Impressum

ISSN 1450-9687 (print edition)
ISSN 2217-9461 (online edition)

Belgrade, Vol. 21, No. 1, September 2021

JOURNAL REDACTION:
Journal Info*theca*
11000 Belgrade, Bulevar kralja Aleksandra 71

+381 11 3370–211
infothecajdh@gmail.com

PRINTED BY:
Mamigo plus
Belgrade

Journal is published twice a year

Contents

Scientific papers

The Enhanced Versions of the Program *Ka minimalnim parovima* (Towards Minimal Pairs)
D. Aleksić & L. Mrkela 7

Application of TXM tools for spatial plan corpus analysis
M. Milinković 32

Digital Resources about Aviation History
Lj. Macura 52

Professional papers

Europeana Data Model: The Key to Cooperation with Europeana
T. Domazet, N. Petrović & T. Butigan Vučaj . . 71

Multimedia document "Welcome to my favorite corner"
V. Ivanović, M. Čeketić, J. Stojičić & D. Galić . 86

Reviews

"Digital transformation and Libraries in Special Circumstances" – conference report
O. Krinulović & V. Petrović 96

Review of the Belgrade European Language Grid Workshop
B. Bašaragin 99

Distant Reading Training School – *Exploring ELTeC: Use-Cases for Information Extraction and Analysis*
A. Marković 103

The Enhanced Versions of the Program *Ka minimalnim parovima* (Towards Minimal Pairs)

UDC 81'322.2
DOI

ABSTRACT: The program KaMP finds word pairs whose members are segmentally (in terms of speech) different only by two selected factors (Deza and Deza 2016, 215), each factor with length 1 or more, e.g. *pêć* ∼ *pêt*, *f̆ilma* ∼ *f̂irma*, *istòrizovati* ∼ *majòrizovati*, *pësničkī* ∼ *polìtičkī*. The paper introduces the faster variants of KaMP with improved sorting and with a supplementary mode.

KEYWORDS: phonetics, phonology, natural language processing, corpus linguistics, Python.

PAPER SUBMITTED: 11 January 2022
PAPER ACCEPTED: 11 May 2022

Danilo Aleksić
danilo.aleksic@fil.bg.ac.rs
University of Belgrade
Faculty of Philology
Belgrade, Serbia

Lazar Mrkela
lazar.mrkela@metropolitan.ac.rs
Belgrade Metropolitan University, Faculty of Information Technologies
Belgrade, Serbia

1 Introduction

According to (Bugarski 2003, 128), minimal pairs are pairs in which two semantically distinct words formally differ in one phoneme only, e.g *bäs* ∼ *čäs*.[1] Ignoring prosody and letter case, in a Serbian corpus, the program Ka minimalnim parovima (Towards Minimal Pairs; Алексић and Шандрих 2021) finds word pairs whose members formally differ from each other by selected substrings[2] only. The corpus needs to be UTF-8 encoded. Apart

1. Ivić (1961–1962, 75) mentions prosodic systems with thousands of minimal pairs of words ie. pairs of forms differentiated by prosodic contrasts exclusively. Such pairs are *v̏ila* ∼ *vîla*, *lôza* (the genitive form of *loz* 'lottery ticket') ∼ *lòza* etc.

2. Globally it holds true that every string is a substring of itself (Partee, Meulen, and Wall 1993, 433; Singh 2009, 33) and that a string can be of length 1 (Partee, Meulen, and Wall 1993, 432; Python 2021b). Therefore, it would not be a misnomer to refer as a substring to (i) the string `"ima"` from the regular expression `\b\S*ima\S*\b` observed in relation to a match `"ima"`, nor to (ii) the character `"a"`.

from the selected substrings, the "words" can contain (i) characters from "A" to "Z", from "a" to "z" and from "Ć" to "ž" in the corresponding Unicode charts and (ii) hyphens in medial position.

Content of the input file	**Selected substrings**	**String for the output**
"Klima-uređaji pre klima-uređaja"	"a", "i"	"klima-uređaja ∼ Klima-uređaji" (or "Klima-uređaji ∼ klima-uređaja")
"α-čestica, α-čestice, α-čestici"	"a", "e"	"čestica ∼ čestice"
"α-čestica, β-čestica"	"α", "β"	"α-čestica ∼ β-čestica"

Table 1. KaMP: Examples of input and output

The program can be of use to teachers of Serbian as a foreign language and to linguists (Алексић and Шандрих 2021, 574–75).

Field	**Selected substrings**	**Utilization**
Teaching Serbian as a foreign language	"c", "č"	Basis for a task in an exercise: "*C* or *č*? 1) Šta bi ti uradio, dragi čitao_e? Naše novine će poštovati svoje čitao_e. 2) [...]"
Derivatology	"auto", "samo"	Data on competition between the segments *auto-* and *samo-*.

Table 2. KaMP: Examples of utilization

In the present paper the authors are publishing and commenting on the improved versions of KaMP which they built, KaMP 2 and KaMP 2.1.[3]

3. V. Appendix 1.

KaMP 2 and KaMP 2.1 are natural language processing tools if natural language processing is taken "in a wide sense to cover any kind of computer manipulation of natural language" (Bird, Klein, and Loper 2009, ix), because the two programs do not accent even a small number of all Serbian words, but process Serbian language superficially. Being somewhat adapted to searching large corpora, the new KaMPs are modest contributions to corpus linguistics as well if it is defined e.g. as "the computer-aided analysis of very extensive collections of transcribed utterances or written texts" (McEnery and Hardie 2012, i).

KaMP 2 and KaMP 2.1 were coded in Python 3.8.2 (Python 2021a). Python is "a high-level, interpreted, general-purpose programming language" (Pajankar 2020, 52). This language is "both elegant and pragmatic, both simple and powerful"; "it's suitable for programming novices as well as great for experts, too" (Martelli, Ravenscroft, and Holden 2017, ix). Python "is becoming more and more popular, and in 2017 it became the most popular language in the world according to IEEE Spectrum" (Shovic and Simpson 2021, 1). Python is "the most widely used language for natural language processing" (Antić 2021, vii). It "may be expected" that Python be "*slow* as compared to compiled languages", but it is faster "[i]f you start the clock to account for developer time, not just code runtime" (Unpingco 2021, 2). Python was created by the Dutch programmer Guido van Rossum in the late 1980s (Cicolani 2021, 41; Rajagopalan 2021, 1).

2 Similar resources

Four tools for finding minimal pairs or "phonological neighbours" (Mairano and Calabrò 2016, 258) which were written before KaMP are listed in (Алексић and Шандрих 2021, 569). The Python 3 package Minpair (PyPI 2021) and the short program in Python 2.7 from the page (Stack Overflow 2021a) can be added to that list.

Minpair looks for "minimal pairs (and minimal sets) for [only monosyllabic – D. A.] US English words". The user selects two or more "vowel phonological element[s]" by which the members of the minimal pairs or minimal sets will differ. Minpair (A) uses `defaultdict` to group the words by the accompanying transcriptions in which (B) the package replaced the chosen vowels with a dot by means of a regular expression and the `enumerate()` function.

The approach A has a general parallel, but somewhat more efficient (v. Appendix 2), in KaMP 2.1. KaMP 2.1 pairs words by means of a standard dictionary (v. Appendix 1).

The approach B has a general parallel, but much more efficient (v. Appendix 3), in KaMP 2 and KaMP 2.1. KaMP 2 and KaMP 2.1 replace the selected elements with the special string by means of the methods `str.replace()` and `str.format()` (v. Appendix 1).

`cmudict` **entry**	**Tuple for grouping**
`("cat", ["K", "AE1", "T"])` `("coat", ["K", "OW1", "T"])`	`("K", ".", "T")`

Table 3. Minpair: An example of input and of the tuple for grouping (if the selected vowels are `"AE"` and `"OW"`)

```
# Minpair: Examples of use 1 and 2
import minpair
print(minpair.vowel_minpair(["AO", "ER"])[12:13])
# Output: [{'AO': 'saw', 'ER': 'sir'}]

print(minpair.vowel_minpair(["AA", "AO", "EH"])[6:7])
# Output: [{'AO': 'dawn', 'EH': 'den', 'AA': 'don'}]
```

Part(s) of speech can also be chosen by the user.

```
# Minpair: Examples of use 3, 4 and 5
import minpair
print(minpair.generator(pos=["ADV"]).vowel_minpair(
        ["AH", "EH"]))
# Output: [{'AH': 'once', 'EH': 'whence'}]

print(minpair.generator(
    pos=["ADJ", "VERB"]).vowel_minpair(
    ["AE", "IH"])[:1])
# Output: [{'AE': 'bad', 'IH': 'bid'}]

print(minpair.generator(
    pos=["ADJ", "VERB"]).vowel_minpair(
    ["AE", "IH"])[22:23])
# Output: [{'AE': 'sang', 'IH': 'sing'}]
```

The word source(s) cannot be chosen by the user. Minpair "depends on a few NLTK's corpora, namely: *brown*, *cmudict*, *universal_tagset*, and *words* corpus".

The code from the page (Stack Overflow 2021a) pairs strings which differ by one character and have the same length. The strings must be inside e.g. a list, but they do not have to meet any natural language conditions (they do not have to come from a specific language or be written in a specific script, or even consist of alphabetic characters). During the execution of the code, every input string is compared to every subsequent input string, character by character. If all characters but one are the same, the pair of strings is printed.

```
# (Stack Overflow 2021a)
for n1,word1 in enumerate(wordlist):
    for word2 in wordlist[n1+1:]:
        if len(word1)==len(word2):
            ndiff=0
            for n,letter in enumerate(word1):
                if word2[n]!=letter:
                    ndiff+=1
            if ndiff==1:
                print word1, word2

"""The program from the page (Stack Overflow 2021a): Example of use 1
(added by D. A.)
Input: ["kula", "kule", "kuli", "kulom"]
Output:
kula kule
kula kuli
kule kuli
"""
```

This code is case sensitive in handling all characters except the differential one.

```
"""The program from the page (Stack Overflow 2021a):
Examples of use 2 and 3
Input: ["kula", "kulE", "kuli", "kulom"]
Output:
kula kulE
kula kuli
kulE kuli

```

```
Input: ["kula", "Kule", "kuli", "kulom"]
Output:
kula kuli
"""
```

When the first string in the input list was followed by its duplicate, the same pair was printed twice.

```
"""The program from the page (Stack Overflow 2021a): Example of use 4
Input: ["kula", "kula", "kule", "kulom"]
Output:
kula kule
kula kule
"""
```

	Minpair	**(Stack Overflow 2021a)**	**KaMP 2 and KaMP 2.1**
The tool prints such string pairs in each of which the strings differ by **any** character in the given position.	×	✓	×
The differential elements are chosen by the user.	✓	×	✓
The number of selected differential elements does not have to be 2.	✓		×
The selected differential elements do not have to be vowels.	×		✓
The words do not have to differ by a sequence of only one phoneme or of only one character.	×	×	✓
The input is chosen by the user.	×	✓	✓
The input does not have to be tokenized.		×	✓

Table 4. KaMP 2 / KaMP 2.1 compared to similar tools

3 Notable new characteristics of KaMP 2 and KaMP 2.1

In **the preparatory part of the algorithm**, tuples for the comparison of words are formed, such as ("Avali", "avali", "▼v▼li"), ("Požeškom", "požeškom", "pož▼škom"), ("sekretarijata", "sekretarijata", "s▼kr▼t▼rij▼t▼"). The first member of the comparison tuple is a word that contains one or both selected substrings. The second member is the casefolded first member. The third member is the second member in which at least one instance of the first selected substring or of the second selected substring has been replaced with the string "▼"[4]. One excerpted word can have one word with the replacement of the selected substrings (see Table 5) or, (i) when the selected substrings have an "overlap" (Lothaire 2005, 7) or (ii) when one selected substring is "a proper substring" (Böckenhauer and Bongartz 2007, 24) of the other selected substring, more than one word with the replacement of a selected substring (see Table 6).

	Comparison strings	
Excerpted word	**Casefolded excerpted word**	**Casefolded excerpted word in which the selected substrings were replaced**
"Knjiga"	"knjiga"	"knjig▼"
"knjigu"	"knjigu"	"knjig▼"
"sveska"	"sveska"	"svesk▼"
"SVESKU"	"svesku"	"svesk▼"
"računaljku"	"računaljku"	"r▼č▼n▼ljk▼"

Table 5. KaMP 2 / KaMP 2.1: Examples of comparison strings (if the selected substrings are "a" and "u")

In **the main part of the algorithm**, the formed tuples are compared. KaMP 2 generates all possible two-member combinations of the tuples with the first selected substring and the tuples with the second selected substring

4. It was chosen because it is conspicuous and relatively rare. Of course, those traits are present in many other strings.

	Comparison strings	
Excerpted word	**Casefolded excerpted word**	**Casefolded excerpted word in which one of the selected substrings was replaced**
"ONA"	"ona"	"on▼"
"Onima"	"onima"	"on▼"
"Onima"	"onima"	"onim▼"
"onimima"	"onimima"	"onim▼"
"onimima"	"onimima"	"onimim▼"

Table 6. KaMP 2 / KaMP 2.1: Examples of comparison strings (if the selected substrings are "a" and "ima")

and then skips the unwanted combinations. This program obtains the possible two-member combinations by calculating the Cartesian product[5] of the two groups of tuples. KaMP 2.1 transforms the tuples with the second selected substring into a hash map (table) and checks whether it contains the words with replacement taken from the tuples with the first selected substring. The map's key is the word with replacement, and the value of the map is a map of the words from which that same key is obtained. – KaMP 2 and KaMP 2.1 print those pairs of excerpted words whose members (i) differ when casefolded and (ii) match by the words with replacement, i.e. pairs of those words which differ by the selected substrings only. For example, if the selected substrings are "a" and "u", and the input file only contains the string "Knjiga, knjigu, sveska, SVESKU", KaMP 2 and KaMP 2.1 will not print the strings "Knjiga ∼ SVESKU" and "knjigu ∼ sveska", since the string "knjig▼" is not equal to the string "svesk▼". The programs will print the strings "Knjiga ∼ knjigu" and "sveska ∼ SVESKU".

KaMP 2 and KaMP 2.1 have (A) a mode in which they ignore case but favor strings of lowercase letters and (B) a mode in which e.g. excerpted strings "vitraž" and "Vitraž" would be processed as separate words (see Table 7). The reason is the justified comment from (Алексић and Шандрих 2021, 574) which raises the question of the importance of case. For example,

5. For example, the Cartesian product of the set of strings {"broj", "ulica"} and the set of strings {"MESTO", "OPŠTINA"} is the set of tuples of strings {("broj", "MESTO"), ("broj", "OPŠTINA"), ("ulica", "MESTO"), ("ulica", "OPŠTINA")}.

in teaching Serbian as a foreign language proper nouns sometimes have priority over non-proper words. The name *Čak* (*Beri*, *Noris*...) is suitable for a pronunciation exercise with photos; what kind of photo would depict the meaning of the uninflected word *čak*? In the mode B, from the corpus `POL`, KaMP 2 and KaMP 2.1 extract not only the pair `"čak ∼ Žak"`, but also the pair `"Čak ∼ Žak"` (alongside `"Čak ∼ ŽAK"` etc.).[6]

Content of the input file	`"EUPRAVE, eUprave, euprave, EUPRAVA, eUprava, euprava"`
String for the output of KaMP	`"EUPRAVA ∼ EUPRAVE"` (or `"EUPRAVE ∼ EUPRAVA"`)
String for the output of KaMP 2 and KaMP 2.1 in the mode A	`"euprava ∼ euprave"`

Strings for the output of KaMP 2 and KaMP 2.1 in the mode B	`"euprava ∼ euprave",`
	`"euprava ∼ eUprave",`
	`"euprava ∼ EUPRAVE",`
	`"eUprava ∼ euprave",`
	`"eUprava ∼ eUprave",`
	`"eUprava ∼ EUPRAVE",`
	`"EUPRAVA ∼ euprave",`
	`"EUPRAVA ∼ eUprave",`
	`"EUPRAVA ∼ EUPRAVE"`

Table 7. KaMP and KaMP 2 / KaMP 2.1: Case (in)sensitivity (if the selected substrings are `"a"` and `"e"`)

The function `corpus_segmentation()`[7] has been reorganized. It no longer reads the corpus using an infinite `while`-loop,[8] but, following the

6. Capital first letter is no guarantee that `"Čak"` can be a name in `POL` (because of sentences like `"Čak sam pronašao i kupca."`). An additional search proves that it can (`"Čak Blekvel"`, `"Čak Dejli"`, `"Čak Noris"`...).

7. This is a generator function which partitions the input corpus so that little RAM is used, in such a way as not to cut individual words apart (Алексић and Шандрих 2021, 572, 581).

8. Cf. the code which was added on May 14th 2021 to the answer which was posted to (Stack Overflow 2021b) on June 11th 2015.

example of the recommended way to call a function until a sentinel value from (Hettinger 2021, 12.27 and onwards), using a `for`-loop, which is "fast and beautiful".

KaMP sorts the found pairs by the Unicode code positions of single letters, while KaMP 2 and KaMP 2.1 sort the found pairs by the positions of single letters in the strings `lower_alphabet` and `upper_alphabet` (see Table 8).

```
# The sorting strings in KaMP 2 and KaMP 2.1
lower_alphabet = "- ~abcčćdđefghijklmnopqrsštuvwxyzž"
upper_alphabet = "- ~ABCČĆDĐEFGHIJKLMNOPQRSŠTUVWXYZŽ"
```

Input list	["nota ∼ note", "đaka ∼ đake", "Bač ∼ Beč"]
Input list after being sorted in the way KaMP sorts pairs	["Bač ∼ Beč", "nota ∼ note", "đaka ∼ đake"]
Input list after being sorted in the way KaMP 2 and KaMP 2.1 sort pairs	["Bač ∼ Beč", "đaka ∼ đake", "nota ∼ note"]

Table 8. KaMP and KaMP 2 / KaMP 2.1: Sorting of the pairs

Truth be told, the new KaMPs use the Unicode code positions as well when they sort pairs, but only for characters which the sorting strings do not contain (see Table 9).

Pair	Sorting list	The origin of the number
α	945	Unicode
-	0	The string
z	32	`lower_alphabet`
r	23	
a	3	
č	6	
e	10	
n	19	
j	15	
e	10	
	1	
∼	2	
	1	
β	946	Unicode
-	0	The string
z	32	`lower_alphabet`
r	23	
a	3	
č	6	
e	10	
n	19	
j	15	
e	10	

Table 9. KaMP 2 / KaMP 2.1: An example of the sorting list

KaMP 2 and KaMP 2.1 sort the members of every pair before joining them into an output string (e.g. `["knjigu", "Knjiga"]` → `["Knjiga", "knjigu"]`).

4 Execution speed

Speed was measured in Python 3.8.2, on Manjaro Linux, with a computer with the processor i5-11600K and two DDR4-3200 CL16 SDRAMs (16 GB

each) and on the POL corpus, which has around 117,900,900 words from 223,308 texts from the *Politika* website (Алексић and Шандрих 2021, 575).[9]

5 A short assessment of the efficiency of KaMP 2[10]

The function which finds pairs in KaMP 2 is based on the Cartesian product of two lists. This approach is an elegant solution in terms of layout and complexity of the code, but it is not efficient enough in the case of lists with large numbers of elements, because of quadratic behavior. The problem is easily noticed in the experimental results, where significantly longer execution time is observed in the case of more frequent substrings (ma-va; cf. Table 10).

6 Further work

In real conditions, which may demand that this program be run on weaker computers, just reading the corpus and excerpting the words which contain the selected substrings can last for too long. For example, reading the corpus POL.xml on an older laptop computer (Acer Aspire 3, Intel Quad Core N3710, 4GB RAM) lasts up to approximately 15 minutes. The recommendation is that the option to create a disk-stored dictionary be added. This processing of the corpus would be conducted just once, and the dictionary would later be used to search for pairs by new substrings.

The next possible step in shortening the execution time is search parallelization. Nowadays, even the weaker computers have several "cores" in their processors (for example, the computer from the previous paragraph has 4 cores). That is why it is possible to execute some parts of the code in parallel and thus additionally speed up the program. A suggestion for simple parallelization is the division of one of the lists into n parts, and then the processing of those parts on separate processors (cores) in parallel. Seeing that the sequential version with hashing is already very efficient, the problem of the slow reading of the corpus should be solved first, and only then should further speeding up be considered.

9. All the results come from measuring the execution speed of the code from the Serbian version of this paper.

10. Sections 5 and 6 were written by L. Mrkela. The other sections (without the two sentences in Section 3 which only concern KaMP 2.1), Appendix 2 and Appendix 3 were written by D. Aleksić.

Selected substrings	Speed in seconds (the average of five successive measurements)				
	KaMP	KaMP 2		KaMP 2.1	
		Mode A	Mode B	Mode A	Mode B
"č", "đ"	334	199	246	67	66
"dž", "đ"	135	81	85		
"nadnad", "supersuper""đ"	6639	**65.79**	**65.51**	66.38	66.09
"ir", "zir"	143	86	90	**67** (!)	66
"ma", "va"	2153	1257	1516		

Table 10. KaMP, KaMP 2 and KaMP 2.1: Execution speed

7 Conclusion

In comparison to KaMP, KaMP 2 and KaMP 2.1 achieve more in less time – they find virtually the same pairs and sort the pairs and the words inside the pairs in a better way.

KaMP 2.1 was inarguably faster than KaMP 2 in the majority of the investigated cases.

