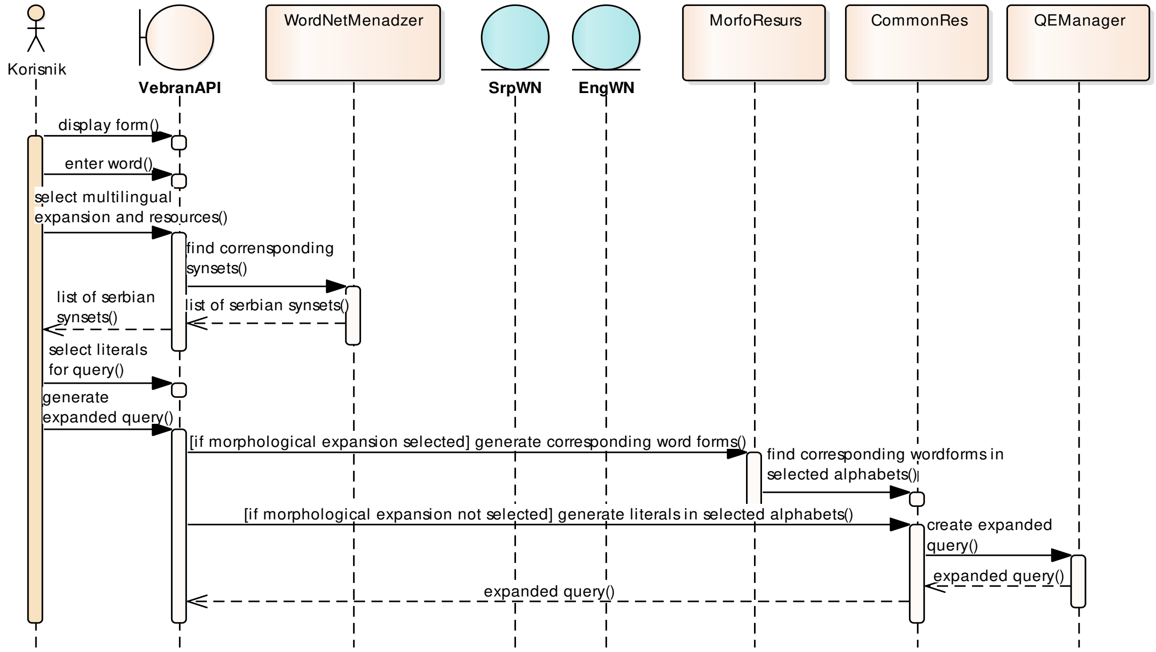

Korisnik
VebranAPI
WordNetMenadzer
SrpWN
EngWN
MorfoResurs
CommonRes
QEManager
display form()
enter word()
select multilingual expansion and resources()
find correnponding synsets()
list of serbian synsets()
list of serbian synsets()
select literals for query()
generate expanded query()
[if morphological expansion selected] generate corresponding word forms()
find corresponding wordforms in selected alphabets()
[if morphological expansion not selected] generate literals in selected alphabets()
create expanded query()
expanded query()
expanded query()

Figure 4. *Sequence diagram a multilingual query expansions*

[17] Cvetana Krstev, Ranka Stanković, Duško Vitas, Ivan Obradović, “The Usage of Various Lexical Resources and Tools to Improve the Performance of Web Search Engines”, in Proceedings of the Sixth Interantional Conference on Language Resources and Evaluation (LREC'08), Marrakech, Morocco, 28-30 May 2008, European Language Resources Association (ELRA), 2008