Appendix 1. KaMP 2 and KaMP 2.1[11]

```
"""KaMP 2.1 is a modified version of KaMP 2. KaMP 2 is
a modified version of KaMP.

The pairing in the functions KaMP_2_1_a() and KaMP_2_1_b()
was written by L. Mrkela, and the rest of the code
(with the functions excerp_(), proc_words_1() and proc_words_2())
was written by D. Aleksić.
"""


def main():
    from functools import partial
    from itertools import product
    import re
    import sys

    sys.stdout.reconfigure(encoding="utf-8")
    """V. (Алексић and Шандрих 2021, 580)."""
    letter_1 = "ma".casefold()
    letter_2 = "va".casefold()
    overlap_ = False
    case_diff = False
    sort_char = chr(1114111)
    lower_alphabet = "- ~abcčćdđefghijklmnopqrsštuvwxyzž"
    upper_alphabet = "- ~ABCČĆDĐEFGHIJKLMNOPQRSŠTUVWXYZŽ"

    def join_strings(*strings):
        return "".join(strings)

    def corpus_segmentation(
            corpus_, size_=8192, separator_="\n"):
        """Cf. (581).
        The function returns parts of the corpus which have
        the specified size. The cuts between the parts are
        made at the specified separator.
        """
```

11. The official documentation, from the site (Python 2021a), was the primary source. The other sources were indicated by directing to links from the References, by directing to the corresponding parts of (Алексић and Шандрих 2021) and by directing to a part of this paper.

```
        remainder_ = ""
        for piece_ in iter(
                partial(corpus_.read, size_), ""):
            """V. (Hettinger 2021, 12.27 and onwards)."""
            piece_ = join_strings(remainder_, piece_)
            if separator_ in piece_:
                pieces_ = piece_.rsplit(separator_, 1)
                """V. (W3Schools 2021)."""
                yield pieces_[0]
                remainder_ = pieces_[1]
            else:
                remainder_ = piece_
        if remainder_:
            yield remainder_

    def lower_first_1(word_):
        """The key for word sorting which favors the words
        consisting of lowercase letters.
        """
        if word_.islower():
            return "!"
        elif word_.istitle():
            return sort_char
        else:
            for letter_ in word_:
                if letter_.isupper():
                    word_ = word_.replace(
                        letter_, sort_char)
            return word_

    def lower_first_2(word_):
        """The key for sorting the found pairs.
        The pairs which contain less uppercase letters will
        be at the top of the list.
        """
        if word_.islower():
            return "!"
        else:
            word_ = word_.replace(" ~ ", "")
            if word_.istitle():
                return sort_char
            else:
                for letter_ in word_:
```

```
                    if letter_.isupper():
                        word_ = word_.replace(
                            letter_, sort_char)
                return word_

    def indexing_for_list(word_):
        """Cf. (Stack Overflow 2021c).
        The key for sorting in the specified order.
        """
        sort_list = []
        for letter_ in word_:
            if letter_ in lower_alphabet:
                sort_list.append(
                    lower_alphabet.index(letter_))
            elif letter_ in upper_alphabet:
                sort_list.append(
                    upper_alphabet.index(letter_))
            else:
                sort_list.append(ord(letter_))
        return sort_list

    def simple_word_repl(word_):
        """The selected substrings of words are replaced by
        the special string.
        """
        if letter_1 in word_ and letter_2 not in word_:
            word_with_repl = word_.replace(
                letter_1, "\u23B2")
        elif letter_1 not in word_ and letter_2 in word_:
            word_with_repl = word_.replace(
                letter_2, "\u23B2")
        elif letter_1 in word_ and letter_2 in word_:
            word_with_repl = word_.replace(
                letter_1, "\u23B2")
            word_with_repl = word_with_repl.replace(
                letter_2, "\u23B2")
        return (word_with_repl,)

    def complex_word_repl(word_, letter_):
        """The specified substrings in words are replaced
        with the special string.
        The cases when there is overlap between the selected
        substrings are covered.
```

```
123         V. (Алексић and Шандрих 2021, 580--81).
124         """
125         output_set = set()
126         word_for_proc = word_.replace(letter_, "{}")
127         for combination_ in product(
128                 [letter_, "\u23B2"],
129                 repeat=word_for_proc.count("{}")):
130             word_with_repls = word_for_proc.format(
131                 *combination_)
132             if word_with_repls != word_:
133                 output_set.add(
134                     word_with_repls)
135         return output_set
136
137     def letter_repl(word_):
138         if not overlap_:
139             return simple_word_repl(word_)
140         else:
141             set_ = set()
142             set_.update(
143                 complex_word_repl(word_, letter_1),
144                 complex_word_repl(word_, letter_2))
145             return set_
146
147     def tokenization_():
148         """The corpus is transformed into a dictionary of
149         words which were gathered by means of a regular
150         expression. The function enables the program to
151         avoid the use of very complex regular expressions
152         in the cases when the selected substrings are
153         long (see Table 10).
154         Cf. Section 6.
155         """
156         dict_ = {}
157         with open(r"/home/.../POL.xml",
158                   "r", encoding="utf-8") as corpus_:
159             pieces_ = corpus_segmentation(corpus_)
160             for piece_ in pieces_:
161                 matches_ = re.findall(
162                     "[A-Za-zĆ-ž-\u00ad]+", piece_)
163                 """V. (573--74)."""
164                 for match_ in matches_:
165                     word_ = match_.strip("-")
```

```
166                     if "\u00ad" in word_:
167                         word_ = word_.replace(
168                             "\u00ad", "")
169                         """V. (Алексић and Шандрих 2021, 581)."""
170                     dict_[word_] = word_.casefold()
171         return dict_
172
173     def excerp_(letter_):
174         """The words which contain the selected substring
175         are taken from the dictionary into which the corpus
176         has been transformed.
177         """
178         return (key_
179                 for key_, value_
180                 in corpus_dict.items()
181                 if letter_ in value_)
182
183     def descartes(list_1, list_2):
184         """The unwanted pairs are eliminated from the
185         Cartesian product of the processed words.
186         """
187         return (
188             (*sorted([b, e]), a, d)
189             for (a, b, c), (d, e, f)
190             in filter(
191             lambda tuple_: tuple_[0][2] == tuple_[1][2]
192                            and tuple_[0][1]
193                            != tuple_[1][1],
194             product(list_1, list_2, repeat=1)))
195
196     def proc_words_1(gen):
197         """The function returns tuples which contain words,
198         casefolded words and words with replacement.
199         This is a case insensitive function.
200         """
201         tuple_list = []
202         counter_ = set()
203         word_list = sorted(
204             list(gen), key=lower_first_1)
205         for word_ in word_list:
206             casefold_word = corpus_dict[word_]
207             if casefold_word not in counter_:
208                 counter_.add(casefold_word)
```

```
                for word_with_repl in letter_repl(
                        casefold_word):
                    tuple_list.append(
                        (word_, casefold_word,
                         word_with_repl))
        return tuple_list

    def proc_words_2(gen):
        """The function returns tuples which contain words,
        casefolded words and words with replacement.
        This is a case sensitive function.
        """
        tuple_list = []
        for word_ in gen:
            casefold_word = corpus_dict[word_]
            for word_with_repl in letter_repl(
                    casefold_word):
                tuple_list.append(
                    (word_, casefold_word,
                     word_with_repl))
        return tuple_list

    def KaMP_2_a():
        """Final case insensitive processing in KaMP 2.
        """
        counter_ = set()
        for tuple_ in descartes(
                proc_words_1(excerp_(
                    letter_1)),
                proc_words_1(excerp_(
                    letter_2))):
            if (tuple_[0], tuple_[1]) not in counter_:
                counter_.add((tuple_[0], tuple_[1]))
                final_set.add((tuple_[2], tuple_[3]))
        pair_list = [
            " ~ ".join(sorted(list(tuple_),
                              key=indexing_for_list))
            for tuple_ in final_set]
        pair_list.sort(key=indexing_for_list)
        for pair_ in pair_list:
            print(pair_)
        print("\n\tTHE NUMBER OF PAIRS:")
        print("\t\t", len(pair_list))
```

```
252
253     def KaMP_2_b():
254         """Final case sensitive processing in KaMP 2.
255         """
256         final_set = {(tuple_[2], tuple_[3])
257                         for tuple_ in descartes(
258                 proc_words_2(excerp_(
259                     letter_1)),
260                 proc_words_2(excerp_(
261                     letter_2)))}
262         pair_list = [
263             " ~ ".join(sorted(list(tuple_),
264                               key=indexing_for_list))
265             for tuple_ in final_set]
266         pair_list = list(set(pair_list))
267         pair_list.sort(key=lower_first_2)
268         pair_list.sort(key=indexing_for_list)
269         for pair_ in pair_list:
270             print(pair_)
271         print("\n\tTHE NUMBER OF PAIRS:")
272         print("\t\t", len(pair_list))
273
274     def KaMP_2_1_a():
275         """Case insensitive pairing and final processing
276         in KaMP 2.1.
277         """
278         list1 = proc_words_1(excerp_(
279             letter_1))
280         list2 = proc_words_1(excerp_(
281             letter_2))
282         map_ = {}
283         for x in list2:
284             if x[2] not in map_:
285                 map_[x[2]] = {}
286             map_[x[2]][x[0]] = x[1]
287         for tuple_ in list1:
288             result = map_.get(tuple_[2])
289             if result is not None:
290                 for k, v in result.items():
291                     if (tuple_[1] != v and (k, tuple_[0])
292                             not in final_set):
293                         final_set.add((tuple_[0], k))
294         list_ = []
```

```
        for pair_ in final_set:
            pair_ = list(pair_)
            pair_.sort(key=indexing_for_list)
            output_ = join_strings(
                pair_[0], " ~ ", pair_[1])
            list_.append(output_)
        list_.sort(key=indexing_for_list)
        for pair_ in list_:
            print(pair_)
        print("The number of pairs: ", len(list_))

    def KaMP_2_1_b():
        """Case sensitive pairing and final processing
        in KaMP 2.1.
        """
        list1 = proc_words_2(excerp_(
            letter_1))
        list2 = proc_words_2(excerp_(
            letter_2))
        map_ = {}
        for x in list2:
            if x[2] not in map_:
                map_[x[2]] = {}
            map_[x[2]][x[0]] = x[1]
        for tuple_ in list1:
            result = map_.get(tuple_[2])
            if result is not None:
                for k, v in result.items():
                    if (tuple_[1] != v and (k, tuple_[0])
                            not in final_set):
                        final_set.add((tuple_[0], k))
        list_ = []
        for pair_ in final_set:
            pair_ = list(pair_)
            pair_.sort(key=indexing_for_list)
            output_ = join_strings(
                pair_[0], " ~ ", pair_[1])
            list_.append(output_)
        list_.sort(key=lower_first_2)
        list_.sort(key=indexing_for_list)
        for pair_ in list_:
            print(pair_)
        print("The number of pairs: ", len(list_))
```

```
338
339     if (letter_1[-1:] == letter_2[:1]
340             or letter_1[:1] == letter_2[-1:]
341             or (letter_1 in letter_2
342                 or letter_2 in letter_1)):
343         overlap_ = True
344     final_set = set()
345     corpus_dict = tokenization_()
346     if case_diff:
347         KaMP_2_1_b()  # KaMP 2 calls the function KaMP_2_b().
348     else:
349         KaMP_2_1_a()  # KaMP 2 calls the function KaMP_2_a().
350
351
352 if __name__ == "__main__":
353     main()
```

Appendix 2. Minpair vs. KaMP 2.1: Pairing speed

```
1  """The Minpair approach.
2  """
3  from collections import defaultdict
4
5  map_1 = defaultdict(lambda: {})
6  for x in list_2:
7      map_1[x[2]][x[0]] = x[1]
8
9  """The KaMP 2.1 approach.
10 """
11 map_2 = {}
12 for x in list_2:
13     if x[2] not in map_2:
14         map_2[x[2]] = {}
15     map_2[x[2]][x[0]] = x[1]
16
17 """The input was a list of tuples like ("subsidiaries",
18 "subsidiaries", "subsidi.ri.s"). The list was made of
19 the words from cmudict which contain the substring "a"
20 and/or the substring "e".
21
22 The KaMP 2.1 approach proved itself around 7% faster in
```

```
Python 3.8.2 on the system which was described in
Section 4. The averages of 500 successive measurements
were compared (55 ms : 51 ms).
"""
```

Appendix 3. Minpair vs. KaMP 2 / KaMP 2.1: Replacement speed

```
"""The Minpair approach.
"""
vowels_regex = re.compile(r'^(?:%s)' % '|'.join(vowels))
matches = [vowels_regex.search(phone) for phone in word]
list_with_repl = []
for i, character in enumerate(word):
    for j, match in enumerate(matches):
        if i == j:
            if match:
                list_with_repl.append(".")
            else:
                list_with_repl.append(character)
string_with_repl = "".join(list_with_repl)

"""The KaMP 2 and KaMP 2.1 approach.
"""
word_with_repl = word_.replace(
    letter_1, ".")
word_with_repl = word_with_repl.replace(
    letter_2, ".")

"""The input consisted of the words from cmudict
which contain the substring "a" and/or the substring "e".

The KaMP 2 and KaMP 2.1 approach proved itself around
95% faster in Python 3.8.2 on the system which was described
in Section 4. The averages of 500 successive measurements
were compared (472 ms : 23 ms). However, it must be pointed out
that the code for replacing in Minpair gets a list and returns
a tuple (e.g. ["L", "UW", "S"] → ("L", ".", "S")), while both
the input and the output of the code for replacing in KaMP 2
and KaMP 2.1 are a string (e.g "teorijska" → "t▛orijsk▜").
"""
```

References

Antić, Zhenya. 2021. *Python Natural Language Processing Cookbook: Over 50 recipes to understand, analyze, and generate text for implementing language processing tasks.* Birmingham: Packt Publishing.

Bird, Steven, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python.* Sebastopol, CA: O'Reilly Media.

Böckenhauer, Hans-Joachim, and Dirk Bongartz. 2007. *Algorithmic Aspects of Bioinformatics.* Berlin: Springer.

Bugarski, Ranko. 2003. *Uvod u opštu lingvistiku.* 2nd ed. Beograd: Čigoja štampa.

Cicolani, Jeff. 2021. *Beginning Robotics with Raspberry Pi and Arduino: Using Python and OpenCV.* 2nd ed. Berkeley, CA: Apress.

Deza, Michel Marie, and Elena Deza. 2016. *Encyclopedia of Distances.* 4th ed. Berlin: Springer.

Hettinger, Raymond. 2021. "Transforming Code into Beautiful, Idiomatic Python." Accessed August 21, 2021. https://www.youtube.com/watch?v=OSGv2VnC0go.

Lothaire, M. 2005. *Applied Combinatorics on Words.* Cambridge: Cambridge University Press.

Mairano, Paolo, and Lidia Calabrò. 2016. "Are minimal pairs too few to be used in pronunciation classes?" In *La fonetica nell'apprendimento delle lingue: Phonetics and language learning,* edited by Renata Savy and Iolanda Alfano, 255–268. Milano: Officinaventuno.

Martelli, Alex, Anna Ravenscroft, and Steve Holden. 2017. *Python in a Nutshell.* 3rd ed. Sebastopol, CA: O'Reilly Media.

McEnery, Tony, and Andrew Hardie. 2012. *Corpus Linguistics: Method, Theory and Practice.* Cambridge: Cambridge University Press.

Pajankar, Ashwin. 2020. *Raspberry Pi Computer Vision Programming: Design and implement computer vision applications with Raspberry Pi, OpenCV, and Python 3.* 2nd ed. Birmingham: Packt Publishing.

Partee, Barbara H., Alice ter Meulen, and Robert E. Wall. 1993. *Mathematical Methods in Linguistics.* Dordrecht: Kluwer Academic Publishers.

PyPI. 2021. "minpair 0.1.3." Accessed October 26, 2021. https://pypi.org/project/minpair.

Python. 2021a. Accessed August 21, 2021. https://www.python.org.

Python. 2021b. "Text Sequence Type — `str`." Accessed August 21, 2021. https://docs.python.org/3/library/stdtypes.html#text-sequence-type-str.

Rajagopalan, Gayathri. 2021. *A Python Data Analyst's Toolkit: Learn Python and Python-based Libraries with Applications in Data Analysis and Statistics.* Berkeley, CA: Apress.

Shovic, John C., and Alan Simpson. 2021. *Python All-in-One For Dummies.* 2nd ed. Hoboken, NJ: John Wiley & Sons.

Singh, Arindama. 2009. *Elements of Computation Theory.* London: Springer.

Stack Overflow. 2021a. "Finding (phonological) minimal pairs with python." Accessed August 31, 2021. https://stackoverflow.com/q/26157361.

Stack Overflow. 2021b. "Lazy Method for Reading Big File in Python?" Accessed August 31, 2021. https://stackoverflow.com/q/519633.

Stack Overflow. 2021c. "Sorting string values according to a custom alphabet in Python." Accessed October 26, 2021. https://stackoverflow.com/q/26579392.

Unpingco, José. 2021. *Python Programming for Data Analysis.* Cham: Springer.

W3Schools. 2021. "Python String rsplit() Method." Accessed October 26, 2021. https://www.w3schools.com/python/ref_string_rsplit.asp.

Алексић, Данило, and Бранислава Шандрих. 2021. "Аутоматска ексцерпција парова речи за учење изговора у настави српског као страног језика." *Српски језик: студије српске и словенске* 26 (1): 567–584. ISSN: 0354-9259. https://doi.org/10.18485/sj.2021.26.1.32. http://doi.fil.bg.ac.rs/pdf/journals/sj/2021-1/sj-2021-26-1-32.pdf.

Ивић, Павле. 1961–1962. "Број прозодијских могућности у речи као карактеристика фонолошких система словенских језика." *Јужнословенски филолог* 25: 75–113.

Application of TXM tools for spatial plan corpus analysis

UDC 811.163.41'322.2
DOI

Milena Milinković
milenam@iaus.ac.rs
Institute of Architecture and Urban & Spatial Planning of Serbia
Belgrade, Serbia

ABSTRACT: The paper presents quantitative textometric research of the PP-TXM corpus of spatial plans in the TXM environment. Thanks to the application of developed methods and technologies in the field of corpus linguistics and SrpNER tools for marking named entities, different linguistic and statistical characteristics of the textual part of planning documents have been pointed out. Using the TXM tool during the research, the results were obtained that provided insight into the frequency of occurrence of different types of words and named entities, both in the whole corpus, and their distribution and progression by partitions of this corpus.

KEYWORDS: TXM, textometry, SrpNER, digital corpora, spatial plans.

PAPER SUBMITTED: 1 May 2021
PAPER ACCEPTED: 1 June 2022

1 Introduction

Quantitative approach to the analysis of texts requires observation of their statistical characteristics based on pre-defined criteria of classification and quantification (Bjekić et al. 2012). The classical approach of quantitative analysis of texts frequently required the cooperation of a large number of researchers but also hard and long work (Utvić 2013), which required significant material and time resources (Mehl and Gill 2010). With the advent of computers and the development of modern computer technologies, such analyzes have become easier. Regardless of whether it is a classical or computer-aided approach, any linguistic analysis requires the formation of adequate and representative corpora, so that research conducted on them

can be applied to other texts that meet pre-set criteria. In linguistics, the corpus can be described as "a collection of texts that are assumed to be representative of a given language, dialect or other subset of languages, and that will be used for linguistic analysis" (Francis 1975, 15). This kind of definition can be applied to each corpus, regardless of the information carrier. However, with the development of computational linguistics and the formation of electronic corpora, a more acceptable definition is "a collection of authentic machine-readable texts that represent a representative sample of a particular language or linguistic variety" (McEnery, Xiao, and Tono 2006, 5). Electronic corpora are, just like pre-electronic corpora, intended for language research, but unlike them, they are adapted to automatic or semi-automatic processing and analysis.

Corpora can be classified in different ways depending on the scope, the number of languages involved, the purpose, the domain, the media and other characteristics (Popović and Vitas 2003). One of the divisions of the corpus is into general-purpose and specialized corpora, respectively domain corpora The Corpus of the Modern Serbian Language, or *SrpKor* for short, is a general-purpose corpus. Its development began in 1981, and the latest version of *SrpKor2021*, was published in 2021. in the length of over 600 million words. This corpus also includes various domain corpora that have been formed over the years. Domain Corpus of Culinary Arts (Витас and Крстев 2016; Витас 2018; Vitas 2019, 2022), domain corpus of legal texts of the National Assembly of the Republic of Serbia (Васильевић 2014), domain corpus in the field of librarianship (Тртовац 2017) and domain corpus in the field of mining (Обрадовић et al. 2017) are just some of the specialized corpora developed over the years by members of the Society for Language Resources and Technologies - JepTex.[1] The last domain corpus formed within the Society, at the end of 2021, is the domain corpus of spatial planning. Research on the SrpNER and TXM tools was performed on a subset of this corpus, which will be presented in the following chapters.

2 TXM tool

TXM[2] is open source software (Heiden 2010) whose application is enabled in three different operating systems: 64-bit Windows, Mac OS X and 64-bit Linux. There are two versions of this software. The desktop version requires

1. Society for Language Resources and Technologies – JePTex
2. What is TXM?

the installation of TXM software, after which the user is allowed to import his corpus on his own computer and perform the desired analysis, while for the TXM version for web software, installation is not required, but the user accesses various online corpora.

If texts are going to be imported into the TXM environment, it is necessary for them to be in a certain format and adequate code layout. With the TXM tool it is possible to analyze documents in TXT plain text format or XML (*eXtensible Markup Language*) format. The first format must be in the recommended UTF-8 code layout, and the second, in addition to the specified code layout, must be in accordance with the TEI (*Text Encoding Initiative*)[3] instructions. The degree of complexity of the analysis, which can be performed in a TXM environment, depends on the level of text representation. Only basic analyzes can be performed on texts in TXT format, since they have the lowest level of representation. In contrast, texts encoded in XML-TEI format can have more or less rich representation, which directly affects the complexity of the analyzes that can be applied to texts.

TXM tool provides users with a variety of techniques for statistical processing of texts, and its working environment is extremely clear and easy to use which makes it suitable for use by both beginners and professionals engaged in statistical research. Since in the TXM environment it is possible to hierarchically organize the text objects on which the research is conducted, the analysis can be performed both on the whole corpus and on a separate subcorp or partition. This is possible because of the fact that corpora are divided into constituent units: textual, structural and lexical units. A textual unit is made of all texts contained in it, marked with the necessary metadata. The structural units of the corpus are chapters, paragraphs, sentences, etc. The lowest level of hierarchical organization consists of lexical units, ie the words of which the text consists (Jaćimović 2019). The results of the given queries are displayed as concordances, tables or graphs.

3 PPTXM corpus

For the purposes of this research, the PPTXM[4] corpus was formed, which consists of texts of spatial plans and is part of the domain corpus in the field of spatial planning. Planning documentation is a specific type of documents characteristic of the field of spatial planning. There are four types of spatial plans that are made on the territory of the state. These are the Spatial

3. TEI (Text Encoding Initiative).
4. Corpus of spatial plans formed for analysis in TXM environment.

Plan of the Republic of Serbia, the regional spatial plan, the spatial plan of local self-government units and the spatial plan of the special purpose area (Службени гласник РС 52/2021). Regardless of the plan in question, after making a decision on the development of the plan, it is essential to determine the scope of the plan, make an analysis of the current situation, and then in accordance with the conditions and guidelines specified in higher level planning documents. This whole process requires the cooperation of spatial planners and scientists and experts from many other fields (architecture, urbanism, demography, ecology, sociology, etc.). In the initial stages of drafting the plan, an elaboration for early public insight is prepared in order to acquaint the general public with the purpose and goals of drafting the plan, with potential concepts and solutions for the development of a specific area and the offered environmental plan. This is followed by preparatory activities, which include obtaining the basis for the development of the graphic part of the plan and obtaining the conditions and data necessary for the preparation of the draft planning document containing the textual and graphic part. The draft plan is subject to expert control that lasts 15 days, and after acting in accordance with the report on the performed expert control, the plan is presented to the public at the public inquiry. During this phase of the plan development, interested natural and legal persons, exclusively in writing, submit any objections to the planning document. After the drafting is completed, the plans are adopted by the competent institutions, and then the textual part of the plan is published in the appropriate official gazette (Службени гласник РС 32/2019).

The developed corpus consists of six spatial plans, four plans of the special purpose areas, one regional spatial plan and one spatial plan of the local self-government unit, namely: *Regional Spatial Plan for the Zlatibor and Moravica administrative districts*, *Spatial plan for the special purpose area of Đerdap National Park*, *Spatial plan for the special purpose area of Grlište Reservoir basin*, *Spatial plan for the special purpose area of Ćelije reservoir basins*, *Spatial plan for the special purpose area of international waterway E-80 – Danube (Pan-Europian Corridor VII)* and *Spatial plan for the municipality of Knjaževac*. The only spatial plan that is not included in this corpus is the *Spatial Plan of the Republic of Serbia*, however, bearing in mind that these plans cover different areas in Serbia, and that they cover all the diversity of geographical features of the Republic, it can be said that this corpus representative.

4 Textometry with TXM tool

Textometry is a methodology for analyzing textual data whose development was started by French theorists Pierre Guiraud, Charles Muller, Jean-Paul Benzécri and others, and it is precisely this that has found its application in quantitative linguistic research (MacMurray and Leenhardt 2011), but also in many other humanities and social sciences (Jaćimović 2019). The textometric analysis, which was applied to this corpus, will indicate the different linguistic and statistical characteristics of the textual part of the planning documents. Since this method was founded in the 1980s, thanks to a rich range of tools and software used first for processing and then analyzing natural languages, as well as a number of international standards for structuring data and text, it represents much more than just counting words (Pincemin, Heiden, and Decorde 2020). Different types of textometric analysis are enabled by TXM software. Some of the exemplified textometric analysis in the TXM environment were conducted on one of the domain corpora developed by the Society for Language Resources and Technology. This is the SrpELTeC[5] corpus (Trtovac, Milnović, and Krstev 2021), which at that time contained 21 prose works in the Serbian language published in the period 1840-1920. The motive for its creation was the inclusion of corpus texts in the multilingual collection of European literary texts (*European Literary Text Collection*). The obtained results provided insight into the great variety of analyzes that can be performed with TXM tools, while pointing out the specifics of using different types of words depending on the gender of the author and the frequency of use of certain lemmas in srpELTeC corpus texts (Jaćimović 2019).

5 SrpNER tool and marking of named entities

One of the many tasks of natural language processing is the recognition of Named Entity Recognition (NER). This type of text processing and analysis refers to the recognition of the names of persons, organizations, locations as well as various numerical expressions including percentages, money, time and dates. Recently, the technology of recognizing named entities has been applied to extracting the names of events, products, book titles and marking e-mail addresses. The development of this segment of natural language processing has been going on for almost three decades (Maurel, Friburger,

5. The complete collection available at SrpELTeC

and Eshkol-Taravella 2014), and the SrpNER system has been developed for many years to recognize named entities in the Serbian language within the Society for Language Resources and Technologies. The manner of work and use of the system for recognizing named entities in the Serbian language was described in detail by prof. Dr. Cvetana Krstev (Krstev et al. 2014).

Prior to any textometric analysis in the PPTXM corpus, the named entities were annotated with the SrpNER tool. For the purposes of this research, geopolitical terms (`top.deogr`, `top.dr`, `top.geo`, `top.gr`, `top.hyd`, `top.reg`, `top.supreg`, `top.ul`), demons (`demonym`) and names of organizations (`org.com`, `org.gen`, `org.pol`, `org.rel`) were tagged. In order for the TXM tool to perform corpus research, certain corrections were necessary, which meant that the dot from the tag for marking the named labels was replaced with an underscore, i.e. the tags were renamed so that the `org.com` tag would be replaced by the `org_com` tag, `top.gr` with `top_gr` label and so on. The only tag that has remained unchanged is the `demonym` tag (Table 1). In addition to the above, it was necessary to eliminate nested tags of named entities, which are produced by the SrpNER tool, and which are not supported in the TXM environment. The annotation of the organization *JKP "Vodovod" Zaječar* with the SrpNER tool was made as follows:

```
<org.com>
  <org.gen>JKP „Vodovod"</org.gen>
  <top.gr>Zaječar</top.gr>
</org.com>
```

Within the external tags used for marking commercial organizations `<org.com>` and `</org.com>` there are also nested tags for general organizations `<org.gen>` and `</org.gen>` which marked *JKP "Vodovod"* and tags for urban settlements or populated places `<top.gr>` and `</top.gr>` which mark *Zaječar*. By removing the nested tags, the full name of the organization is marked with the `<org.com>` tag.

6 Textometric processing and analysis of PPTXM corpora in TXM environment

The first step after importing the corpus into the TXM environment involved automatic segmentation, tokenization, lematization, and finally word type labeling (Jaćimović 2019). This is made possible by the TreeTagger[6]

6. TreeTagger - a part-of-speech tagger for many languages

TXM tags	Explanation of tags	Example
`demonym`	nouns denoting population, ethnic groups and adjectives derived from geographical names	*beogradski*
`org_com`	commercial organizations	*JKP "Vodovod" Zaječar*
`org_gen`	general organizations (unclassified organizations)	*Institute of Architecture and Urban & Spatial Planning of Serbia*
`org_pol`	political organizations	
`org_rel`	religious buildings (or organizations)	*Church of St. Mark*
`top_deogr`	part of urban settlements or inhabited areas	*Dedinje*
`top_dr`	country	*Serbia*
`top_geo`	geographical features (lowlands, mountains, hills, plateaus...)	*Avala*
`top_gr`	urban settlements or inhabited areas	*Belgrade*
`top_hyd`	Hydronyms (rivers, lakes, springs...)	*Danube*
`top_reg`	region within a state	*Podunavlje*
`top_supreg`	supranational region	*Europe*
`top_ul`	streets or city locations in general	*Knez Mihailova Street*

Table 1. TXM tags with explanations and examples

tool integrated into TXM software that associates each word in the text with the lemma and associated morphosyntactic categories. Such an annotated corpus is suitable for analysis by the *Corpus Query Processor* (CQP), which instead of a character as a search unit includes a corpus word, which provides the opportunity for various phraseological research.

In order to obtain basic corpus structure data within the TXM tool, a search was performed using CQP queries that resulted in a list of concordances. The query for paragraphs `/region[p]` resulted in 11,895 concordances, while the query `/region[s]` related to sentences yielded 17.063 concordance lines.

By tokenizing the texts, it was determined that there are 487,213 tokens in the PPTXM corpus (Table 2). Of that number, there are 32,565 different corpus words, and 14,903 different lemmas. When the punctuation marks are subtracted from the total number of tokens, the corpus consists of 404,218 words. The number of words as well as the number of tokens are not equally distributed among the partitions of this corpus (Figure 1, Table 2).

Name of the spatial plan	**Scope of the plan**	**Dimensions of partitions in number of tokens**
Regional Spatial Plan for the Zlatibor and Moravica administrative district	9.184 km^2	90.724
Spatial plan for the special purpose area of Đerdap National Park	1542 km^2	83.058
Spatial plan for the special purpose area of Grlište Reservoir basin	400 km^2	53.907
Spatial plan for the special purpose area of Ćelije reservoir basins	935 km^2	66.892
Spatial plan for the special purpose area of international waterway E-80 - Danube (Pan-Europian Corridor VII)	4.536 km^2	104.284
Spatial plan for the municipality of Knjaževac	1.202 km^2	88.348

Table 2. Basic geographical and linguistic characteristics of the PPTXM corpus

As already mentioned, the PPTXM corpus consists of six plans (Table 2). The territorial coverage of plans differs significantly from plan to plan, as do the dimensions of texts measured by the number of tokens. The territories covered by the plans range from 935 km2 to 9,184 km2, and the dimensions of the texts range from 53,907 to 104,284 tokens. However, it is noticeable that the size of the territory covered by the plan is not proportional to the dimension of the planning texts, which can be seen in the example *RSP for the Zlatibor and Moravica administrative districts*, which covers 9,184 km2 of territory and whose plan contains 90,724 tokens. Also, *Spatial plan for the special purpose area of international waterway E-80- Danube (Pan-Europian Corridor VII)* whose territory is almost twice as small in area, and the text of the plan contains about 14,000 tokens more. A similar example is with *Spatial plan for the special purpose area of Đerdap National Park* and *Spatial plan for the municipality of Knjaževac*. The first of these plans is larger in terms of territory, and the second contains a larger number of tokens.

The next step in the analysis of the corpus was to determine the frequency of occurrence of punctuation marks and different classes of words in texts. This was achieved through the use of morphological tags (*Part of Speech*

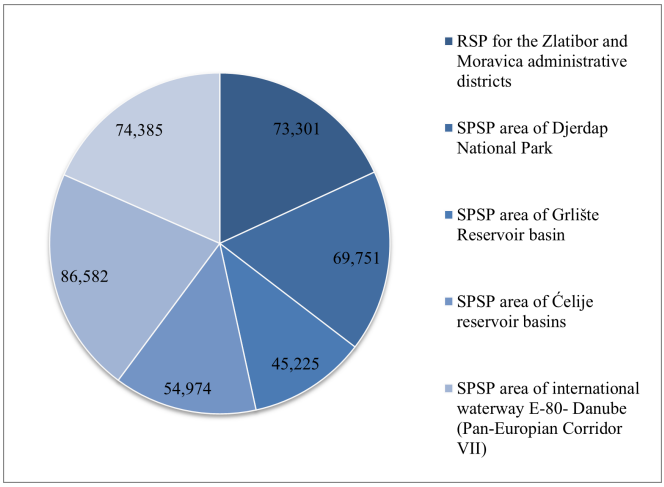

Figure 1. Extracted number of words per partition in the PPTXM corpus.

tags) for the Serbian language. Based on the obtained results (Table 3), it can be seen that the frequency of nouns is the highest, followed by the frequency of punctuation marks, and then the absolute frequencies of other pars of speech are listed. In addition to the total number of nouns, the query `[srpos = "NOUN"]` determined that there are 110,587 different forms, as well as 5,204 different lemmas.

The same query established which nouns are the most common in the entire corpus, and the five most frequent ones are shown in Figure 2, which shows that the frequency of nouns *područje* (area), *zaštita* (protection) and *razvoj* (development) in the corpus is higher than 3,000 and that they are used much more often in plans.

As already mentioned, there are a number of tokens within each partition. In the same way, it is possible to determine how each part of speech is represented in each of the texts of the PPTXM corpus. Considering the ratio of the number of nouns and tokens per corpus partition, it can be concluded that this relationship is disturbed only when it comes to the *RSP of the Zlatibor and Moravica administrative districts* and the *SP of Knjaževac*. According to the number of tokens, the *SPSP of the Zlatibor and Moravica administrative districts* is on the second place, and according to the number of nouns on the third place, while in the case of *SP of Knjaževac*, the situation

srpos	F	srpos	F
N (noun)	153.857	DET (determiner)	7.624
PUNCT (punctuation mark)	82.995	ADV (adverb)	7.367
ADJ (adjective)	78.039	AUX (auxiliary verbs)	6.991
ADP (prepositions)	42.125	PART (particle)	5.917
CCONJ (coordinate conjuction)	33.368	X (other)	3.878
PROPN (proper name)	27.500	SCONJ (subordinate conjunctions)	2.261
VERB (verb)	17.865	PRON (pronoun)	893
NUM (number)	15.873	INTJ (interjection)	660

Table 3. Frequency list of all represented values of the `srpos` position attribute in the PPTXM corpus

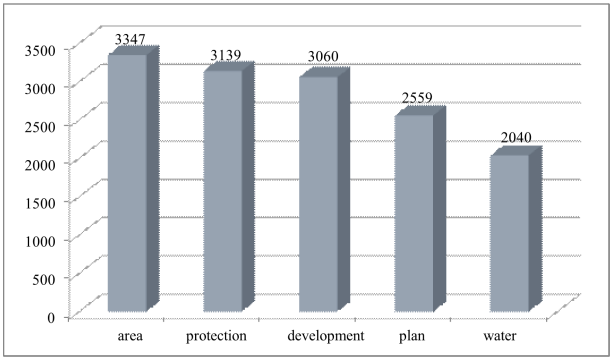

Figure 2. Distribution of nouns whose frequency in the corpus is higher than 2,000.

is the opposite. Other plans have maintained a proportional relationship between tokens and nouns (Figure 3).

The situation is somewhat different when it comes to the relationship between tokens and pronouns in PPTXM corpus partitions. The total number of tokens, as already determined, in the whole corpus is 487.213, and the total number of pronouns obtained by the query `[srpos = "PRON"]` is 893. The representation of tokens by partitions is not proportional to the representation of pronouns. Figure 4 shows the relationship between tokens and pronouns in each of the six spatial plan texts. The highest frequency of use of pronouns is in the *SP for the special purpose area of Ćelije reservoir basins*, although the number of tokens in this plan exceeds only the

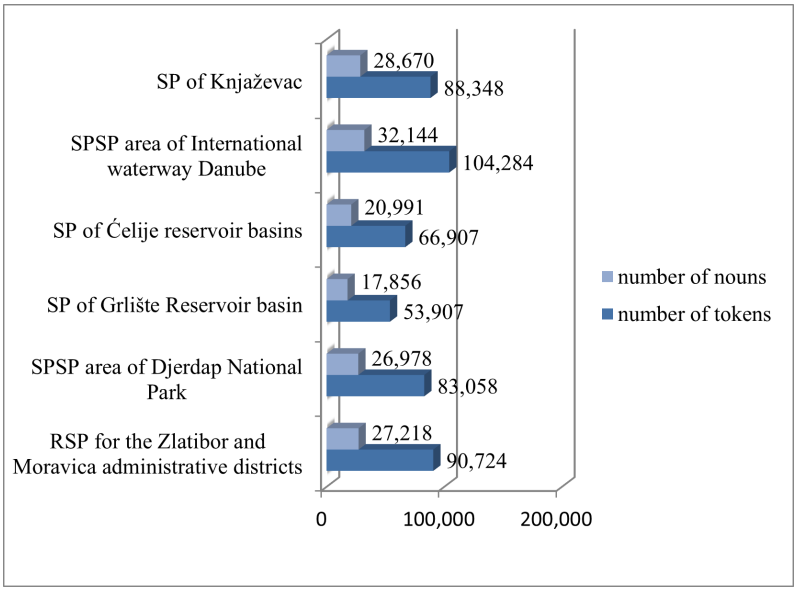

Figure 3. Relationship between tokens and nouns by PPTXM corpus partitions.

number of tokens in the *SP for the special purpose area of Grlište reservoir basin*. On the other hand, the *Spatial plan for the special purpose area of international waterway E-80- Danube (Pan-Europian Corridor VII)* as the largest partition calculated on the basis of the number of tokens, is only in fourth place in terms of the percentage of pronouns. *Spatial plan for the municipality of Knjaževac* and *SP for the special purpose area of international waterway E-80- Danube (Pan-Europian Corridor VII)* according to the number of tokens, take the third and fourth place, respectively, with a difference of only 5,000 tokens. However, in the *SP for the special purpose area of international waterway E-80- Danube (Pan-Europian Corridor VII)* the representation of pronouns is almost two and a half times higher than in the *SP for the municipality of Knjaževac*.

Previous analyzes are the basic form of statistical processing of texts. Tools in the TXM environment enable much more complex analysis of various forms of frequency and representation in texts. What is characteristic of the TXM environment is the fact that in the corpus, in addition to the total appearance of a certain part of speech and the frequency of individual forms,

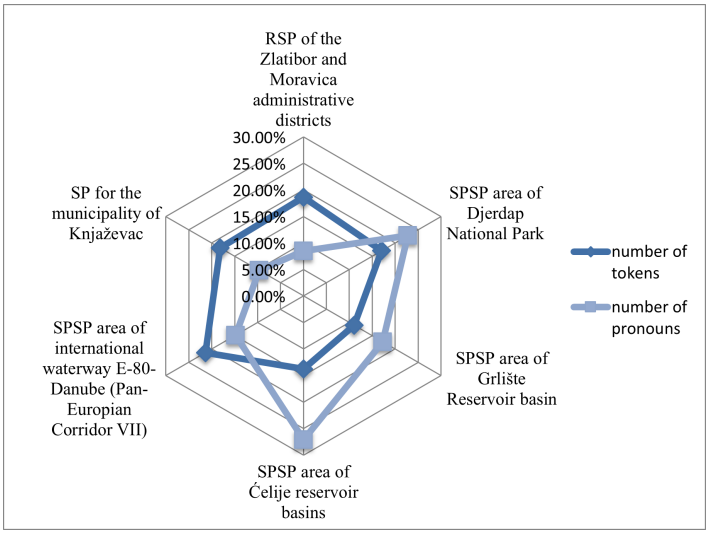

Figure 4. Relationship between tokens and pronouns by PPTXM corpus partitions.

it is possible to monitor the progression and cumulative frequency of different part of speech through corpus components and throughout the corpus.

The graph of progression, shown in Figure 5, gives examples of five nouns: *планина* (mountain), *експлоатација* (exploitation), *водоток* (watercourse), *пристаниште* (port), *споменик* (monument). The red line shows the cumulative frequency of the noun *планина* (mountain). One can notice that it most often occurs in *SP for the municipality of Knjaževac*, while its appearance in *SPSP area of Đerdap National Park* (5) and *SPSP area of international waterway E-80 - Danube (Pan-European Corridor VII)* (3) is at the level of statistical error. In the remaining three plans, the noun *планина* (mountain) is the least used in the *SPSP area of Ćelije raservoir*, somewhat more often in the *SPSP area of Grlište reservoir basin*, and most often in the *RSP of Zlatibor and Moravica administrative districts*. This frequency of the noun *планина* (mountain) according to the partitions of the PPTXM corps is expected because the area of municipality of Knjaževac, as well as the *RSP of the Zlatibor and Moravica administrative districts* have a pronounced hilly-mountainous relief.

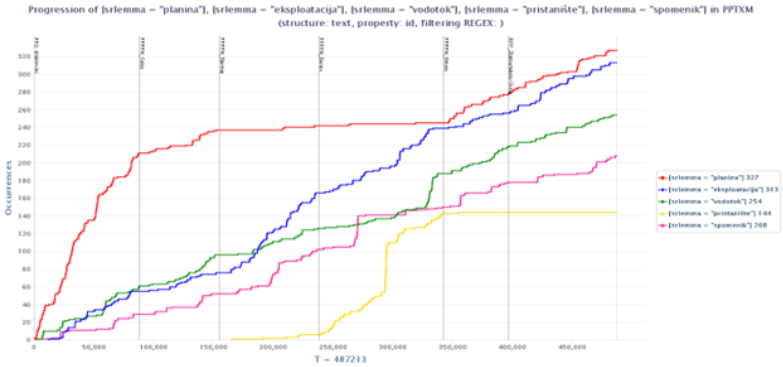

Figure 5. The progression of five nouns in PPTXM corpus texts; from left to right partitions are: *SP for the municipality of Knjaževac*, *SPSP area of Ćelije raservoir*, *SPSP area of Đerdap National Park*, *SPSP area of international waterway E-80 - Danube*, *SPSP area of Grlište*, *RSP of the Zlatibor and Moravica districts.*

The noun *пристаниште* (port) (yellow line) appears 144 times in the whole corpus, however, as can be seen in the graph, this noun was not used in *SP for the municipality of Knjaževac*, *SPSP area of Ćelije reservoir basin and RSP of the Zlatibor and Moravica administrative districts*, barely mentioned in *SPSP area of Grlište reservoir basin*, slightly higher in the *SPSP area of Đerdap National Park*, and the only spatial plan in which the presence of this noun is prominent is the *SPSP area of international waterway E-80 - Danube (Pan-European Corridor VII)*. Having in mind that this spatial plan covers the entire area of the Danube region in the Republic of Serbia, i.e., the entire course of the Danube in the country, it is completely understandable that the noun *пристаниште* (port) is used throughout the Plan.

The noun *водоток* (watercourse) (green line) is represented in all six plans, and although its appearance in *SP for the municipality of Knjaževac* and *the SPSP area of international waterway E-80 - Danube (Pan-European Corridor VII)* is more prominent, the use of this noun is not negligible in the remaining four plans. Since the entire territory of the Republic of Serbia is characterized by exceptional hydro potential and a developed river system, such a frequency in all six corpus texts could have been expected.