The search option is on the operator page. There are two ways of search Narrowed search that includes the search on specified fields, collections and kinds, user who added a resource and the geographic address. This search method searches only on the basis of the given word, without changing the form of words. Extended search includes morphological and semantic search. Morphological search includes search of all inflected forms of specified word that retrieve from SrpMD (Serbian morphological dictionary). For nouns, grammatical forms include case and number for example for kuća (eng. House) *kuće, kućama, kući, etc*. for adjective additionally comparison, for verbs person, times etc. Semantic search involves the expansion of the query by searching semantic network Serbian WordNet.[18] This semantic network is based on the concepts among which are semantic relations. With the simple search with keyword *napad* (eng. attack), the system will find only exact match of the word *napad*, while with the extended search for the same word the system will search also for the words: *agresija, agresijama, agresije, …, akcija, akcijama, akcije,…, inicijativa, inicijativama, inicijative, …, napad, napada, napade, …, nasrtaj, nasrtaja, nasrtaje,…, , navala, navalama, navale, …, агресија, агресијама, агресије, …, акција, акцијама, акције, …, иницијатива, иницијативама, иницијативе, … , напад, напада, нападе,…, насртај, насртаја, насртаје,…, навала, навалама, навале, …,* that means all synonyms, both alphabets (Cyrillic and Latin) and all infected forms.

If we want to focuss on word *napad* only, but analyse different contexts of occuraces, more sophisticated query can be requested. The following expression: *<A><napad><PREP><N+Hum>* is an example of morphological and semantic expression search in Unitex system. This query is retrieving any inflective form of lemma *napad* (attack), preceded by an adjective (<A>) and followed by preposition and noun (<N>) that is human (depicted by the semantic mark +Hum), retrieving output concordances like:

se označa-va mesto na kome se javljaju česti napadi na žrtve škole, kada je pitanju imovinski
išljajno ubistvo, kidnapovanje ili neki drugi napad na lica ili slobodu međunarodno zaštićenog
samo kada je to neophodno da se spreči fizički napad na službeno lice, drugog maloletnika ili
me, posle ukazivanja na podatke o broju fizičkih napada na policajce u SAD-u u razdoblju −. go
unima sa navijačima protivničkih ekipa; fizičkim napadima na građane, igrače, službena lica i p
oružja i pretnjom njegove upotrebe ili fizičkim napadom na žrtvu. Od oružja se najčešće kor
svaki oblik fizičkog zlostavljanja ili fizičkog napada na dijete kojim se izaziva ili se može
eno krivično delo iz člana a KZRS, zbog fizičkog napada na sudiju udaranjem pesnicom ruke u pre
i zovu se krivična dela.{S} Mnogi napadi na pojedinca predstavljali su krivično del
pomenuto, očekivan od ljudi osuđenih za najteže napade na pripadnike organa reda, Ali, s dr
ikom pokušaja krivičnog dela, dolazi do neposrednog napada od strane učinioca na društvene odno
nu ili kolektivnu samoodbranu u slučaju oružanog napada protiv člana Ujedinjenih nacija, dok Sa
iz tri naša kaznena zavoda.{S} Rizik ponovljenog napada na pripadnike policije Sa puno osnov
dgovor na pitanja prvo: da li se radi o ponovljenom napadu na službena lica, u drugom slučaju d
UMESTO ZAKLjUČKA Mali broj istraživanja posvećenih napadima na policajce uslovio je potrebu da
jive maštovitosti, teorija koja podvodi preventivni napad na suverenu državu pod izgovorom tero
ativne baze za izvršavanje sistematskih razbojničkih napada na putnike i trgovce. (T. Taranovsk
rorizma koji se sastoji u samoubilačkim terorističkim napadima na ljude i imovinu, svesnim žrt
rezultata kriminološkog istraživanja o teškim napadima na policajce u Srbiji koje je sproveden
odologija kriminološkog istraživanja o teškim napadima na policajce u Srbiji, koje je sproved
anja poslova, smatramo da ubuduće svaki teži napada na službena lica zahteva potpunu analizu ko
adicijom, data dva ilustrativna primera ubilačkih napada na policajce, koji mogu poslužiti kao
Kada se radi o dobu dana, sve podatke o vremenu napada na policajce podelili smo u vremenske p

The Serbian morphological dictionaries cover large lexica, but each special domain has characteristic words that are occurs occasionally in ordinary texts, but frequently in domain specific texts. That is the case with presented collection. Among unrecognized tokens were terms:

[18] I. Obradović, R. Stanković, “Wordnet Development Using a Multifunctional Tool”. Proceedings of the International Workshop Computer Aided Language Processing (CALP) '2007, Borovets, Bulgaria, C. Orasan, S. Kuebler (eds.), pp. 25-32, September 2007.

psihoaktivni, podstrekavati, situacijski, narkokartel, kriminalnopolitički, izvršilaštvo, zakonopisac, procesnopravni, geoprostorni, protivvazduhoplovni, delikvencija,... (eng. psychoactive, incite, site, cartel, criminal and political, complicity, legislator, procedural, geospatial, antiaircraft, delinquency,…).