The distribution of natural and cultural heritage on the territory of the state of Serbia can be confirmed by the frequency of the appearance of the noun *споменик* (monument) (purple line) at the level of the entire corpus, but also its constituent parts. The noun *експлоатација* (exploitation) (blue

line) is another noun whose presence is noticeable in all PPTXM corpus texts. Of course, this was influenced by the fact that Serbia has certain reserves of different types of mineral raw materials and that their exploitation is present in all parts of the Republic.

The TXM tool provides another important type of analysis that provides the ability to track the distribution of the frequency of occurrence of certain words by partitions. Figure 6 shows the *планина* (mountain), *слив* (basin) и *подручје* (area). Each of them has a different degree of representation in the whole corps. With 3,139 occurrences, the noun *подручје* (area) is the most common noun at the level of the entire corpus, while the noun *слив* (basin) with 620 and *планина* (mountain) with 327 occurrences is much less used in planning documents. The results, after processing in the TXM environment, show a very high frequency of the noun *планина* (mountain) in *SP for the municipality of Knjaževac*, significantly above average compared to its use at the level of the whole corpus, while its frequency is below average in the remaining five spatial plans, especially in *SPSP area of international waterway E-80 - Danube (Pan-European Corridor VII)*. The use of the noun *слив* (basin) is most pronounced in the *SPSP area of Ćelije reservoir basin* and there is a positive score in the *SPSP area of Grlište reservoir basin*. In all four remaining plans, the use of this noun is below average, and as with the noun *планина* (mountain), it is most prominent in the *SPSP area of international waterway E-80 - Danube (Pan-European Corridor VII)*. The noun *подручје* (area) has a negative representation, when comparing with the entire corpus, in the *RSP of the Zlatibor and Moravica administrative districts* and the *SP for the municipality of Knjaževac*. In other partitions, the representation of this noun is above average.

Bearing in mind the fact that the TXM platform imported a corpus previously marked with named entity tags, it can also analyse, in addition to determining the frequency of different part of speech, the degree of representation of named entities, using the `Index` option and posing different queries, such as `/region[top_gr]`, `/region[top_hyd]`, `/region[demonym]` and others. Figure 7 shows the six most frequent named entities in the PPTXM corpus. In addition to the total frequency, the numbers of different lemmas and different forms for each named entity are shown. By far the most represented named entity type is `top_gr`, which is used for urban settlements and other inhabited areas.

The query `/region[org_rel]` determined that there are 33 different forms of named entities `org_rel`, which marked religious objects, and that 29 different lemmas with a total of 56 occurrences were recorded for them.

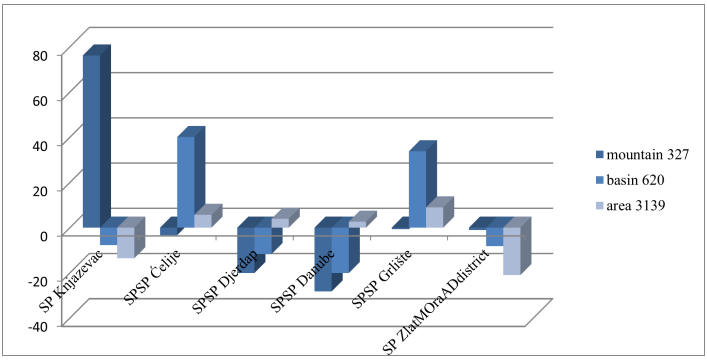

Figure 6. : Specificity of the use of nouns in the PPTXM corpus.

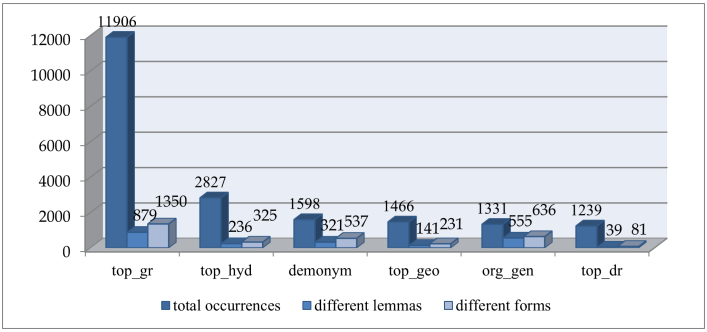

Figure 7. Most represented named entities in the PPTXM corpus.

The most frequent are *Манастир Милешева* (Mileseva Monastery) (6), *Црква Св. Петра* (the Church of St. Peter) (4), *Црква Св. Николе* (Church of St. Nikola), *Св. Тројице* (St. Trinity), *Св. Ђорђа* (St. George) and *Манастир Рача* (Monastery of Rača) with 3 appearances, then eight religious buildings that appear in the corpus 2 times and the remaining named entities `org_rel` that appear only once. However, in the TXM environment, it is possible to formulate more complex queries. Thus, the query `/region[org_rel][]0,2/region[top_gr]` determines the frequency of named entities `org_rel` followed by the named entity `top_gr`. The result is a completely different order in the number of appearances of religious buildings, as shown in Figure 8. In this case, the most frequently listed are *Црква Св. Илије* (the Church of St. Elijah) in Brekovo, *Црква Светог*

Никоље (the Church of St. Nicholas) in Brezova and *Црква Светог Саве* (the Church of St. Sava) in Savinac.

Query /region[org_rel][]{0,2}/region[top_gr]

word	Frequency
crkva Sv . Ilije u Brekovu	2
crkva Svetog Nikole u Brezovi	2
crkva Svetog Save u Savincu	2
crkva Sv . Bogorodice u Donjoj Kamenici	1
crkva Sv . Nikole - Stari Slankamen	1
crkva Sv . Preobraženja , Pančevo	1
crkva Svete Bogorodice u Donjoj Kamenici	1
crkva Svete Bogorodice u Prilipcu	1
crkva Svete Trojice u Bistrici	1
crkva Svetog Đorđa u Godoviku	1
crkva Svetog Đorđa u Knjaževcu	1
crkva Svetog Đorđa u Užicu	1
manastira Mileševo (Prijepolje	1
manastir Vojlovica , Pančevo	1

Figure 8. Frequency of religious objects based on query `/region[org_rel][]0,2/region[top_gr]`.

7 Conclusion

The paper presents various linguistic and statistical analyzes, using TXM tools, on the texts of spatial plans in which the named entities were previously marked using SrpNER, a system for recognizing named entities in the Serbian language. The common characteristics of the texts of planning documentation were pointed out, as well as certain particularities of individual spatial plans, which are included in the Serbian corpus *SrpKor2021*, which currently contains more than 600 million words, as part of the domain corpus of spatial planning.

Spatial plans, which are the main focus of this paper, are a specific type of documents characteristic in the field of spatial planning. In order to enable more detailed research of these texts, the PPTXM corpus was created, which includes six documents dealing with different regions of the Republic

of Serbia: Zlatibor and Moravica administrative districts, basins of reservoirs "Grlište" and "Ćelije", Đerdap National Park, International Waterway E-80 - Danube (Pan-European Corridor VII) and the municipality of Knjaževac.

The analysis covered more than 400,000 words, whose representation ranges from 11.18% in the *PPPPN of the "Grlište" reservoir basin* to 21.42% in the *PPPPN of the international waterway E-80 - Danube (Pan-European Corridor VII)*. The corpus contains more than half a million tokens, a third of which are nouns.

It should be noted that, in addition to the presented analyzes, the TXM environment provides other possibilities and that their use as well as the use of presented textometric approaches depends on pre-set requirements as well as the needs and preferences of researchers. What is indisputable is that the TXM platform enables the display of the features of the corpus as a whole, as well as its components. For that reason, the application of these methods is very important for various quantitative analyzes and determination of characteristic features of texts. Some of the presented results, such as the representation of specific nouns in the entire corpus and the frequency of their occurrence in individual plans, could be deduced in advance on the basis of the plan coverage area. However, the obtained results on the representation of pronouns by corpus partitions cannot be given a logical explanation only on the basis of knowledge about the scope of the plans; for their better understanding it is necessary to conduct additional research and analysis.

Acknowledgment

This research was conducted in the scope of the work on a doctoral dissertation at the Faculty of Philology, University of Belgrade entitled "The Development of library and language resources for organizing and finding information on spatial planning." The tools used during the research were specifically developed for the Serbian language (SrpNER) or were adapted for various linguistic research of Serbian texts (TXM) within the Society for Language Resources and Technologies. The author owes special gratitude to professors Dr. Cvetana Krstev and Dr. Ranka Stanković, without whose personal engagement, the research presented in this paper would not have been possible.

References

Bjekić, Jovana, Ljiljana Lazarević, Milica Erić, Elena Stojimirović, and Teodora Đokić. 2012. "Razvoj srpske verzije rečnika za automatsku analizu teksta (LIWCser)." *Psihološka istraživanja* XV (1): 85–110. https://doi.org/10.5937/PsIstra1201085B.

Francis, W Nelson. 1975. "Problems of Assembling, Describing, and Computerizing Corpora." *Research Techniques and Prospects. Papers in Southwest English,(1),* 15–38.

Heiden, Serge. 2010. "The TXM Platform: Building Open-Source Textual Analysis Software Compatible with the TEI Encoding Scheme." In *Proceedings of the 24th Pacific Asia Conference on Language, Information and Computation,* 389–398. Tohoku University, Sendai, Japan: Institute of Digital Enhancement of Cognitive Processing, Waseda University. https://aclanthology.org/Y10-1044.

Jaćimović, Jelena. 2019. "Textometric methods and the TXM platform for corpus analysis and visual presentation." *Infotheca - Journal for Digital Humanities* 19 (1): 30–54. https://doi.org/10.18485/infotheca.2019.19.1.2.

Krstev, Cvetana, Ivan Obradović, Miloš Utvić, and Duško Vitas. 2014. "A system for named entity recognition based on local grammars." *Journal of Logic and Computation* 24 (2): 473–489. https://doi.org/10.1093/logcom/exs079.

MacMurray, Erin, and Marguerite. Leenhardt. 2011. "Textometry and Information Discovery: A New Approach to Mining Textual Data on the Web." In *Proceedings of the 2011 International Conferece on Artifical Intelligence, ICAI 2011,* edited by H.R. Arabnia, D. Fuente, E.B. Kozerenko, and J.O. Olivas, II:605–611. Las Vegas: Worldcomp'11.

Maurel, Denis, Nathalie Friburger, and Iris Eshkol-Taravella. 2014. "Enrichment of Renaissance Texts with Proper Names." *Infotheca* 15 (1): 29a–41a. https://infoteka.bg.ac.rs/index.php/en/archives/2014/1/infoteka-15-1-2014-30-41.

McEnery, Tony, Richar Xiao, and Yukio Tono. 2006. *Corpus-based Language Studies: An Advanced Resource Book.* London: Routledge. https://www.lancaster.ac.uk/fass/projects/corpus/ZJU/xCBLS/CBLS.htm.

Mehl, Matthias R., and Alastair J. Gill. 2010. "Automatic text analysis." In *Advanced methods for conducting online behavioral research,* edited by S. D. Gosling and J. A. Johnson, 109–127. Washington: American Psychological Association. https://doi.org/10.1037/12076-008.

Pincemin, Bénédicte, Serge Heiden, and Matthieu Decorde. 2020. "Textometry on Audiovisual Corpora Experiments with TXM software." In *JADT 2020 : 15es Journées internationales d'Analyse statistique des Données Textuelles.* http://lexicometrica.univ-paris3.fr/jadt/JADT2020/jadt2020_pdf/PINCEMIN_HEIDEN_DECORDE_JADT2020.pdf.

Popović, Ljubomir, and Duško Vitas. 2003. "Konspekt za izgradnju referentnog korpusa srpskog standardnog jezika." In *Naučni sastanak slavista u Vukove dane.* Beograd, Novi Sad: MSC.

Trtovac, Aleksandra, Vasilije Milnović, and Cvetana Krstev. 2021. "The Serbian Part of the ELTeC - from the Empty List to the 100 Novels Collection." *Infotheca - Journal for Digital Humanities* 21 (2): 7–25. https://doi.org/10.18485/infotheca.2021.21.2.1.

Utvić, Miloš. 2013. "Izgradnja referentnog korpusa savremenog srpskog jezika." PhD diss., Univerzitet u Beogradu, Filološki fakulte. https://nardus.mpn.gov.rs/handle/123456789/4091.

Vitas, Duško. 2019. "Food as Text." *Infotheca - Journal For Digital Humanities* 19 (2): 139–161. https://doi.org/10.18485/infotheca.2019.19.2.7.

Vitas, Duško. 2022. "From Onions to Champagne – Food and Drink in the SrpELTeC Corpus." *Infotheca - Journal for Digital Humanities* 21 (2): 88–118. https://doi.org/10.18485/infotheca.2021.21.2.5.

Васиљевић, Небојша М. 2014. "Аутоматска обрада правних текстова на српском језику." PhD diss., Универзитет у Београду, Филолошки факултет. https://nardus.mpn.gov.rs/handle/123456789/4091.

Витас, Душко, and Цветана Крстев. 2016. "Оглед из гастрономатике [Experiments in gastronomatics]." In *Теме језикословне у србистици кроз дијахронију и синхронију [Linguistic topics in Serbian through diachrony and synchrony],* edited by Јасмина Дражић, Исидора Бјелаковић, and Дејан Средојевић, 1–10. Novi Sad: Филозофски факултет.

Витас, Душко М. 2018. “Храна из нежељене поште : (анатомија језика брзе хране [Spam Food: (Anatomy of Fast Food Language)]).” In *Српски језик и његови ресурси: теорија, опис и примене Научни састанак слависта у Вукове дане,* edited by Божо Ћорић and Александар Милановић, 21–35. Београд: Међународни славистички центар, Филолошки факултет, Универзитет у Београду. https://doi.org/10.18485/msc.2018.47.3.ch2.

Обрадовић, Иван, Александра Томашевић, Ранка Станковић, and Лазић Биљана. 2017. “Увођење доменских и семантичких маркера за област рударства у српске електронске речнике.” In *Научни састанак слависта у Вукове дане - Српски језик и његови ресурси: теорија, опис и примене,* edited by Рајна Драгићевић and Александар Милановић, 147–158. Београд: Међународни славистички центар на Филолошком факултету. https://doi.org/10.18485/msc.2017.46.3.ch10.

Службени гласник РС. 52/2021. *Закон о планирању и изградњи.* Бр, 72/2009, 81/2009 - испр., 64/2010 - одлука УС, 24/2011, 121/2012, 42/2013 - одлука УС, 50/2013 - одлука УС, 98/2013 - одлука УС, 132/2014, 145/2014, 83/2018, 31/2019, 37/2019 - др. закон, 9/2020 и 52/2021), 52/2021.

Службени гласник РС. 32/2019. *Правилнику о садржини, начину и поступку израде докумената просторног и урбанистичког планирања.* Бр. 32/2019), 32/2019.

Тртовац, Александра. 2017. “Аутоматизација библиотека у Србији - историјски преглед.” *Библиотекар,* no. 2, 101–114. https://bds.rs/wp-content/uploads/2018/01/bibliotekar-2-2017-7.pdf.

Digital Resources about Aviation History

UDC 004.738.5:027]:656.7(091)
DOI

Ljiljana V. Macura
ljiljana.macura@gmail.com
The National Library of Serbia
Belgrade, Serbia

ABSTRACT: In this paper the author is trying to find out if there is any digital library of aviation history in general and describe digital resources in the field of aviation history. Both the world history of aviation and the history of aviation in Serbia are the subject of the research in the paper. Results of research carried out between April 2019 and November 2020 showed that digital library of aviation history of particular country or nation is not available. There are different resources, that is, web sites where various artifacts of aviation history can be found.
KEYWORDS: Aviation history, digital library, cultural heritage, metadata.

PAPER SUBMITTED: 23 March 2022
PAPER ACCEPTED: 04 April 2022

1 Introduction

The central topic of this paper are online resources with its main objective to nurture the history of aviation.[1] During the work on the topic, no similar researches or theoretical works in the field of digitization of aviation history were found, but a large number of network resources that are fully or partially, professionally or thematically determined for aviation history were found. While some of the resources present the history of aviation in the world, other resources present the history of domestic aviation or the history of aviation in a particular region. The paper "Digital Resources about Aviation History" is written for experts in the field of librarianship

1. The text was written along with the creation of doctoral dissertation "Conditions for creation and initial records (material) of the digital library of the history of aviation in Serbia", the proposal submitted in 2019 at the Faculty of Philology, University of Belgrade.

and information sciences, as well as for the experts in aviation. The topic can be interesting to a wider audience as well as to traffic experts in general, enthusiasts and audience with an affinity for aviation. The target groups of readers in the field of aviation are the ones in education, lecturers and aviation professionals.

The need for a digital library of aviation history is confirmed by a multitude of different digital resources that contain one thing in common and that is (the nurturing of) the aviation tradition. In addition to the organization of aviation events, the digitization of various documents also contributes to nurturing the aviation tradition. Text documents, audio and video recordings, various visual materials are all digitized, and it can be assumed that computer programs for aviation purposes are stored in the same way, e.g., aircraft manufacturer program, airline software for flight simulators or airline commercial activity. The type and scope of material as well as the type of services depend on goals and the target group of users of such library. Publications containing aeronautical technical information, as well as photographs, drawings and notes that lean towards the social history of aviation have been digitized. There are digital resources from the history of aviation that exclusively digitize photographs, while others mostly digitize video recordings. In general, there are digital resources whose collections vary from posters, bookplates and postcards with motives from aviation. There are also digitized catalog sheets (bibliographic records)[2] related to aviation. It can be stated that generally, digital resources often contain information from aviation, but the success in searching is much greater if textual records are searchable by words from the text.

Unfortunately, no theoretical works have been found on the digitization of aviation history so far, that is, no theoretical works on the topic of digital libraries of aviation history or collections from aviation history. However, if the elements of the IFLA / UNESCO Manifesto for Digital Libraries are followed[3], they can point out the importance of digital resources or digital library of aviation history. Digital libraries have a "strategic role" in society by helping to organize and "disseminate information". Thus, digital sources of aviation history would preserve and present material that represents cultural and scientific heritage. The digital library of aviation history would represent, in a certain way, the systematized history of aviation and the organization of knowledge in that field. As a "collection of digital objects

2. Retrospective catalog conversion.

3. All quotations in this paragraph are related to the Manifesto (IFLA/UNESCO 2012).

of certified quality" and "services necessary for finding and using" of digital sources, the digital library would provide systematized information in a "fast and cost-effective way". Information for researchers, lecturers and people in the educational process, as well as the general public (persons outside the aviation profession), who thus put aviation activities in a certain social context, in the case of our social reality, would raise awareness of the importance and comparability of Serbian aviation globally.

2 Digital sources of aviation history – the course of research

There are many resources in digital media that deal with the history of aviation in different ways. The research was conducted through the Google search engine, with different dynamics, from April 2019 to November 2021. Queries containing the following terms: AERO*, AERONAUTIC, AIR FORCE, AIR TRAFFIC, AIR TRANSPORT, AIR*, AVIATION HISTORY, AVIATION HISTORY AND DIGITAL LIBRARY, AVIATION MUSEUM, AVIATION, CIVIL AVIATION AUTHORITY, CIVIL AVIATION, DIGITAL LIBRARY, FLIGHT, FLYING, HISTORY, LIBRARY, MUSEUM and TRADITION were performed.

The research started with the queries AVIATION HISTORY and AVIATION HISTORY AND DIGITAL LIBRARY and provided many results which were not (arranged) in any order. Introduction of the terms AERO*, AERONAUTIC, AIR FORCE, AIR TRAFFIC, AIR TRANSPORT, AIR*, AVIATION MUSEUM, AVIATION, CIVIL AVIATION AUTHORITY, CIVIL AVIATION, DIGITAL LIBRARY, FLIGHT, FLYING, HISTORY, LIBRARY, MUSEUM and TRADITION also provided various results. Finally, the research was completed by introducing country parameters. The name of the country in English was combined with the terms AIR FORCE, CIVIL AVIATION AUTHORITY, AVIATION MUSEUM and DIGITAL LIBRARY.

The use of the last of those methodologies mentioned before, provided the best results. However, the search according to that principle could not be fully applied to all states, i.e. the territories to which it referred. If there was no other way to get an answer, the search was performed according to the state parameter on various web pages (*Global security*, *Flickr*, *Flock Aero*)[4] that appeared in response during one of the previous attempts of research.

4. More about *Global Security* and *Flock Aero*, in the text below; *Flickr* will not be described since it is general type resource.

In the end, a list of organizations, institutions or companies that present the history of aviation in electronic form in different ways was obtained.

3 Research results

Since the middle of the 1990s, there were websites that continued the previous tradition of nurturing the history of aviation. Textual recordings and photographs were most often in electronic form, but over time, recordings from other media were also included in the content. Today, the history of aviation is presented on the World Wide Web, originally preserved on two-dimensional and three-dimensional media, audio and video recordings, i. e., in the form of images, text, sound and video. The history of aviation in electronic form is also represented by individual items on the web pages of museums or libraries as well as web pages created exclusively for the purpose of presenting the history of aviation.

In this paper, some cases of network resources of the history of aviation are presented. They are presented by: name, territorial scope, i. e., history of domestic aviation, regional aviation or aviation in the world, as well as the ways of searching, ways of presenting search results, comments on the functionality of the site, registration conditions and opportunities provided by registration on the site if available or necessary.

The following web pages are listed alphabetically by country of origin[5] and the search results (digital objects) are described by metadata that vary in range (Тртовац 2017).

3.1 United Kingdom

The Heritage collection comes from the website of the Royal Aeronautical Society (Royal Aeronautical Society, n.a.). It contains 14 thematic collections and seven biographies of people who are the pioneers of the theory and practice of British aviation. Researchers and other visitors to the *Heritage Collection* are instructed on how to use the resource while respecting copyright. Special emphasis was placed on publishing photos, drawings and other digitized content from the site, which is not disputable if it refers to the resource, that is, to the *Heritage Collections* page. Searching is simple, by entering a query in the provided field. The access page shows the most

5. Alphabetical order refers to the resources from all countries except Serbia whose resources will be described afterwards.

sought-after objects (Top Terms), systematized by terms (Taxonomy terms), i.e. metadata and the numbers of items marked. In other words, collections of photographs, drawings and notebooks, models of airplanes, constructions of airplanes, gliders and propulsion units are shown, as well as collections related to persons and sciences of importance to the history of the pioneering era of aviation. Each selected object contains metadata that relates to the digital document itself (Item Details) and to the Collection (Collection Details) to which it belongs. These are the following metadata: description, time, and often different terms (Taxonomy terms) - for each unit of digital library material (digital object), i.e., collection name, collection, description and terms (Collection Terms) - for a collection, where the metadata for the collection is always the same (since they describe the same collection). The site is clear and functional. There is a option for registration on the main page (website of the Company).

Aeroflight (Aeroflight, n.a.) is a site registered in June 1996, and since October 2019 this site also functions as *Aeroflight Cloud* (Aeroflight Cloud, n.a.), representing a space for collecting and disseminating information in the field of aviation. These are current information and information from history of aviation, related to air force, sport and civil aviation, museums, institutions and organizations, aircraft owners, etc. For the needs of the site but based on crowdsourcing activities, a database of monograph publications is created, primarily in English. These are publications with ISBNs that were published until 2000, regardless of the original form. The database was created in Microsoft Excel, and consists of three sheets: title, attribute definitions and attribute values. There is a great similarity between the metadata attributes that describe digital documents on the *Aeroflight* page and those in bibliographic databases that use some of the MARC formats. In addition to the database of monographs, there is also a database of serial publications. Regarding information retrieval, search results represent articles, lists, and links to other sites. At *Aeroflight*, you can also find content from the history of aviation in Serbia, under the designation Yugoslavia and Serbia. The page is loaded with various (and numerous) contents within the topic, but it kept its functionality. There is an option for registration on this page.

Flock Aero (FltCtrl Ltd, n.a.) is the world's largest airline services marketplace. For the paper, this page is interesting in terms of evidence of the existence of any sign of aviation in certain countries or territories (tab Countries at (FltCtrl Ltd, n.a.)), in cases if searching by metadata (terms such as AVIATION HISTORY AND DIGITAL LIBRARY or just AVIATION HISTORY) did

not provide satisfactory results. In other words, this page is interesting in the sense of civil aviation authorities.

Air Team Images was registered in 2003 (AirTeamImages, n.a.). Its initiator is also a reputable photographer, focused on aviation. *Air Team Images* is an online library of aircraft photographs, which is set up with the aim of providing the site visitor or client with the highest quality "aerial photography" ("About Us" at (AirTeamImages, n.a.)). Although the emphasis is on aircraft, this page contains photographs that refer to all aspects of aviation. In addition to photos, videos and texts can also be found on the website. A slide show consisting of fifty newly arrived photos, as well as a list and links to the items that the visitors of the page most often searched for (Popular Photo Searches) are shown at the Home page. The search is performed by metadata and can simply be carried out by entering a query in the Search field. Also, the search can be done by selecting the metadata attributes, from the available fields, if the Popular Photo Search option is selected. In that case, a large number of attributes are available. Some of them are: aircraft manufacturer, operator, country, photo theme, photo orientation, photographer, sorting and photo ID. The results of the query are often photos, marked with an ID number. If a photo is opened, there is a broader description.[6] In that case, photos are not only noticeable, but there's a detailed description given in the form (attributes and values) of numerous metadata: manufacturer, aircraft type and model, operator code, place or country, date the photo was taken, photo ID and annotation. The image link is below each digital document (in this case, a photograph). *Air Team Images* also contain material from the history of aviation in Serbia, that is, the aviation of Yugoslavia and Serbia. The site is clear and functional. There is an option for registration on the site.

3.2 Canada

Avialogs (Avialogs, n-a-) online community and library was registered in 2010 with the aim of preserving publications in the field of aviation and providing the widest audience with online access. For the needs of this library, monographic and serial publications in the field of aeronautical technology have been digitized. *Avialogs* operates in the form of several collections related to aircraft, powertrains, engines, equipment, parts, construction, manuals and serial publications. For now, the search is performed by metadata. A new released version of *Avialogs* will not include searching by words from

6. All photographs are interactive locations.

text. However, the new version will include a complete transcript (software improvement),[7] drawings and diagrams, unlimited download for all users, whether they are legal entity (aviation facilities, museums, collectors) or person, and finally, a monthly plan. Speaking about monographic publications, digital documents are described by metadata that usually has only the following attributes: year of publication, number of pages, language of publication, type of publication and name of the source on the page where the object was published, that is, situated on *Avialogs*. If it is an object from the Aircraft Register collection, the documents are described by metadata with several attributes. These are: the name of the series, the designation of the series, the name of the factory in which the aircraft was manufactured, the name of the engine type, as well as the name of the source. Digital objects (publications) are fully accessible to registered visitors. Registered users (site visitors) can access entire publication, while others can view the first ten pages of the selected publication. Registered users also have the choice to read publications free of charge or with one of the available membership fees, gain the opportunity to edit, download or print the page. Before the step of "membership", i.e. subscription for use, each visitor has access to all digital objects. Users are not required to join as long as they do not need a particular publication. They may take a look which publications are available through this resource without any commitment. There are three publications in Serbian which are, in fact, manuals, about the use of some planes of the post-war Yugoslav aircraft factory Soko, which have also been digitized. The site is clear and functional.

3.3 Germany

Flugzeugbilder (Flugzeugbilder, n.a.) was registered in 2001. It represents aircraft photos; amateur photos and photos taken by professional photographers. The home page, which is in German can be switched to English user interface, while the French is under construction. The search is performed on metadata that has the following attributes:

- airline,
- airport,
- email,
- aircraft type,
- photo id,

7. Author's note.

– photographer,
– comment,
– recording date,
– aircraft registration plate.

One can view 10, 15, 25 or 50 photos per page. The search can start or end with a certain term, perform with the exact term or by entering keywords. The search can be simple and combined. Search results give a list of objects. These are photographs of airplanes with attribute values: photo ID, number of views, date taken, photographer, operator, type of aircraft and its registration number, sorted by number or date of view, in ascending or descending order. Airport also appears among the attributes and is described by the three-letter and four-letter designation of the airport; location and name of the airport; the state. If a specific photo is selected, a larger view with the same metadata is obtained. *Flugzeugbilder* contains photos of the Serbian Air Force.[8] This website is clear and functional but also contains the optimal amount of metadata. The visitors of the site are also allowed to upload their photos. The instructions on the procedure can be found on the Home page. There is no registration on the site.

3.4 United States

The Aviation History Online Museum (Aviation History Online Museum 1996) is a resource of aviation history that offers textual, visual and audio-visual records. This online resource was launched in 1996, for educational purposes. The predecessor of this page is Aviation Models, a page dedicated to aircraft modeling, launched in 1995. On the Home page of the Aviation History Online Museum there is the content of the page grouped by areas: aircraft, engines, theory, aviation, modeling, World War I, pioneer era, videos, links to other pages and briefly about the page. Each area is described with several metadata. Aircraft engines have three attributes: the name of the manufacturer, the type of engine and the country of production. In the case of aircraft, there are attributes related to the manufacturer, type of aircraft, country of manufacture, text and color attributes; text and color have the values "yes" and "no". The content of the site is easy to understand, it is rich in metadata describing aviators, pioneering age and theory, while other topics are modestly described. *The Aviation History Online Museum*

8. Query that retrieves data about Serbian aviation (accessed 17. 1. 2022).

page is the optimal source of information for aviation history and is very functional. There is neither search field nor registration on the site.

The Digital Collections of the WWII Museum (The National WWII Museum, n.a.) is website of the *World War II Museum* (WWII Museum), which was founded as an institution in 2003 in New Orleans.[9] This online resource contains photographs and testimonies (of World War II participants) in the form of video recordings of interviews. The search is simple and is performed by entering a query in the search field on the access page. Search is facilitated by metadata sorted by the branch of service and theater of operations criteria.[10] From the access page through the browser, one can find numerous metadata that make searching easier. They are sorted by service in the army, area of operation and terms (subject terms) from dictionaries related to a particular topic.[11] The terms from the dictionary are numerous, and the dictionary itself consists of hypertext connections, that is, the term and the number of objects to which the term refers. A term can be a single word or it can be a phrase. Using terms makes it easier to form queries. For example, two items are scored on the term friendships, data type video: (1) the *Dabney Montgomery's video* (part of *Early Life*) and (2) the *Donald Wilson's video* (*The Naval Battle of Guadalcanal*). Each object contains a document and metadata. In the first case, attribute values are Dabney Montgomery and Early Life. Also, an extensive annotation was made. Descriptive terms are Personal / family relationships, Prewar life, Experience with death, Great Depression, Race relations, Ethnic / racial identity, Education, Reaction to Pearl Harbor, Enlistment and Friendships. *The Digital Collections of the WWII Museum* also contain records relating to the history of the Yugoslav Air Force.[12] *The Digital Collections of the WWII Museum* site is quite simple, well organized and very useful. There is also the possibility of registration

9. Originated from The D-Day museum founded 2000; intended to commemorate American veterans, that is, participants in the Battle of Normandy More details on the origins of these institutions (accessed 13. 1. 2022).

10. Following terms: *US Army Air Forces*, *US Navy*, *US Army*, *US Marine Corps* and *Civilian US Merchant Marine* are metadata attributes referred to service, while terms: *China Burma India Theater* (CBI), *Pacific Theater of Operations* (PTO), *Atlantic Theater of Operations* (ATO), *European Theater of Operations* (ETO), *Mediterranean Theater of Operations* (MTO) and *Home Front* are metadata atributes refered to zone of action. Available (accessed 11. 1. 2022)

11. Browsing page (accessed 7. 2. 2022)

12. Eben Smith – Final mission and capture, Luke McLaurine – Loss of aircraft and capture (accessed 13. 1. 2022)

on the page, which, judging by the values from the available list, is intended for the citizens of the United States of America.[13]

Flight Journal (Flight Journal, n.a.), an aviation magazine, was launched in Ridgefield in the 1990s in a classic form (ISSN 1095-1075) and at the beginning of the next decade it acquired its online edition. Flight Journal presents the history of aviation from the pilot's perspective: "we put readers in the cockpit". The basic metadata of the page are: Air Force, Civil Aviation, News and Stories (various events and topics; restoration, personalities) and the directory of the United States Air Force Museum, in alphabetical order, from A to W (Alabama - Wyoming). On the website one can also find items from the history of aviation in Serbia, presented in the form of an article.[14] The publication offers the possibility of subscription and registration.

Global Security (Global Security, n.a.) is a continuous online resource, in fact, a serial publication (ISSN 2769-8947), registered in 2000. It covers security issues (on the planet and extraterrestrial space), that is, issues of military and intelligence nature. Global Security broadcast news is updated every hour.[15] A search on the page can be performed from the access page, by selecting the country or territory, or entering a query in the search field. One of the metadata for countries is aviation. The basic metadata on *Global Security* are: land forces, warships, airplanes, missiles, laser technology, intelligence, space; war zones; para-military forces. This page is not relevant to the history of aviation but was the only source on the existence of aviation in the case of individual states or territories.

The Aviation History Collection - Selections from the George Hardie Collection, of the University of Wisconsin (UWM Libraries Digital Collections, n.a.) was digitized in 2006. It consists of ten titles from the collection mentioned above, that is, digitized pages of publications related to aviation, in the period from 1865 to 1930.[16] Digitization has been carried out in this institution since 2001,[17] which enables remote access to library and information materials. In the year 2022, the material that includes over seventy collections has been digitized. There is a basic and advanced search on the page. The search results are described by the following attributes: title, author, year, title of the publication, annotation, subject matter, type of content, place of publication, name of publisher, medium, notes, scope of publica-

13. The registration page (accessed 13. 1. 2022)
14. On This Day in Aviation History (accessed 17. 1. 2022)
15. Company History (accessed 18. 1. 2022)
16. Keyword search (accessed 11. 1. 2022)
17. About UWM Libraries Digital Collections (accessed 18. 1. 2022)

tion, language, donor, year of digitization, format as well as other elements. Each page of the digitized publication is also described by its own title, year, publication title, copyright and collection name. Unfortunately, this well and easily organized page does not contain data from the history of aviation in Serbia.[18] There is no registration on the site.

The Latin American Aviation Historical Society (LAAHS) was founded in 1997 and registered in Miami as a non-profit organization (The Latin American Aviation Historical Society, n.a.). It displays various textual and visual forms: articles, columns, reviews, Facebook forums related to the history of Latin American aviation and is the only tool the Society uses to publish, preserve and disseminate the research results of its members.[19] In addition to the articles, the Home page also contains information about the Society, instructions for authors, donors and site visitors, as well as search fields. The search is basic and is performed by entering a query in the field provided for this type of search. In addition to the basic search, there is also a search that is performed by the criteria of the category of articles (article, column, review). Text and images are accompanied by labels. Within the (Categories, Articles) entry point, articles are grouped into *Caribbean Aviation*, *Central American Aviation*, *General Latin America Aviation*, *Mexican Aviation* and *South American Aviation*. There is also an archive of articles on the pages covering the period 2018-2021 and the search is performed by selecting the month of a particular year. Talking about the category of Latin American aviation, the Latin American air force is divided by the territorial principle. The forum on the page is restricted and intended only for members of the group (on the social network Facebook). There is no registration on this page.

Historynet, LLC, founded in 2015,[20] as the largest publisher of history magazines, published the magazine *Aviation History* (HistoryNet, n.a.), and provides among other things daily information and a photo gallery on the site. A search on a page is basic, by entering a topic, event, time period or place in the search box. One of the responses to the query Serbia AND aviation history was an article with a title *F-117A: The Black Jet*.[21] Articles are not accompanied by metadata or labels. Metadata for aviation history on *Historynet* are: aviation history magazine, aviation history, Mag. Aviation

18. Queries Serbian AND aviation and Yugoslavia AND aviation returned no match.

19. About LAAHS (accessed 19. 1. 2022)

20. Historynet (accessed 18. 1. 2022)

21. F-117A: The Black Jet (accessed 14. 2. 2022).

History Aircraft, Mag. Aviation History Featured, Mag. Aviation History Hero, Mag. Aviation History Modeling, Mag. Aviation History Mystery and Mag. Aviation History Personalities. There is no registration on the site but there is a subscription option.

3.5 Netherlands

Aviation News (Aviation News, n.a.) is a website, launched in 2006 in the Netherlands, which provides the latest aviation news in general. Following websites originate from the *Aviation News*: *Aviation Museum database*, *Airshow reports* and the unofficial *Paris Airshow website*.[22] These are online resources that are focused both on the contemporary and the history of aviation. The most relevant for this paper is Aviation Museum website. It covers Africa, Asia, Europe, Latin America, the Middle East, North America and Oceania. There is also a sitemap; a list of hypertext links to each country that is relevant to this page. Each of the countries on the list can also be represented by a new list (of museums' or aviation institutions' website). If the request is Serbia, the response is the "Aviation Museum" in Belgrade, "Soko Jastreb" and "Mig-21" (two military planes located in the yard of the Aviation Academy in Belgrade). For all three *Aviation News* websites, the search is simple and performed in the same way. It is supposed to enter a query in the search field and get an answer. The response is presented in the form of text and images, and in most cases with a link to other pages and labels and therefore a large number of attachments can impair the visibility and functionality of the site. There is no registration on the site, but there is a subscription to the newsletter and contact form.

3.6 Network sources in Serbia

Digitalna NBS (Digital National Library of Serbia, hereinafter *Digital NLS*)[23] is an online resource of the National Library of Serbia which began digitization in 2004 (Butigan-Vučaj 2003). The search is performed on the basis of metadata and on the basis of words from the text. One can make thorough examination using advanced search which is performed by assigning values to metadata attributes such as: title, author, topic and identifier. It is also possible to enter the name of publisher, place of publication, year

22. Aviation museum, Airshow reports, Paris Airshow / (accessed 13. 1. 2022.)

23. Previous version of Digital NLS, available on (Narodna biblioteka Srbije, n.a.).

or range of years relevant to the search and select the language in which the publication was published, select the type of library material (monographic publications, serial publications, articles or other components of the publication, sound recordings, cartographic material, music material, pictorial material, manuscript book and manuscript material) or enter value of attributes in "name and surname of the collection owner". Finally, search can be performed by words from the text, which are introduced later and is much better solution that contributes to the quality of the digital library. *Digital NLS* contains digitized publications and non-book library and information material in the field of aviation. On the *Digital NLS*, you can also find photos where you can see, for example, the announcement of the upcoming air rally. The document is accompanied by a simple description, with only two attributes. These are the title and identification number of the record in the bibliographic-catalog database COBISS.SR or COBISS.SR-ID, that is, the values of the two attributes. Attribute values, for example, may be: Belgrade, Terazije, Terazié, or 95356428.[24] ID 95356428 refers to a bibliographic record that provides a broader description using the following attributes: comparative title, language, type of material, year, issue and production, physical description, notes, subject terms, UDC number, link and abstract. Digitized monographic publications are associated with several attributes: title, subject, author, publisher or printing agency, date, language and COBISS.SR-ID. A more detailed metadata view contains the following attributes: title translation, scope, format, and spatial determinant. The document is accompanied by the attribute: Collection, with the value: Books, which means that it is a collection of digitized monographic publications. In addition, digitized domestic newspapers are potential source on the history of aviation. A significant number of articles on aviation can be found in searchable newspapers, for example, in *Политика* (Politika) (Injac 2006).[25] Thus, important events for the aviation were covered. For the purposes of the paper, only some of them will be listed: the crash of the Merćep-Rusjan plane near the Kalemegdan fortress (Belgrade) on the third day of Christmas in 1910,[26] than, the first registered plane crash in civil aviation which occurred on the Paris-London route on April 7th, 1922 and the first domestic aviation exhibition, held in Belgrade in 1925, as well as the first international aviation exhibition

24. There are many photos of Yugoslav Airforce on Digital NLS. They are available on: Balkan war 1912-1913. 1/25, Balkan war 1912-1913. 1/27, Balkan war 1912-1913. 1/28, History of Yugoslav Aiation (accessed 15. 1. 2022).

25. Politika (accessed 15. 1. 2022)

26. The Julian calendar.

in Belgrade in 1938. A search performed within this resource indicates domestic digitized newspapers as a possible resource. *Digital NLS* is functional and user-friendly. No registration form on the site.

There are works that deal with the history of aviation which can be find in social science, culture and technology journals. The Serbian Library Consortium for Coordinated Acquisition (hereinafter KoBSON) (Kosanović 2003) is an excellent source of scientific papers in the field of aviation. KoBSON may provide high response but low precision. The keyword search, through the EBSCO Discovery Service was performed as follows. The following inquiries provided a great recall (1,277,719, 1,261,374, 111,588 and 1,181 hits) in the form of articles from journals, that is, serial and monographic publications:

- AVIATION HISTORY,
- AVIATION AND HISTORY,
- AVIATION HISTORY,
- AVIATION HISTORY with options: Full-Text (online), Peer Reviewed.

By applying the appropriate strategy, it is possible to obtain precise answer. Although the time interval 1900-2021 was set, narrowing of the search (to scientific journals and applying the subject term AVIATION HISTORY) , produced way less number of hits (117) but increased relevance.

Papers dealing with the history of domestic aviation can often be found in journals such as *Istorija 20. veka* (Journal of the Institute of Contemporary History)[27] and *Наслеђе* (Journal *Heritage*)[28] which are both on the DOAJ (*Directory of Open Access Journals*) list. For example: Bojan B. Dimitrijević[29] stands out among the authors of works from the history of domestic aviation in the *Istorija 20. veka*. In Наслеђе, with several articles in the

27. Istorija 20. veka (Cobiss); The journal is publicly available online and has following metadata: Institut za savremenu istoriju, History (General) and History of Europe: History of Eastern Europe, 20th century, Balcan,contemporary history, Serbia, Yugoslavia, World War 2 (accessed 19.02.2022)

28. Nasleđe (Cobiss); The journal is publicly available online and has following metadata: Zavod za zaštitu spomenika kulture grad Beograda, Fine arts: Architecture, architecture, Belgrade, conservation, cultural monuments, heritage (accessed 19.02.2022).

29. Bojan B. Dimitrijević appears as an author and co-author of articles, reviews and other contributions in *Istorija 20. veka*; he is active in the editorial board of the journal Istorija 20. veka (приступљено 20. 2. 2022). He is author, co-author and reviewer of some monographs as well. More information Bojan Dimitrijević (Cobiss) (accessed 20. 2. 2022).

field of architecture, in which the topic of the buildings were aviation institutions, especially the aviation industry, the authors Biljana Mišić, Hajna Tucić and Saša Mihajlov must not be left out.[30] One of the latest works from the history of domestic aviation, which is on the DOI list, is the article by Ilija Kukobat "The beginnings of air transportation in post-war Yugoslavia 1945–1947", published in *Istorija 20. veka*.[31]

4 A possible model of a digital library on the history of aviation

According to previous experiences with websites presenting the history of aviation, as well as *Europeana* and *Digital NLS*, the digital library of the history of domestic aviation could be all-in-one, consisting of all the mentioned web resources. First of all, there should be a service for communication with users, i.e. site visitors, for the purpose of professional consultations as well as improving the quality of other services. Furthermore, following the example of the *Digital Collections of the WWII Museum*, the digital library could become the initiator of the oral history of domestic aviation. The search in the library may be performed on metadata as well as on all words in text type documents and the characteristics of *Avialogs* in terms of conciseness, *Europeana* in terms of encompassing various forms of cultural heritage artifacts, *Digital NBS* in terms of focus on the preservation of national heritage, *Digital Collections of the WWII Museum* in terms of oral history and organization of searches, metadata and the existence of a kind of terminological dictionary, *Air Team Images* in all aspects of aviation photography.