These words are examples of word candidates for enrichment of morphological dictionary. Their addition will enhance the search performace and criminalistics text analysis. In this research, the analysis included extraction of multi-word units using LeXimir[19] that system retrieved as most frequent:

nasilje u porodici žurnal za kriminalistiku, *izvršenje krivičnog dela*, *policijski službenik*, *policijska akademija*, *pravno lice*, *ljudsko pravo*, *pranje novca*, *radnja izvršenja, država članica, trgovina ljudima,* (eng. domestic violence, journal of criminology, committing an offense, police officer, police academy, legal person, human right, money laundering, action execution, member states, trafficking).

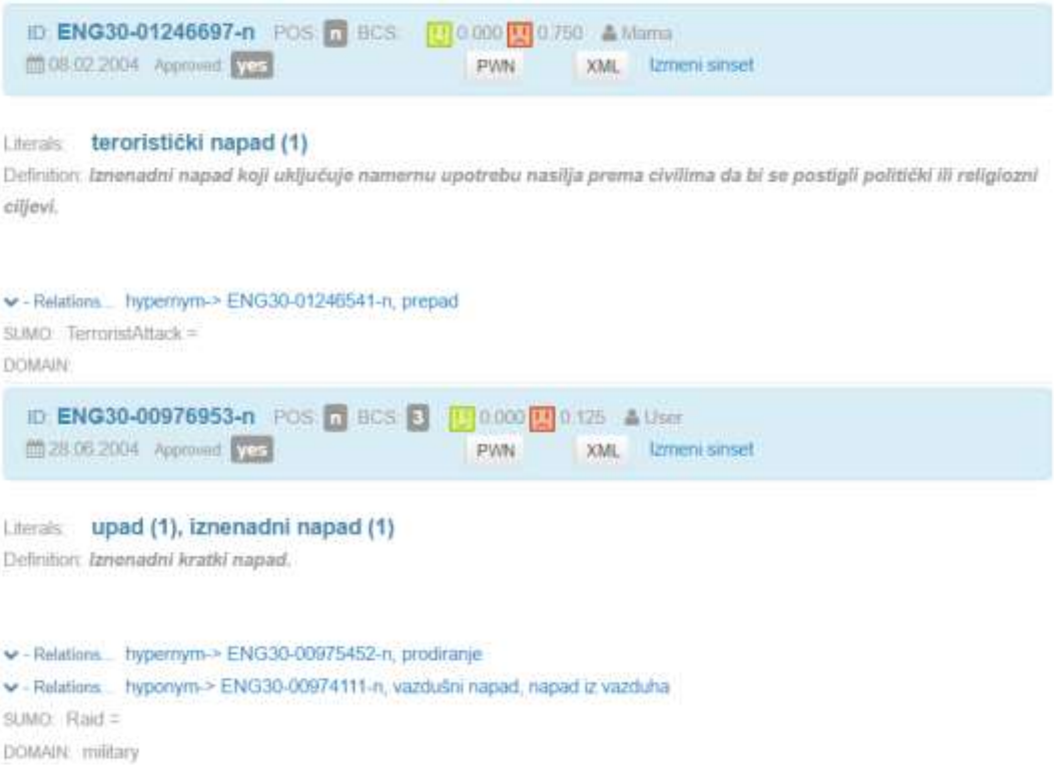
SENTIMENT ANALYSIS AS A NEXT STEP IN A STUDY OF FORENSIC TEXTS

The semantic network Serbian WordNet (SWN) is a lexico-semantic resource that has been developed based on the idea of the Princeton WordNet (PWN), a mental lexicon that helps scientists working on psycholinguistic projects. SWN is a set of above 22.000 concepts called synsets where a concept is represented by the set of synonym word forms that have the same or similar meaning. Synsets respect the syntactic categories noun, verb, adjective, and adverb and can be interconnected by semantic relations, and word forms by lexical relations. Synonymy is WordNet's basic relation, because WordNet uses sets of synonyms (word forms with similar meaning) to represent a concept; Antonymy (opposing-name) is a symmetric relation between two word forms with oposite meaning; Hyponymy (sub-name) and its inverse, hypernymy (super-name), are transitive relations between synsets and they organize the meanings of concepts into a hierarchical structure. Meronymy (part-name) and its inverse, holonymy (whole-name) distinguish component parts. Troponymy (manner-name) is for verbs what hyponymy is for nouns. For the purpose of enriching SWN with the data concerning sentiment measurement, SentiWordNet, a lexical resource for opinion mining based on the Princeton WordNet, is used.20 It assigns three sentiment scores: positivity, negativity and objectivity to each PWN synset, but in SWN two SentiWordNet sentiment scores (positive and negative) used for each SWN synset. A sentiment lexicon is produced using word forms defined in SWN that have positive or negative sentiment scores. This kind of a lexicon is applied in sentiment polarity classification tasks on Serbian texts, achieving 97.1% accuracy over cross-validated datasets and 84.9% and 79,1% over different test datasets. [21] In that terms, SWN can be used in query expansion to give semantic meaning to terms that are looking for. For example, in Figure 5 it is shown that SWN synsets defining notions „teroristički napad“ (terrorist attack) and „upad“ (intrusion) have negative sentiment polarity scores (0.75 and 0.125) respectively, which makes possible classify texts containning these terms as „forensic texts“.