This library could also function as a virtual institute for the history of aviation in Serbia.

Regarding that the digital library could become a significant source for researchers in the field of domestic aviation history, a large number of description elements, that is, metadata would also be important.The following attributes would be welcomed: the owner of collection or single physical object, the name of collection, title, the title translated in English, id, author, spatial determinant (subject term), time determinant (subject term), scope, description, UDC classification, subject classification, annotation, medium,

30. Biljana Mišić, Hajna Tucć, Saša Mihajlov – articles in the journal *Nasleđe*

31. Ilija Kukobat - the article from the journal *Istorija 20. veka* Vol. 38, no. 2/2020, pp. 173 – 186.

format, date of digitization, as well as aircraft type and version, airport (airfield), institution, event (manifestation), contributor (associate, predecessor, successor, family member).

The most important title in the digital library of the history of aviation in Serbia would be History of the Yugoslav Air Force (Микић 1933), the monograph that could be linked along with the memoirs of the same author (Микић 2000). Considering serial publications, it would be good to digitize *Vazduhoplovni glasnik* (Aviation Herald)[32] with its two appendices. The digital library would use links to other network locations of facilities related to the history of domestic aviation and cooperate with organizations that digitize history of domestic aviation artifacts, as well as with owners of private collections or family collections. Cooperation with the RS Ministry of Defense would be especially welcomed in terms of joint activities of the Military Archives and the *Aviation Museum* in Belgrade, in terms of digitization and promotion of archival and museum materials that are of great importance for the history of domestic aviation (Macura 2017). In case of digital libraries, teamwork and cooperation are taken for granted, thus establishing a kind of virtual community of admirers of the history of domestic aviation.

5 Conclusion

Digital, that is, network sources show the history of aviation in different ways. The content on the website is greatly influenced by different target groups of visitors: the professional public, the general public and those who are in the process of education. There are websites created by aviation professionals, enthusiasts and aviation enthusiasts themselves, as individuals, groups or societies. There are also academic resources, that is, websites of universities and cultural institutions that contain material from the history of aviation. Websites that present the history of aviation do so for educational or recreational purposes or to present their content as scientific information. The first electronic sources dealing with the history of aviation dates back to mid-1990s while organizations and associations have not stopped posting content from the history of aviation to this day. Each of the mentioned digital sources represents the history of aviation in its own way, but among them there is no source that according to its elements represents a digital library of aviation history. In that sense, *Avialogs* stands out from

32. Vazduhoplovni glasnik (Cobiss)

the resources mentioned in the text. Although declared as an online library and community, *Avialogs* is closest to the idea of digital aviation history library. However, *Avialogs* does not offer a wide range of services but focuses on digitizing publications, accessing digital collections and possibly reproducing them for non-commercial purposes (printing or downloading). The new version of the site intend to provide a full-text search, a more detailed description by including more metadata, preferably based on some of the existing standard bibliographic descriptions, as well as an aviation history consulting service. Being enhanced this way, *Avialogs* would fulfill the purpose of a digital library. It would provide a range of different information services to those who access, whether they are the aviation professionals or just students, researchers, lecturers or general public.

References

Butigan-Vučaj, Tamara. 2003. "Digitalna Narodna biblioteka Srbije (DNBS)." *Pregled Nacionalnog centra za digitalizaciju,* no. 1, 40–42.

IFLA/UNESCO. 2012. *Manifesto for Digital Libraries – Манифест за дигиталне библиотеке.* Accessed 3. 4. 2022. https://cdn.ifla.org/wp-content/uploads/files/assets/digital-libraries/documents/ifla-unesco-digital-libraries-manifesto-sr.pdf.

Injac, Vesna. 2006. "Narodna biblioteka Srbije u projektu "Evropska biblioteka" (TEL–The European Library)." *Glasnik Narodne biblioteke Srbije,* no. 1, 37–45.

Kosanović, Biljana. 2003. "Konzorcijum biblioteka Srbije za objedinjenu nabavku: da li dovoljno koristimo elektronske izvore informacija." *Infoteka* 4 (2): 107–113.

Macura, Ljiljana. 2017. "Zbirke vazduhoplovnih institucija u očuvanju kulturne baštine: Muzej vazduhoplovstva - Beograd." In *Programi zaštite, promocije i približavanja kulturne baštine široj javnosti : zbornik radova = Programs of protection, promotion and convergence of cultural heritage to general public : conference proceedings.* Sarajevo: Asocijacija informacijskih stručnjaka, bibliotekara, arhivista i muzeologa = Assiciation of Information Professionals – Librarians, Archivists / Museologists.

Микић, Сава Ј. 1933. *Историја југословенског ваздухопловства: са 615 слика у тексту од којих две Њ. В. Краља и Престолонаследника Петра у бојама.* Стр. V-VII: Предговор / Мил. Ј. Николајевић. Штампарија Драг. Грегорића.

Микић, Сава Ј. 2000. *Недовршени мемоари.* Стр. 9-10: Предговор / Александра Вранеш. Стручна књига.

Тртовац, Александра. 2017. *Проналажење информација у дигиталним библиотекама.* Vol. 4. Универзитетска библиотека "Светозар Марковић". ISBN: 978-86-7301-103-5.

Web sources

Aeroflight. n.a. *Aeroflight – Home page.* www.aeroflight.co.uk. Editor John Hayles; accessed 11. 1. 2022.

Aeroflight Cloud. n.a. *Aeroflight Cloud – Home page.* https://docs.google.com/presentation/d/10wZReN2_38bvTBAFLYJiCpgSFItpYD0Y4UvCnWLbako/edit#slide=id.g9d40c95f98_0_25. Editor John Hayles; accessed 11. 1. 2022.

AirTeamImages. n.a. *Aviation Image Library.* www.airteamimages.com. Accessed 6. 2. 2022.

Avialogs. n-a-. *Aviation Library.* https://www.avialogs.com/. Created by Benoit de Mulder; accessed 11. 1. 2022.

Aviation History Online Museum. 1996. *Aviation History Online Museum - Home page.* http://www.aviation-history.com. Web site creator Larry Dwyer; accessed 18. 1. 2022.

Aviation News. n.a. *International aviation news, airshow reports, airceaft facts, worlds largest aviation museum dabase.* https://aviationnews.eu/. Accessed 13. 1. 2022.

Flight Journal. n.a. *Flight Journal (online).* https://www.flightjournal.com/. Accessed 17. 1. 2022.

FltCtrl Ltd. n.a. *Flock aero – Home page.* https://flock.aero/. Accessed 17. 1. 2022.

Flugzeugbilder. n.a. *Aviation photography at its best.* https://www.flugzeugbilder.de/v3/index.php. Accessed 17. 1. 2022.

Global Security. n.a. *Global Security (online).* https://www.globalsecurity.org/index.html. Accessed 18. 1. 2022.

HistoryNet. n.a. *Aviation History.* https://www.historynet.com/magazine/aviation-history/. Accessed 18. 1. 2022.

Narodna biblioteka Srbije. n.a. *Digitalna Narodna biblioteka Srbije.* https://digitalna.nb.rs//. Accessed 15. 1. 2022.

Royal Aeronautical Society. n.a. *Heritage Collection.* https://aerosocietyheritage.com/. Accessed 20. 1. 2022.

The Latin American Aviation Historical Society. n.a. *The Latin American Aviation Historical Society.* https://www.laahs.com/. Accessed 19. 1. 2022.

The National WWII Museum. n.a. *Digital Collections of the WWII Museum.* https://www.ww2online.org/. Accessed 13. 1. 2022.

UWM Libraries Digital Collections. n.a. *Aviation History Collection – Selections from the George Hardie Collection.* https://uwm.edu/libcollections/aviation_history/. Accessed 11. 1. 2022.

Europeana Data Model: The Key to Cooperation with Europeana

UDC 025.3:004.9
DOI

ABSTRACT: Europeana Data Model (EDM) is a comprehensive data model created with the idea of presenting versatile digital objects within a single metadata schema to ensure interoperability. It is based on the principle of linked open data and the semantic web. The paper deals with the application of this model in the National Library of Serbia through direct cooperation with Europeana, within which the library performs the role of a national aggregator for Europeana, as well as a data provider.
KEYWORDS: EDM, digital library, Europeana, metadata, metadata aggregator.

PAPER SUBMITTED: 21 March 2022
PAPER ACCEPTED: 7 June 2022

Tatjana Domazet
tatjana.domazet@nb.rs

Nataša Petrović
natasa.petrovic@nb.rs

Tamara Butigan Vučaj
tamara@nb.rs

National Library of Serbia
Belgrade, Serbia

1 Introduction

Europeana is the most important initiative and the greatest investment of the European Commission into the digitisation of cultural heritage. It was founded in 2008 to transform the world with culture and strengthen the European identity among its citizens through common cultural heritage in the digital form. The core of Europeana is its platform,[1] which provides access to over 51 million digital objects from more than 3,500 libraries, museums, archives and galleries across Europe (March 2022). The Europeana Initiative is an ecosystem consisting of three interconnected expert organisations: Europeana Foundation, Europeana Aggregators' Forum and Europeana Network Association. The Europeana Foundation, seated at the Royal Library of the Netherlands, is an independent, non-profit organisation with around 60 employees engaged in the development of the Europeana platform and

1. Europeana, „Welcome to Europeana"

cooperation with other cultural heritage digitisation initiatives. The Europeana Aggregators' Forum consists of 40 accredited metadata aggregators, out of which 13 are domain and thematic and 27 are national and regional. The Europeana Network Association is an open and democratic community of experts involved in the digitisation of cultural heritage. It boasts over 3,000 experts from cultural heritage institutions, the IT sector and educational and research institutions. The Europeana mission statement says: "Europeana empowers the cultural heritage sector in its digital transformation. We develop expertise, tools and policies to embrace digital change and encourage partnerships that foster innovation. We make it easier for people to use cultural heritage for education, research, creation and recreation. Our work contributes to an open, knowledgeable and creative society." The Europeana vision statement says: "Europeana imagines a cultural heritage sector powered by digital and a Europe powered by culture, giving it a resilient, growing economy, increased employment, improved well-being and a sense of European identity."

Since Europeana was not envisioned as a repository of digital content, but rather as an aggregator of its metadata, the Europeana Data Model (EDM) – the metadata schema on which the database in its background is based – is the backbone of its platform. The EDM enables the interoperability of metadata originating from different institutions and is the least common denominator for all metadata schemas used by different cultural institutions to describe their digital objects, such as METS in libraries, LIDO in museums or EAD in archives.

Since Europeana's foundation, the National Library of Serbia (Narodna biblioteka Srbije – NBS) has been cooperating with it as a metadata provider. Initially, the aggregation was performed through The European Library (TEL), which acted as an aggregator for all the European national libraries until its closure in December 2016. In 2015, the NBS began working on establishing the national aggregator. The same year, the Data Exchange Agreement (DEA) was signed between the library and Europeana. The aggregation infrastructure (Aggregator for Europeana) was developed in partnership with the Slovenian company Semantika (Butigan-Vučaj 2019). The Aggregator, which is housed at the Digitisation Department of the National Library of Serbia, has been operational since 2018. That year, the National Library of Serbia became part of the Europeana Aggregators' Forum, having previously fulfilled all the necessary conditions to obtain the status of an accredited aggregator and a trusted partner of Europeana. The Aggregator for Europeana has been recognised as a provider of quality metadata, and it

has also achieved admirable standards in terms of content description, metadata mapping, copyright licence attribution for digital objects, accessibility, multilingualism and the use of controlled vocabularies. It has also become a source of information, guidance and support in the digitisation process. In addition to all this, the Aggregator for Europeana is the only aggregator operating from a non-EU country. So far, the NBS has submitted datasets in the EDM format in preparation for their publication on the Europeana platform as part of the following European projects: Europeana Collections 1914-1918, Collections of South and East Europe in Europeana (CSEEE), The Rise of Literacy, Migration in the Arts and Sciences and Europeana Common Culture. The European Library was the metadata aggregator for the first two projects, whereas the NBS Aggregator for Europeana was already operational during the remaining three projects. The Europeana Common Culture project,[2] which involved 23 aggregators from 21 European countries, was vital for metadata standardisation and the improvement of the entire aggregation infrastructure. As part of it, the NBS maximised the quality of its metadata for all the datasets that had been published on Europeana up to that point.

2 Europeana Data Model (EDM)

Europeana provides access to versatile materials coming from different cultural heritage institutions across Europe. Those institutions rely on different metadata standards, which mainly depend on the type of material. All those resources are extremely important. However, the diversity of data formats, as well as the efforts to preserve as much information as possible, with as little data overlap, gave rise to a need to establish a common metadata standard, which would be easily applicable to the materials stored in libraries, museums, archives and institutions in possession of audio-visual collections. As a result, it was decided to introduce a common metadata model – one that would enable data linking to facilitate search based on certain criteria.

Europeana Semantic Elements[3] (ESE) was the early stage in the development of this idea. Its further elaboration led to the establishment of the Europeana Data Model[4] (EDM) (Drege 2019; Klimpel and Ojler 2019).

2. Europeana, "Europeana Common Culture", accessed 18. 3. 2022

3. Europeana, „Europeana Semantic Elements Documentation", accessed 17. 3. 2022

4. Europeana, „Europeana Data Model", accessed 17. 3. 2022

EDM[5] constitutes an improvement of the previous model in that it enables the integration of semantic elements and descriptions of hierarchical structures. Also, metadata can now be connected with the digital reproductions of the material they relate to. Besides, EDM offers the possibility of free text entry to ensure the highest content description quality possible. EDM also enables the entry of values that relate only to the digital copies in addition to those that relate to the physical objects.

As EDM combines materials from different cultural institutions, it is a much better overall solution compared to individual standards compatible with specific types of material. When converting metadata into EDM, there is no data loss. EDM is based on an open framework enabling participation in the semantic web. Thanks to it, the data related to persons, space and subject areas can be connected regardless of their provenance. This stimulates mutual data enrichment between institutions and facilitates the discovery of new resources from the perspective of users interested in a determined topic, which, in turn, leads to an increase in the traffic to cultural heritage objects kept at different institutions.

The EDM format is based on the following standards: OAI ORE (Open Archives Initiative Object Reuse & Exchange),[6] SKOS (Simple Knowledge Organization System),[7] DC (Dublin Core)[8] and CIDOC-CRM.[9] Furthermore, it also relies on various normative databases and dictionaries. That way, new connections between objects are established and language barriers are overcome. Through EDM, data is collected, linked and enriched. Since it is based on open data and the semantic web, cultural heritage from institutions which cooperate with Europeana can reach a wider audience. The semantic web provides a broader metadata context and enables the representation of complex relationships between different objects as well as different parts of an object while encouraging data reuse and reliance on external resources available as related data. The principles of the semantic web have been implemented in the form of the RDF[10] (Resource Description Framework) standard.

5. Europeana, „Europeana Data Model – Mapping Guidelines v2.4“, accessed 17. 3. 2022
6. „Open Archive Initiative - Object Exchange and Reuse“
7. „SKOS Simple Knowledge System Namespace Document“
8. „DCMI Metadata Terms“
9. „Classes & Properties Declarations of CIDOC-CRM“
10. World WideWeb Consortium „RDF Primer“

EDM consists of three main classes. They represent three different parts of a single record. The <edm:ProvidedCHO> class refers to a physical object. It contains the data about the author(s), publisher, year of publication, format, publication language, unique object identifier and, also, geographic and temporal headings. The <ore:Aggregation> class contains data on the object, the institution where the object is kept and the institution which delivered the data to Europeana. This class also contains links to the digitised object and the attributed licence, which serves to let the users know how the digitised material can be used. The <edm:WebResource> class refers to the digital reproduction of the object. In addition to those three main classes, some other classes can be included in a record to represent entities that are somehow related to the object: <edm:Agent>, <edm:Place>, <edm:TimeSpan>, <skos:Concept>, etc. The classes contain properties with descriptive elements. The classes and their properties are defined based on RDFS (Resource Description Framework Schema)[11] and OWL (Web Ontology Language)[12] elements.

The National Library of Serbia uses the COMARC/B metadata standard (Figure 1) within the COBISS platform for the metadata related to the materials published on its digital presentation (Digitalna NBS). For data delivery to Europeana, the data is exported from COBISS and converted into the EDM format using various internal tools for data verification, enrichment, refinement and conversion. They facilitate data transfer from the relevant fields of a comarc.xml record to the required fields within an EDM record. Then, the obtained EDM record is uploaded to the Aggregator for Europeana, which performs a further transformation, after which the record is ready to be harvested by Europeana and published on the platform.

The <ore:aggregates> element contains all the data related to a digitised object (a book, an issue of a periodical or a photograph). The <edm:ProvidedCHO> class contains the link toward the digital object at Digitalna NBS[13] (Europeana does not store the digitised material from cultural heritage institutions. Instead, it presents the metadata and directs the users to the websites of those institutions via links). The <dc:creator> property relates to the author(s). Since the integration of the normative database of authors' names (CONOR)[14] into the COBISS system, two <dc:creator> properties have been introduced into EDM (Cyrillic and Latin; also, the

11. „RDF Schema 1.1“
12. „owl:Ontology“
13. National library of Serbia, "Digitalna NBS"
14. Izum, "CONOR normative database"

```
<ore:aggregates>
    <edm:providedCHO rdf:about="http://www.digitalna.nb.rs/wb/NBS/Tematske_kolekcije/Industrijsko_nasledje/borski_rudnici/Krvavo_zlato/GM_85">
        <dc:contributor xml:lang="sr">Јован Поповић (1905-1952)</dc:contributor>
        <dc:contributor xml:lang="sr">Јован Поповић (1905-1952)</dc:contributor>
        <dc:contributor rdf:resource="http://www.wikidata.org/wiki/Q6296683" xml:lang="sr"/>
        <dc:creator xml:lang="sr">Ђорђе Андрејевић-Кун (1904-1964)</dc:creator>
        <dc:creator xml:lang="sr">Ђорђе Андрејевић-Кун (1904-1964)</dc:creator>
        <dc:creator rdf:resource="http://www.wikidata.org/wiki/Q1283347" xml:lang="sr"/>
        <dc:description xml:lang="sr">Тираж 250.</dc:description>
        <dc:description xml:lang="sr">Str. 1-11: Umjetnik i njegovo delo / J. Popović.</dc:description>
        <dc:format xml:lang="sr">12 str., 28 listova s tablama</dc:format>
        <dc:format xml:lang="sr">drvorezi, crno-belo</dc:format>
        <dc:format xml:lang="sr">25 cm</dc:format>
        <dc:identifier>170197516</dc:identifier>
        <dc:language>sr</dc:language>
        <dc:publisher xml:lang="sr">Đ. Andrejević-Kun</dc:publisher>
        <dc:subject xml:lang="sr">Ђорђе Андрејевић-Кун</dc:subject>
        <dc:subject xml:lang="sr">Индустријско наслеђе</dc:subject>
        <dc:subject xml:lang="en">Industrial Heritage</dc:subject>
        <dc:title xml:lang="sr">Krvavo zlato : 28 originalnih drvoreza Đ. Andrejevića-Kuna</dc:title>
        <dcterms:issued>1937</dcterms:issued>
        <edm:currentLocation rdf:resource="http://sws.geonames.org/792680/"/>
        <edm:type>TEXT</edm:type>
        <dc:type xml:lang="sr">књига</dc:type>
    </edm:providedCHO>
    <ore:Aggregation rdf:about="http://www.digitalna.nb.rs/wb/NBS/Tematske_kolekcije/Industrijsko_nasledje/borski_rudnici/Krvavo_zlato/GM_85">
        <edm:aggregatedCHO rdf:resource="http://www.digitalna.nb.rs/wb/NBS/Tematske_kolekcije/Industrijsko_nasledje/borski_rudnici/Krvavo_zlato/GM_85"/>
        <edm:provider>National Library of Serbia</edm:provider>
        <edm:dataProvider>National Library of Serbia</edm:dataProvider>
        <edm:rights rdf:resource="http://creativecommons.org/licenses/by-nc-sa/4.0/"/>
        <edm:isShownAt rdf:resource="http://www.digitalna.nb.rs/wb/NBS/Tematske_kolekcije/Industrijsko_nasledje/borski_rudnici/Krvavo_zlato/GM_85"/>
        <edm:isShownBy rdf:resource="http://www.digitalna.nb.rs/attach/NBS/Tematske_kolekcije/Industrijsko_nasledje/borski_rudnici/Krvavo_zlato/GM_85/output.pdf"/>
        <edm:Object rdf:resource="http://www.digitalna.nb.rs/wb/NBS/Tematske_kolekcije/Industrijsko_nasledje/borski_rudnici/Krvavo_zlato/GM_85?pageIndex=thumb"/>
    </ore:Aggregation>
</ore:aggregates>
```


Figure 1. An example of a record from Digitalna NBS after the conversion from COMARC/B into EDM

names of foreign authors are written etymologically). If available, the birth and death years are also included. One of those fields contains the link to the Wikidata[15] database entry related to the author. Subsidiary authors (contributors) are placed within <dc:contributor>. The <dc:format> property is also repeatable and contains information about the physical copy of the material (dimensions, number of pages, etc.) and so is <dc: description>, in which the data on circulation and certain descriptive elements are entered, depending on availability. Subject headings are entered into <dc:subject> in Serbian and/or English for better data linking (language is marked within the xml:lang attribute). The <dc:identifier> property is a unique object identifier (the NBS uses the COBISS ID for this purpose). The metadata language is entered as xml:lang (a mandatory attribute for textual metadata), whereas <dc:language> is the language of the publication (for multi-language publications, this field is repeatable). Two-letter language codes are used, as per the ISO 639-1 standard.[16] The publisher is placed within <dc:publisher> and <dc:title> contains the object title. The <dcterms:issued> property is the year of publication, and <edm:currentLocation> is the place where the physical copy of the material is kept. This field contains a link to the

15. „Wikidata“
16. Wikipedia, „List of ISO 639-1 codes“

Geonames[17] database. Defined values "TEXT" or "IMAGE" are entered into <edm:type>, depending on the type of material.

The <ore:Aggregation> class contains links to the webpage of the digitised object. The link toward the object is entered into <ore:Aggregation>, <edm:aggregatedCHO> and <edm:isShownAt>. The <edm:isShownBy> property contains the link to the PDF file (for textual material) or the full screen view of the image viewer. The <edm:object> property contains the link to the thumbnail. In addition to those links, the class also contains the <edm:rights> property, in which the assigned Creative Commons licence[18] link is entered. The <edm:provider> property refers to the aggregator through which the data was submitted to Europeana, while <edm:dataProvider> is for the library that submitted the data to the aggregator.

Once ready, the EDM dataset is uploaded to the Aggregator for Europeana, which carries out an additional transformation after which the dataset is ready for publication on Europeana (Figure 2).

3 Europeana Publishing Framework

Europeana has established content and metadata description standards. The quality of content and metadata determines the way an object can be presented and promoted on the Europeana portal and defines if it can be downloaded. There are three levels of metadata[19] quality: A, B and C.[20] The A level is the lowest, whereas the C level is the highest. Europeana's principle is "The more you give, the more you get",[21] i.e. the richer the data, the better the search results. The A level means the content can be found on the pages of Europeana and the B level means it can also be used for different thematic collections and promotional activities (e.g. social media posts). The C level is the highest, meaning that it has the richest data. In addition to all that applies to the previous two levels, the C-level metadata can be used for various projects, applications and services. The records the NBS has so far delivered to Europeana predominately have the B-level metadata quality.

17. „GeoNames"
18. „Creative Commons"
19. Europeana, „Europeana Publishing Framework: Metadata", accessed 15. 3. 2022
20. Tier A, Tier B, Tier C.
21. Europeana, „The more you give, the more you get", accessed 18. 3. 2022

```
<OAI-PMH xmlns="http://www.openarchives.org/OAI/2.0/" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/
 http://www.openarchives.org/OAI/2.0/OAI-PMH.xsd">
   <responseDate>2022-03-14T10:50:18Z</responseDate>
   <request metadataPrefix="edm" set="ECC_kun1_EDM_okt2_20_NP" verb="ListRecords">http://agregator.nb.rs/oai/OAIHandler</request>
  <ListRecords>
    <record>
      <header>
        <identifier>2354823</identifier>
        <datestamp>2020-10-19T00:00:00Z</datestamp>
      </header>
      <metadata>
        <rdf:RDF xmlns:ore="http://www.openarchives.org/ore/terms/" xmlns:owl="http://www.w3.org/2002/07/owl#" xmlns:skos="http://www.w3.org/2004/02/skos/core#"
          xmlns:rdaGr2="http://rdvocab.info/ElementsGr2/" xmlns:edm="http://www.europeana.eu/schemas/edm/" xmlns:ns13="http://www.w3.org/ns/ddr1/2/"
          xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#" xmlns:ns12="http://xmlns.com/foaf/0.1/" xmlns:ns11="http://rdfs.org/sioc/services#"
          xmlns:ns10="http://www.ebu.ch/metadata/ontologies/ebucore/ebucore#" xmlns:dcterms="http://purl.org/dc/terms/" xmlns:ns16="http://www.europeana.eu/schemas/edm/enrichment/"
          xmlns:ns15="http://usefulinc.com/ns/doap#" xmlns:ns14="http://creativecommons.org/ns#" xmlns:dc="http://purl.org/dc/elements/1.1/">
          <skos:Concept rdf:about="/direct/2354833">
            <skos:prefLabel xml:lang="sr">Индустријско наслеђе</skos:prefLabel>
          </skos:Concept>
          <edm:Agent rdf:about="/direct/2354828">
            <skos:prefLabel xml:lang="sr">Јован Поповић</skos:prefLabel>
            <skos:note xml:lang="en">Latin script</skos:note>
            <rdaGr2:dateOfBirth>1905</rdaGr2:dateOfBirth>
            <rdaGr2:dateOfDeath>1952</rdaGr2:dateOfDeath>
          </edm:Agent>
          <skos:Concept rdf:about="/direct/2354826">
            <skos:prefLabel xml:lang="sr">Ђорђе Андрејевић-Кун</skos:prefLabel>
          </skos:Concept>
          <edm:Agent rdf:about="/direct/2354829">
            <skos:prefLabel xml:lang="sr">Ђорђе Андрејевић-Кун</skos:prefLabel>
            <skos:note xml:lang="en">Cyrillic script</skos:note>
            <rdaGr2:dateOfBirth>1904</rdaGr2:dateOfBirth>
            <rdaGr2:dateOfDeath>1964</rdaGr2:dateOfDeath>
          </edm:Agent>
          <edm:Agent rdf:about="/direct/2354832">
            <skos:prefLabel xml:lang="sr">Јован Поповић</skos:prefLabel>
            <skos:note xml:lang="en">Cyrillic script</skos:note>
            <rdaGr2:dateOfBirth>1905</rdaGr2:dateOfBirth>
            <rdaGr2:dateOfDeath>1952</rdaGr2:dateOfDeath>
          </edm:Agent>
          <edm:Agent rdf:about="/direct/2354827">
            <skos:prefLabel xml:lang="sr">Đorđe Andrejević-Kun</skos:prefLabel>
            <skos:note xml:lang="en">Latin script</skos:note>
            <rdaGr2:dateOfBirth>1904</rdaGr2:dateOfBirth>
            <rdaGr2:dateOfDeath>1964</rdaGr2:dateOfDeath>
          </edm:Agent>
          <edm:Agent rdf:about="/direct/2354830">
            <skos:prefLabel xml:lang="sr">http://www.wikidata.org/wiki/Q1283347</skos:prefLabel>
          </edm:Agent>
          <skos:Concept rdf:about="/direct/2354825">
            <skos:prefLabel xml:lang="sr">књига</skos:prefLabel>
          </skos:Concept>
          <edm:Agent rdf:about="/direct/2354824">
            <skos:prefLabel xml:lang="sr">http://www.wikidata.org/wiki/Q6296683</skos:prefLabel>
          </edm:Agent>
          <skos:Concept rdf:about="/direct/2354831">
            <skos:prefLabel xml:lang="en">Industrial Heritage</skos:prefLabel>
          </skos:Concept>
          <edm:ProvidedCHO rdf:about="URN:RS:NAE:d27476c7-f752-4a15-901b-e7e802b5f7eecho">
            <dc:identifier>URN:RS:NAE:d27476c7-f752-4a15-901b-e7e802b5f7eecho</dc:identifier>
            <dc:identifier>170197516</dc:identifier>
            <dc:language>sr</dc:language>
            <dcterms:issued>1937</dcterms:issued>
            <edm:currentLocation rdf:resource="http://sws.geonames.org/792680/"/>
            <dc:title xml:lang="sr">Крвavo zlato : 28 originalnih drvoreza Đ. Andrejevića-Kuna</dc:title>
            <dc:description xml:lang="sr">Tiraž 250.</dc:description>
            <dc:description xml:lang="sr">Str. 1-11: Umjetnik i njegovo delo / J. Popović.</dc:description>
            <dc:publisher xml:lang="sr">Đ. Andrejević-Kun</dc:publisher>
            <dc:format xml:lang="sr">12 str., 28 listova s tablama</dc:format>
            <dc:format xml:lang="sr">drvorezi, crno-belo</dc:format>
            <dc:format xml:lang="sr">25 cm</dc:format>
            <dc:subject rdf:resource="/direct/2354833"/>
            <dc:contributor rdf:resource="/direct/2354828"/>
            <dc:subject rdf:resource="/direct/2354826"/>
            <dc:creator rdf:resource="/direct/2354829"/>
            <dc:contributor rdf:resource="/direct/2354832"/>
            <dc:creator rdf:resource="/direct/2354827"/>
            <dc:creator rdf:resource="/direct/2354830"/>
            <dc:type rdf:resource="/direct/2354825"/>
            <dc:contributor rdf:resource="/direct/2354824"/>
            <dc:subject rdf:resource="/direct/2354831"/>
            <edm:type>IMAGE</edm:type>
          </edm:ProvidedCHO>
          <ore:Aggregation rdf:about="URN:RS:NAE:d27476c7-f752-4a15-901b-e7e802b5f7eecho">
            <edm:aggregatedCHO rdf:resource="http://www.digitalna.nb.rs/wb/NBS/Tematske_kolekcije/Industrijsko_nasledje/borski_rudnici/Krvavo_zlato/GM_85"/>
            <edm:dataProvider>National Library of Serbia</edm:dataProvider>
            <edm:isShownAt rdf:resource="http://www.digitalna.nb.rs/wb/NBS/Tematske_kolekcije/Industrijsko_nasledje/borski_rudnici/Krvavo_zlato/GM_85"/>
            <edm:isShownBy rdf:resource="http://www.digitalna.nb.rs/attach/NBS/Tematske_kolekcije/Industrijsko_nasledje/borski_rudnici/Krvavo_zlato/GM_85/output.pdf"/>
            <edm:object rdf:resource="http://www.digitalna.nb.rs/wb/NBS/Tematske_kolekcije/Industrijsko_nasledje/borski_rudnici/Krvavo_zlato/GM_85?pageIndex=thumb"/>
            <edm:provider>National Library of Serbia</edm:provider>
            <edm:rights rdf:resource="http://creativecommons.org/licenses/by-nc-sa/4.0/"/>
          </ore:Aggregation>
        </rdf:RDF>
      </metadata>
    </record>
  </ListRecords>
</OAI-PMH>
```


Figure 2. The same record after the transformation carried out by the Aggregator for Europeana

When it comes to the content,[22] the same principle applies. The better the quality, the better the reach of the content and the possibilities of its reuse. There are four levels of content quality: 1, 2, 3 and 4.[23] Just like with metadata, the lowest level only enables the objects to be found on Europeana, whereas the highest level enables reuse for commercial purposes, as well as the use on different platforms and applications. For some content to achieve level 4, it must have a licence enabling free use and reuse (PDM, CC-0, CC-BY, CC-BY-SA) (Kasadeval and Butigan Vučaj 2020). The content the NBS has delivered so far is mainly tier 3 or 4 (depending on the type of licence). Another factor influencing content quality is the presence of PDF files for textual materials (NBS implemented optical character recognition (OCR) as a mandatory step for textual materials in the PDF format). The library also adheres to the rule for the scanned material to be in the JPG format, resolution 0.95 MPIX, i.e. minimum size 1200x800 pixels[24] (which is a precondition for the tiers 3 and 4).

Since Europeana collects data from cultural heritage institutions across Europe and presents them within a single platform, it aims to minimise copyright limitations standing in the way of digitisation and free access. When it comes to copyright, the key document is Europeana Licensing Framework,[25] which serves to standardize and harmonize copyright-related information and practices so that the relationship between the end-users and data-submitting institutions is clear and transparent. Europeana uses the Creative Commons licences and the Rights Statements[26] licences, for which it is a co-founder.

4 Aggregation Infrastructure Improvement

The Europeana Common Culture project, which lasted from the beginning of 2019 until the end of 2020, was focused on the improvement of the aggregation infrastructure. As we have already mentioned, the project aimed at establishing common recommendations and standards, enriching metadata and enabling free access to the content, as much as possible. Europeana fully supports open access. Within the project, the NBS submitted 8 new collections and updated metadata for 15 collections, which had been incorporated into the Europeana platform through earlier projects. The aggregation

22. Europeana, „Europeana Publishing Framework: Content“, accessed 16. 3. 2022
23. Tier 1, Tier 2, Tier 3, Tier 4.
24. Europeana, „Europeana publishing guide“, accessed 16. 3. 2022
25. Europeana, „Europeana Licensing Framework“, accessed 16. 3. 2022
26. „Rights Statements“

infrastructure first had to be improved to enable updating the collections. Semantika, a Slovenian company, worked on the Aggregator for Europeana as a technical partner of the NBS. As part of the project, the Aggregator was updated in such a way as to be able to process geographic and temporal headings, metadata related to authors' and contributors' dates of birth and death, Wikidata links and scripts in which their names are written (this is important because the Serbian language has two scripts).

Figure 3 shows what the updated metadata on the author and contributor look like. The <dc:creator> and <dc:contributor> properties contain the years of birth and death, in addition to the names, and the xml:lang attribute contains the value sr, which stands for Serbian. The link toward the Wikidata database entry is entered as the value of the rdf:resource attribute.

```
<dc:contributor xml:lang="sr">Jovan Popović (1905-1952)</dc:contributor>
<dc:contributor xml:lang="sr">Јован Поповић (1905-1952)</dc:contributor>
<dc:contributor rdf:resource="http://www.wikidata.org/wiki/Q6296683" xml:lang="sr"/>
<dc:creator xml:lang="sr">Đorđe Andrejević-Kun (1904-1964)</dc:creator>
<dc:creator xml:lang="sr">Ђорђе Андрејевић-Кун (1904-1964)</dc:creator>
<dc:creator rdf:resource="http://www.wikidata.org/wiki/Q1283347" xml:lang="sr"/>
```

Figure 3. Metadata enrichment in a sample with a Serbian author and contributor

Figure 4 better illustrates the difference between the values of the xml:lang attribute. In this example, each author has three <dc:creator> properties in which their names are written in German (xml:lang="de"), Serbian Cyrillic and Serbian Latin. If there is no Wikidata entry for an author, that value is skipped (such is the case with Ludwig Erminy).

```
<dc:creator xml:lang="de">Ludwig Erminy</dc:creator>
<dc:creator xml:lang="sr">Ludvig Ermini</dc:creator>
<dc:creator xml:lang="sr">Лудвиг Ермини</dc:creator>
<dc:creator xml:lang="de">Franz Wolf (1795-1859)</dc:creator>
<dc:creator xml:lang="sr">Franc Volf (1795-1859)</dc:creator>
<dc:creator xml:lang="sr">Франц Волф (1795-1859)</dc:creator>
<dc:creator rdf:resource="http://www.wikidata.org/wiki/Q20772339" xml:lang="sr"/>
```

Figure 4. Metadata enrichment in a sample with foreign authors

Those two examples are parts of two EDM.xml records. After an EDM record is generated, it is uploaded to the Aggregator for Europeana where it is further transformed. Each field is transferred into its equivalent field(s). The end result looks like this (Figure 5):

```
▼<edm:Agent rdf:about="/direct/2354828">
    <skos:prefLabel xml:lang="sr">Jovan Popović</skos:prefLabel>
    <skos:note xml:lang="en">Latin script</skos:note>
    <rdaGr2:dateOfBirth>1905</rdaGr2:dateOfBirth>
    <rdaGr2:dateOfDeath>1952</rdaGr2:dateOfDeath>
  </edm:Agent>
▼<skos:Concept rdf:about="/direct/2354826">
    <skos:prefLabel xml:lang="sr">Ђорђе Андрејевић-Кун</skos:prefLabel>
  </skos:Concept>
▼<edm:Agent rdf:about="/direct/2354829">
    <skos:prefLabel xml:lang="sr">Ђорђе Андрејевић-Кун</skos:prefLabel>
    <skos:note xml:lang="en">Cyrillic script</skos:note>
    <rdaGr2:dateOfBirth>1904</rdaGr2:dateOfBirth>
    <rdaGr2:dateOfDeath>1964</rdaGr2:dateOfDeath>
  </edm:Agent>
▼<edm:Agent rdf:about="/direct/2354832">
    <skos:prefLabel xml:lang="sr">Јован Поповић</skos:prefLabel>
    <skos:note xml:lang="en">Cyrillic script</skos:note>
    <rdaGr2:dateOfBirth>1905</rdaGr2:dateOfBirth>
    <rdaGr2:dateOfDeath>1952</rdaGr2:dateOfDeath>
  </edm:Agent>
▼<edm:Agent rdf:about="/direct/2354827">
    <skos:prefLabel xml:lang="sr">Đorđe Andrejević-Kun</skos:prefLabel>
    <skos:note xml:lang="en">Latin script</skos:note>
    <rdaGr2:dateOfBirth>1904</rdaGr2:dateOfBirth>
    <rdaGr2:dateOfDeath>1964</rdaGr2:dateOfDeath>
  </edm:Agent>
▼<edm:Agent rdf:about="/direct/2354830">
    <skos:prefLabel xml:lang="sr">http://www.wikidata.org/wiki/Q1283347</skos:prefLabel>
  </edm:Agent>
▼<skos:Concept rdf:about="/direct/2354825">
    <skos:prefLabel xml:lang="sr">књига</skos:prefLabel>
  </skos:Concept>
▼<edm:Agent rdf:about="/direct/2354824">
    <skos:prefLabel xml:lang="sr">http://www.wikidata.org/wiki/Q6296683</skos:prefLabel>
  </edm:Agent>
```

Figure 5. Part of an EDM record after the transformation carried out by the Aggregator (Serbian author and contributor)

The years of birth and death are placed in <rdsGr2:dateOfBirth> and <rdsGr2:dateOfDeath> in the <edm:Agent> class (representing the authors). If the years are not available, those properties are left out. The Wikidata link is placed within <skos:prefLabel>, while <skos:note> contains information on the script (Cyrillic or Latin). The programme is capable of detecting the script automatically.

The link to the record after being processed by the Aggregator is submitted to Europeana. The end result[27] is the following view of the metadata (Figure 6):

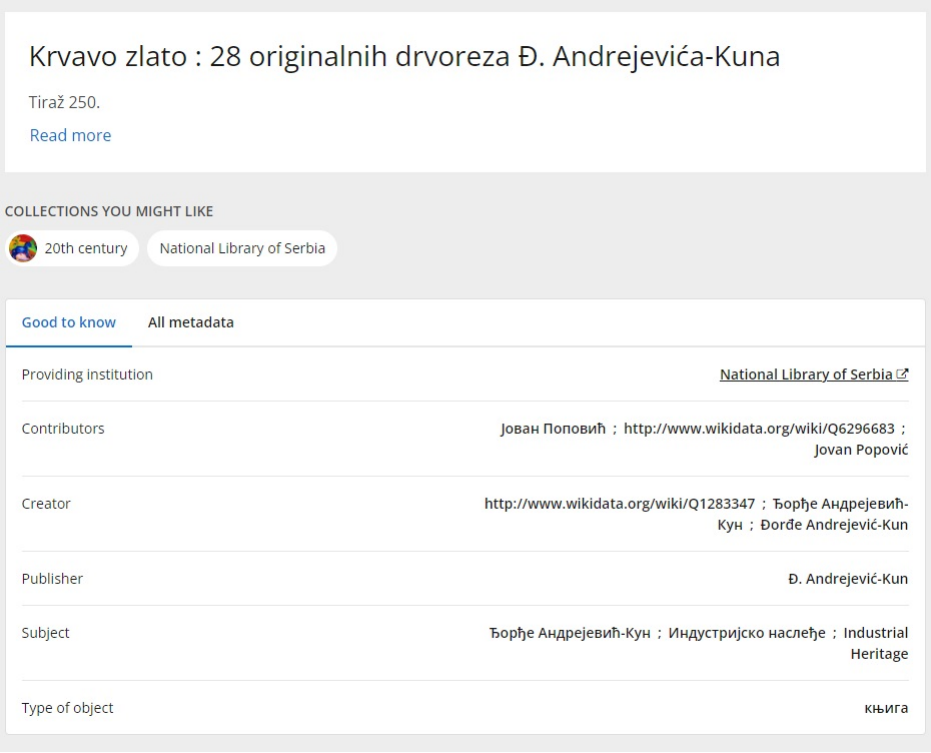

Figure 6. Metadata for a single object on Europeana

27. „Krvavo zlato: 28 originalnih drvoreza Đ. Andrejevića-Kuna“

Geographic and temporal headings are also new. EDM records now contain <dcterms:spatial> and <dcterms:temporal>, which are transformed by the Aggregator into <edm:Place> and <edm:TimeSpan>.

A sample of a geographic and a temporal heading in an EDM record is given in Figure 7.

```
<dcterms:issued>1921</dcterms:issued>
<dcterms:spatial xml:lang="sr">Југославија</dcterms:spatial>
<dcterms:temporal xml:lang="sr">1921</dcterms:temporal>
<edm:currentLocation rdf:resource="http://sws.geonames.org/792680/" />
```

Figure 7. A geographic and a temporal heading in an EDM record

After being processed by the Aggregator, the record looks as shown in Figure 8.

```
▼<edm:Place rdf:about="/direct/2352735">
   <skos:prefLabel xml:lang="sr">Југославија</skos:prefLabel>
 </edm:Place>
▼<edm:TimeSpan rdf:about="/direct/2352732">
   <skos:prefLabel xml:lang="sr">1921</skos:prefLabel>
 </edm:TimeSpan>
```

Figure 8. Record shown in Figure 7 after being processed by the Aggregator

The most recent aggregation infrastructure update at the NBS consists in enabling the processing of IIIF[28] records. This is an open standard for the online delivery of high-quality digital objects. The abbreviation also refers to the international open community for the development and implementation of this technology with the support of consortia made up of the leading cultural heritage and academic institutions. The initiative to perform this update originated from the Milutin Bojić Library in Belgrade. After the update, the Aggregator will be able to process the datasets coming from the libraries that have implemented the IIIF standard. The EDM specification for IIIF entails 4 steps:

28. "IIIF"

1. Submitting a IIIF Resource as <edm:WebResource>, which practically means listing a WebResource identifier, the purpose of which is to "review" the object and identify the connection type between a IIIF Resource and the object itself;
2. Indicating that the <edm:WebResource> is IIIF compliant, which means to link the WebResource to a source of the <svcs: Service> type, as well as to state that the WebService corresponds to the IIIF profile;
3. Indicating the level of IIIF implementation via <doap:implements>;
4. Enabling access to the IIIF manifesto directly or indirectly via <dcterms:isReferencedBy>.