[19] Ranka Stanković, Cvetana Krstev, Ivan Obradović, Biljana Lazić, and Aleksandra Trtovac, "Rule-based Automatic Multi-word Term Extraction and Lemmatization", Proceedings of the 10th International Conference on Language Resources and Evaluation, LREC 2016, Portorož, Slovenia, 23--28 May 2016, 2016, eds. Nicoletta Calzolari et al., ISBN 978-2-9517408-9-1

[20] Mladenović, M., & Mitrović, J. (2014). Semantic Networks for Serbian: New Functionalities of Developing and Maintaining a WordNet Tool. In G. Pavlović Lažetić, C. Krstev, I. Obradović & D. Vitas Natural Language Processing for Serbian – Resources and Application, 1-11. Matematički fakultet, Beograd.

21 Mladenović, M., Mitrović, J., Krstev, C., & Vitas, D. (2015). Hybrid Sentiment Analysis Framework For A Morphologically Rich Language. Journal of Intelligent Information Systems, Volume 46, Issue 3, pp 599–620



ID: ENG30-01246697-n POS: n BCS: 0.000 0.750 Mama
08.02.2004 Approved: yes PWN XML Izmeni sinset

Literals: teroristički napad (1)
Definition: *Iznenadni napad koji uključuje namernu upotrebu nasilja prema civilima da bi se postigli politički ili religiozni ciljevi.*

- Relations... hypernym-> ENG30-01246541-n, prepad
SUMO: TerroristAttack =
DOMAIN:

ID: ENG30-00976953-n POS: n BCS: 3 0.000 0.125 User
28.06.2004 Approved: yes PWN XML Izmeni sinset

Literals: upad (1), iznenadni napad (1)
Definition: *Iznenadni kratki napad.*

- Relations... hypernym-> ENG30-00975452-n, prodiranje
- Relations... hyponym-> ENG30-00974111-n, vazdušni napad, napad iz vazduha
SUMO: Raid =
DOMAIN: military

Figure 5. *WordNer interface*

CONCLUSION

The paper presented the digital library from criminalistics domain available at http://master-kpa.rgf.rs/, as a document collection organised in several categories: Journal for criminalistics and right, Archibald Reiss, Doctoral Dissertation, and other (final process). Various methods for improvement of keyword based simple search is demonstrated on text prepared for text analysis and terminology extraction. Implementation details of Omeka, including add-in customisation, integration with Vebran, LeXimir and Unitex is discussed and presented n few examples. Having in mind that the metadata are a core for the development of digital libraries, that explains, locates, or otherwise makes it easier to retrieve, use or manage an information resource, metadata classification and management of metadata is elaborated. Paper concludes with Sentiment Analysis as a next step in a study of forensic texts.

REFERENCES

1. Cvetana Krstev. Processing of Serbian – Automata, Text and Electronic Dictionaries, Faculty of philology, Belgrade, 2008.
2. Cvetana Krstev, Duško Vitas, “Corpus and Lexicon - Mutual Incompletness ”, in Proceedings of the Corpus Linguistics Conference, 14-17 July 2005, Birmingham, eds. Pernilla Danielsson and Martijn Wagenmakers, ISSN 1747-9398, http://www.corpus.bham.ac.uk/PCLC/, 2005.
3. Cvetana Krstev, Ranka Stanković, Duško Vitas, Ivan Obradović, “The Usage of Various Lexical Resources and Tools to Improve the Performance of Web Search Engines”, in Proceedings of the Sixth Interantional Conference on Language Resources and Evaluation (LREC'08), Marrakech, Morocco, 28-30 May 2008, European Language Resources Association (ELRA), 2008
4. Duško Vitas, Cvetana Krstev, Ivan Obradović, Ljubomir Popović, Gordana Pavlović-Lažetić”,An Processing Serbian Written Texts: An Overview of Resources and Basic Tools ”, in Workshop on Balkan Language Resources and Tools, 21 Novembar 2003, Thessaloniki, Greece, eds, S. Piperidis and V. Karkaletsis, pp. 97-104, 2003.
5. Miljana Mladenović, Jelena Mitrović, Cvetana Krstev, “Developing and Maintaining a WordNet: Procedures and Tools”, In The Proceedings of Seventh Global WordNet Conference 2014, eds. Heili Orav, Christiane

Fellbaume, Piek Vossan, University of Tartu, Tartu, Estonia, January 25-29, 2014, pp. 55-62, 2014, ISBN 978–9949–32–492–7

6. TRTOVAC, Aleksandra S. Deskriptori metapodataka i sadržaja u pronalaženju informacija u digitalnim bibliotekama. 2016. PhD Thesis. Univerzitet u Beogradu-Filološki fakultet.
7. Ranka Stanković, Cvetana Krstev, Ivan Obradović, Biljana Lazić, and Aleksandra Trtovac, “Rule-based Automatic Multi-word Term Extraction and Lemmatization”, Proceedings of the 10th International Conference on Language Resources and Evaluation, LREC 2016, Portorož, Slovenia, 23--28 May 2016, 2016, eds. Nicoletta Calzolari et al., ISBN 978-2-9517408-9-1.
8. HODGE, G., 2001. Metadata made simpler, Niso Press.
9. MILENKOVIĆ, M., 2003. Dublin Core Metadata Initiative (DCMI). Review of the National Center for Digitization, 70-79.
10. TRTOVAC, A. S., 2016. Deskriptori metapodataka i sadržaja u pronalaženju informacija u digitalnim bibliotekama. Univerzitet u Beogradu-Filološki fakultet.
11. John Olsson (2008). Forensic Linguistics, Second Edition. London: Continuum ISBN 978-0-8264-6109-4
12. John Olsson (2004). An Introduction to Language Crime and the Law. London: Continuum International Publishing Group
13. C.S. Michell (2013). Investigating the use of forensic stylistic and stylometric techniques in the analysis of authorship on a publicly accessible social networking site (Facebook) (MA in Linguistics thesis). University of South Africa
14. C. Hardaker (2015). The ethics of online aggression: Where does “virtual” end, and “reality” begin? BAAL Conference on The Ethics of Online Research Methods. Cardiff
15. Falzini, Mark W. (9 September 2008). "The Ransom Notes: An Analysis of Their Content & "Signature""
16. https://omeka.org/
17. https://www.gnu.org/licenses/gpl-3.0.en.html
18. Ranka Stanković, Cvetana Krstev, Ivan Obradović, Biljana Lazić, and Aleksandra Trtovac, “Rule-based Automatic Multi-word Term Extraction and Lemmatization”, Proceedings of the 10th International Conference on Language Resources and Evaluation, LREC 2016, Portorož, Slovenia, 23--28 May 2016, 2016, eds. Nicoletta Calzolari et al., ISBN 978-2-9517408-9-1