A segment of a sample EDM record pertaining to a digital object from the Milutin Bojić Library (*Beležnica* Milutina Bojića,[29] a manuscript) illustrating what the Aggregator is required to submit for publication on Europeana is presented in Figure 9.

```
<!--
    edm:WebResource class - contains all information about the
    digital version of the CHO only-->
<edm:WebResource
rdf:about="https://milutinbojic.digitalna.rs/servlet/ImageTileRenderer?doc_id=0fbe53e1-56af-416f-9d98-79f165e07d4c/00000001/ruko
pisi/00000021&amp;pg_seq=1&amp;search_doc=&amp;scale=0.2&amp;rotation=0">
    <dc:format>jpg</dc:format>
    <dcterms:isReferencedBy
rdf:resource="https://milutinbojic.digitalna.rs/iiif/api/presentation/2/0fbe53e1-56af-416f-9d98-79f165e07d4c%252F00000001%252Fru
kopisi%252F00000021/manifest"/>
    <svcs:has_service
rdf:resource="https://milutinbojic.digitalna.rs/iiif/api/image/2/0fbe53e1-56af-416f-9d98-79f165e07d4c%252F00000001%252Frukopisi%
252F00000021%252F0"/>
</edm:WebResource>
<!--
    svcs:Service class - contains all information about the
    IIIF resource-->
<svcs:Service
rdf:about="https://milutinbojic.digitalna.rs/iiif/api/image/2/0fbe53e1-56af-416f-9d98-79f165e07d4c%252F00000001%252Frukopisi%252
F00000021%252F0">
    <dcterms:conformsTo rdf:resource="http://iiif.io/api/image"/>
    <doap:implements rdf:resource="http://iiif.io/api/image/2/level2.json"/>
</svcs:Service>
```

Figure 9. EDM segment for IIIF

5 Conclusion

Thanks to the cooperation with Europeana and the work on developing the Aggregator, the National Library of Serbia has adopted certain rules

29. Milutin Bojic Library – Digital Repository, "Beležnica Milutina Bojića"

and recommendations set forth by Europeana (EDM, minimum dimensions of scanned objects, OCR for PDF documents, copyright licences). This way, the library has implemented European standards in the area of cultural heritage. More importantly, it has proven that this can be done even with a small expert team. The following steps include the publication of new content on Europeana and continuous work on further improvement of content and metadata quality and the Aggregator. Further plans will evolve in two directions: including new data-submission partners to present Serbian culture within the European framework and implementing the rules and standards that Europeana adopts in the future.

References

Butigan-Vučaj, Tamara. 2019. "Nacionalni agregator za Evropeanu u Srbiji: mali korak bliže Evropskoj uniji." *Bibliozona* 7:15–19.

Drege, Evelin. 2019. "Budućnost prošlosti: Kulturno nasleđe u digitalnom svetu." Chap. Od ESE do EDM i dalje: na koji način Evropeana omogućava pristup objektima kulutrnog nasleđa, edited by Paul Klimpel i Elen Ojler, 87–111. Beograd: Narodna biblioteka Srbije.

Kasadeval, Arijadna Matas, and Tamara Butigan Vučaj. 2020. "Evropeana kroz objektiv autorskih prava." *Glasnik NBS* 22:41–56.

Klimpel, Paul, and Elen Ojler. 2019. *Budućnost prošlosti: kulturno nasleđe u digitalnom svetu.* Beograd: Narodna biblioteka Srbije.

Multimedia document „Welcome to my favorite corner“

UDC 004.55:378.147]:02(497.11)
DOI

Valentina Ivanović
valentinaivanovic2@gmail.com

Milica Čeketić
sarangfree@gmail.com

Dušica Galić
dusicagalic7@gmail.com

Jovana Stojičić
jovanastojicic98@gmail.com

University of Belgrade
Faculty of Philology
Belgrade, Serbia

ABSTRACT: The subject of this paper is the presentation of the project development process within the course Multimedia Documents, which is attended in the final year of undergraduate studies at the Department of Library and Information science, Faculty of Philology in Belgrade. The project is the result of the work of students of the 2017/2018 generation under the supervision of professor Branislava Šandrih Todorović and teaching assistant Milica Ikonić Nešić, with the help of prof dr. Cvetana Krstev. *"Welcome to my favorite corner"* was the topic of the project, which presented sites that were familiar or had sentimental value to students as a web page.

KEYWORDS: informatics, multimedia documents, HTML, CSS, librarianship

PAPER SUBMITTED: 16 November 2022
PAPER ACCEPTED: 01 April 2022

1 Introduction

Multimedia, as a combination of text, image and sound, is an efficient way of using modern technology in order to preserve the achievements of society. Multimedia is, in fact, information presented or stored in the form of a combination of text, graphics, sound, animation and video that are combined using computer programs. It allows the user to navigate and get acquainted with the certain topic, motivates the user and keeps his attention, but also helps to clarify doubts (Trtovac 2010).

As a gathering place of knowledge, the library must be equipped for handling all types of information that are produced and go hand in hand with

all forms of technology development and user needs. Therefore, literacy by its definition („The ability to read and write"[1]) is not enough for a librarian, it is never uniformed and it must always reflect the products of change that the new age brings, and one of those products is multimedia.

Accordingly, the program of the Department of Library and Information science within four years includes many IT subjects, such as Informatics for Librarians, Informatics Practicum, Digital Text, Databases, Information Structures, which aim to provide IT training for future work in IT (but also in information) field. The Multimedia Documents course is the ultimate test of knowledge acquired in the aforementioned courses, through the development of a joint project. The importance of this project is reflected in the fact that, in addition to testing knowledge, students develop both team and research spirit, which are important for any kind of training.

The final year students of the Department of Library and Information science of the Faculty of Philology in Belgrade - generation 2017/2018 worked on this project and here they brought together all the interesting sites of their hometowns that they would visit or show to people who would come to visit their city. The theme of this year's project is "Welcome to my favorite corner" ("Dobrodošli u moj najdraži kutak"). The mentors of the project were professor dr. Branislava Šandrih Todorović and Milica Ikonić Nešić, teaching assistant. The document shows the following cities and towns: Belgrade, Zemun, Užice, Smederevo, Šabac, Braničevo, Golubac, Obrenovac and Grocka. Students who participated in the project were: Valentina Ivanović, Milica Čeketić, Danica Kričković, Jovana Stojičić, Dušica Galić, Sanja Ignjatović, Maja Savić, Marija Đinić and Dragiša Stokić.

2 Inspiration for the project

"Welcome to my favorite corner" was not initially the topic intended for the final year students of the 2017/2018 generation. The original topic for students was "Famous personalities from my region", where each student would choose up to 10 well known individuals from their hometown to present interesting information about the particular person's life. However, soon there were doubts about the scope of the topic (is it just a person who was born there and became famous there, or may it also be applied to a person who became famous there, even though s/he was not born there, or a person who lived there after becoming famous, but was not born there ...).

1. „Literacy" Cambridge Dictionary (посећено 25. 3. 2022).

For that reason professor Branislava Šandrih Todorović suggested a more specific topic, "Welcome to my favorite corner", which was well accepted by students. The basic idea for this topic is simple: choose sites that we'd gladly show to the guest who came to visit us for the first time. Without any limitations, students could have picked any site, no matter for its historical or cultural context such as museums, monuments, etc. They were given freedom to present sites that were integral part of their daily routine, making it more beautiful or having some other kind of sentimental value for them (Kalenić market that the student used to visit as a child with her grandmother, a field called Bare near the village of Braničevo where young people gather in the summer . . .).

Also, throughout the project students got the opportunity to find out more about the sites they chose, something they may not have known before (Pavilion of the Great Refractor of the Astronomical Observatory was acquired by Milan Nedeljković for war reparations after World War II (Mihajlov 2010), but also, thanks to their colleagues, to get acquainted with sites they have not heard of before or knew little about. It is important to emphasize that, due to pandemic situation, it was decided to avoid restaurants or similar catering facilities.

3 Classification of sites

Before initial creation of the website, each student had to choose 5 to 10 sites presented in PowerPoint, which included basic information about the chosen sites, with explanations about what makes each particular site special for the student.

Professor Šandrih Todorović set up a table within Google Drive where students entered their suggestions, which were subsequently classified according to the criteria agreed with the professor. The classification included 8 categories and 21 subcategories:

1. CULTURAL CENTERS- libraries, museums, theaters/ cinemas, scientific institutions, galleries
2. SITES FOR RECREATION - sports centers, playgrounds
3. MONUMENTS - fortresses, towers
4. SQUARES
5. RELIGIOUS OBJECTS - churches, monasteries
6. WATER OASIS - beaches, quays, rivers, lakes
7. CRAFTS AND TRADE - craft shops, markets

8. GREEN OASES - parks, promenades, gardens, lookouts

After the classification, the material for presentations and website was collected.

4 Content collection

To collect material for this project for a period of four months, each student went to the particular site s/he chose for the project. Students made photographs for each site they visited, while the texts that describe places were taken from numerous sources, which included books (such as "Colorful World of Narration" by Jovan Đorđević or Darinka Lekić's book "Belgrade forever"), periodicals (articles from daily newspapers, such as Politika), oral sources, webpages, etc. Collecting texts included going to local libraries, researching home libraries, information on the Internet and official webpages if any. Oral sources included local stories and legends related to the specific site where the students grew up.

While collecting the content, the students encountered obstacles such as "under construction", renovations or inconvenient working hours of the site, which in some cases, prevented them from taking the necessary photographs. If such obstacles could not be overcome, students would find a new site to be included in the project.

5 Developing and designing the webpage

For the main design of the webpage the team was inspired by the honeycomb. The idea behind this choice was that even though we come from different parts of the country we all belong to the same bee hive, which is the Library and Information science department. The whole page was divided into three levels for practical reasons and so that the tasks could be delegated more efficiently. The first level is the Homepage, which is the starting point for further page navigation where general categories are placed. The second level contains the names of sites that are connected to general categories from the first level. The third level is represented by individual student pages that contain sites which represent their favorite corners. During the initial stage of the website creation, students used technologies and knowledge they acquired during the second year, as part of the Digital text 2

course. HTML [2] is the standard markup language for creating webpages,[3] while CSS [4] is used to describe the way HTML elements are displayed on a website [5]. The choice to use HTML and CSS proved to be very practical given the students are well acquainted with these technologies and since they are easy to use. Also, HTML in combination with CSS provides many creative solutions for projects like this one for Multimedia Documents since the content of the webpage contains a lot of images, animations, music and other forms of media.

5.1 Development of the first level of the webpage

In early stages of development, the team chose a honeycomb layout for the first level combined with honey and chocolate hues that give off a pleasant appearance. Team member Valentina Ivanović was tasked with the development of the Homepage and template for other two levels as well. For the "Welcome to my corner" page header (Figure 1) the team used "Sacramento" font that was downloaded from the Google Fonts platform, which provides free access to numerous fonts, while CSS was used to decorate it using a neon sign animation.

An animation was applied to each honeycomb field so when the cursor hoovers over it the name of the general category disappears and an image that represents the category is revealed. The author of the photographs was Valentina Ivanović. Honeycomb fields are also buttons that are linked to the second level of the webpage (Figure 2).

The Homepage also has an animated button that leads to the "About" page that is placed on the second level. Team members also wanted to add music to the Homepage which was done using HTML and the song "Night" by the Italian pianist Ludovico Einaudi was chosen.

A subtle orange gradient animation was also placed on the Homepage as a background. This effect was accomplished by using a CSS attribute, animation, within the body tag.

2. Acronym for *Hyper Text Markup Language*
3. "HTML introduction" (visited 8.11.2021).
4. Cascading Style Sheets
5. "CSS Introduction" (visited 8.11.2021).

Figure 1. Homepage

Figure 2. Honeycomb animation, before and after

5.2 Development of the second and third level of the webpage

In the early development stages of the project, team members decided to use the existing template created for the Homepage, to keep a uniform look, while the third level team members had full artistic freedom.

For the design of the second level, students also had a bit of creative freedom, so they picked colors they would use to decorate the template. Every page of the second level has a header of the general category which it represented, honeycomb fields that are buttons with links connected to certain sites on the third level. To ensure that the user will land on the right part of the page on the third level, students used anchors. The second level pages also contain three animated buttons that lead to the previous general category, Homepage and to the next general category (Figure 3).

Figure 3. The second level

The "About" page is also located on the second level that contains information about the project, team members, their photographs (that are linked to their personal pages on the third level) and also photographs of project mentors prof. dr. Branislava Šandrih Todorović and teaching assistant Milica Ikonić Nešić (Figure 4).Their photographs are also linked to their personal websites.

For the third level, which contains individual team member profile pages, some students used the provided template, while others used their own cre-

Figure 4. The about page

ative solutions. Most of the team members used HTML and CSS, except for Dušica Galić, the team member who used a little bit of JavaScript as well. Each individual page from the third level contains two buttons that are linked to the Homepage and a page which contains information about the author of the page and references they used. On their individual pages, team members presented their favorite corners, usually from their hometowns, which made this project diverse and provided the opportunity for other team members to learn more about the country we live in (Figures 5 and 6).This way, the project joined other similar projects created by previous generations of students for the Multimedia document course, like "Al' se nekad dobro jelo" that the students of 2012/2013 generation developed about traditional Serbian cuisine, or the "Igra rokenrol cela Jugoslavija" project that the students of 2014/2015 generation developed for the purpose of "presenting the most comprehensive material about the beginning and development of the rock scene in Yugoslavia in one place".[6]

6 Conclusion

Thanks to the Multimedia Documents course students had the opportunity in their final year, to merge the knowledge acquired during the years

6. „About the project", The whole Yugoslavia is dancing rock'n'roll (visited 25.3.2022).

Figure 5. A personal page from the third level

Figure 6. A personal page from the third level

as undergraduates. Lessons that may have seemed irrelevant and abstract during the study, information that may have seemed vague, were shaped and fully adopted within the work on this project.

Through personal efforts, harmony and diligent work of all students, as well as the help of professors, the final project - "Welcome to my favourite corner" is a valuable experience gained during the summer semester of the year 2020/21. Being motivated and eager to present different parts of our country and sites that have exceptional value for them, students easily and quickly agreed on all parts of the project. There were almost no disagreements, so the project itself was completed on time without major difficulties.

Even though the sentimental value was the major criteria for selecting sites for students, many of the presented sites are actually of historical and cultural significance to the entire community. Having these facts in mind, the project can also be seen as a tourist guide that can lead viewers of the site to interesting places they could visit themselves.

We hope that our project will inspire you to visit some of our favourite places and that it will become familiar to you too.[7]

Acknowledgement

We owe a great deal of gratitude to professor Dr. Cvetana Krstev, professor Dr. Branislava Šandrih Todorović and teaching assistant Milica Ikonić Nešić, for patient and professional guidance, willingness to always take time for our questions and for all the knowledge transferred, not only during the project, but also during the previous four years of our academic education.

References

Mihajlov, Saša. 2010. *Astronomska opservatorija.* Posećeno 25.3. 2022. Beograd: Zavod za zaštitu spomenika kulture grada Beograda. http://beogradskonasledje.rs/wp-content/uploads/2012/06/astronomska_opservatorija.pdf.

Trtovac, Aleksandra. 2010. "Multimedijalni dokument kao predmet-projekat na Katedri za bibliotekarstvo i informatiku Filološkog fakulteta u Beogradu." *Infoteka* 11 (2): 45–54. http://infoteka.bg.ac.rs/pdf/Srp/2010-2/INFOTHECA_XI_2_December2010_45-54.pdf.

7. "About the project", Welcome to my favourite corner (visited 11.11.2021).

"Digital transformation and Libraries in Special Circumstances" – conference report

Oja Krinulović
okrinulovic@unilib.rs

Vera Petrović
veramilanovic
petrovic@gmail.com

University library "Svetozar Marković" Belgrade, Serbia

PAPER SUBMITTED: 07 March 2022
PAPER ACCEPTED: 16 May 2022

About the conference

The days of special and higher education libraries have been held for many years now. The date of this conference is at the end of April. Initially, the venue was Opatija, a coastal little town in the Republic of Croatia, however, since 2015, the venue has been moved to Lovran, not far from Opatija. This conference is a meeting of librarians from special and higher education libraries. The organizers are: the Croatian Library Association (HKD), the Section for Special and Higher Education Libraries, the National University Library (NSK) from Zagreb, as well as the Library Association and the University Library from Rijeka. The goal is to present higher education libraries that are part of various institutes as well as other institutions whose activities are mainly related to scientific research projects with the Ministry of Science, Education and Sports.

Previous conferences covered topics related to the concept and organization of special libraries, as sources and places of using knowledge. In addition, attention is paid to the education of users in education, as well as continuous training of library staff. The library system of Croatian libraries was also analyzed, as was the partnership between libraries in Croatia and those in the region.

Guests from abroad, colleagues from university libraries, representatives of publishers and experts from related professions were all invited to the conference. The program included lectures by invitation, presentations of posters on a given topic, workshops and a presentation of sponsors.

Topics from some of the previous conferences

Extremely interesting topics related to higher education and special libraries were discussed at the previous conferences. Some of them are:

- Libraries, founders, community: dialogue for change
- Connecting and collaborating: opportunities and challenges
- Libraries: where and what is the path forward?
- Establishing quality in special and higher education libraries- libraries in the Bologna area
- A step towards the new Voyager library system
- New strategies in special and higher education libraries, promotion and fundraising
- Partnerships in special and higher education libraries

17 Days of special and higher education libraries

Due to the global pandemic caused by the Covid 19 coronavirus, the 2021 conference was postponed until the end of September, so in 2021 the 17 Days of Special and Higher Education Libraries were held, but in a slightly changed format.

The seventeen days of special and higher education libraries were held in the period from 30th September to 2nd October 2021 in Lovran, in the form a hybrid conference, meaning that the participants were able to attend the conference live in person and virtually. The theme of the conference was "Digital Transformation and Libraries in Special Circumstances". The presentations were within the following subtopics:

1. Copyright and related rights – new circumstances and digital environment
2. Competence of librarians the work in a digital environment
3. The role and position of libraries in the digital environment
4. Space, services and materials of libraries in the digital environment

The presentations were in the form of invited presentations, round tables, workshops and poster galleries.

The conference was opened by a round table entitled "The role of librarians in supporting scientists with regard to changes in the ways of scientific communication." Two exhibitors were invited to participate in the round table: Professor Dr. Ivana Kunda, Vice Dean for Science at the University of

Rijeka, also a professor at the Faculty of Law, with a paper on "Hybrid Law Librarian or Librarian Lawyer: however you please in the digital environment" and Professor Dr. Miroslav Rajter, Director of the Research Office at the University of Zagreb with the topic: "How rankings rank us and how we rank the rankings". Two invited exhibitors were present at the opening of the conference as well: Professor Dr. Tatjana Aparac Jelušić, President of the Coratian Library Council with her work on the topic "Normative framework in the field of higher education libraries" as well as Loida Gracija-Febo, President of the American Library Association 2018-2019 with her topic on "Libraries and digital transformation: supersonic library services in the 2020s".

The conference was concluded by invited speaker Professor Dr. Miroslav Rajter with a peper on "Trends in the development of Croatian universities from a science perspective".

All of the works were extremely interesting and useful, however one work needs to be especially emphasized, and that is the work of Renata Petrušić from the National and University Library in Zagreb on the topic "Works which are not available on the market: copyright, digitization, use". The paper presents the changes brought on by the new Law on Copyright and Related Rights, which relate to works that are not available on the market (out-of-commerce works). The new law enables institutions dealing with cultural heritage to digitize and use works that are not available on the market. The interesting thing is that this law was adopted at the time of the Conference. Considering that in Serbia, out-of-commerce works can only be photocopied, not digitized, perhaps we should consider the possibility of introducing such a law in our country, which would enable digitization of works that are not available on the market.

At the conference, 40 papers were presented as well as 14 poster presentations. The conference was attended by 140 colleagues, 50 of which were physically present and 90 listened to the lectures online. We were especially pleased to learn that during the Conference, the Ministry of Culture and Media of the Republic of Croatia adopted the Standards for Digital Libraries.

Review of the Belgrade European Language Grid Workshop

Bojana Bašaragin
bojana.basaragin@ivi.ac.rs
Research and Development Institute for Artificial Intelligence of Serbia
Belgrade, Serbia

PAPER SUBMITTED: 19 April 2022
PAPER ACCEPTED: 16 May 2022

March 11, 2022 was marked by the first in a series of dissemination events within the EU project European Language Grid (ELG)[1]. This was an online ELG Workshop organized by the Faculty of Philology in Belgrade and the Society for Language Resources and Technology JeRTeH. This event announced the recent and successful completion of the project in mid-2022.

The ELG project (2019-2022) was created with the idea of joining the academic community and industry in the field of language technologies and strengthening Europe's position in this regard compared to other continents. The result of the project is a unique ELG platform that collects resources and tools for processing European languages, aiming to make them more visible and accessible to academia, business partners, NGOs, and the public sector. One of the project goals is to help address the issue of digital language vulnerability by providing help to those European languages that are not sufficiently supported through language technologies. According to data from January 2022, the platform gathers over 10,000 services, tools, data sets, resources, and language models for 87 languages, and that number is continually increasing. Following the completion of the project, the plan is for the platform to continue its life, providing access to both non-commercial and commercial language technologies.

The workshop lasted three hours and was held online, via the Zoom platform, while the recording of the workshop can be found on the JeRTeH Society YouTube channel[2]. The official language of the workshop was English. After the welcoming speech of the organizer, prof. Cvetana Krstev, the project coordinator prof. Georg Rehm gave the introductory presentation. Prof. Rehm made a detailed review of the ELG project and its results, as well as perspectives for further growth and development of the platform. He

1. European Language Grid project (ELG).
2. JeRTeH YouTube channel.

also presented the problem of language inequality in terms of language technologies for most European languages, except Spanish, German, and French (Figure 1). He addressed the situation of the Serbian language, which, according to the European Language Equality (ELE) project report for Serbian[3], still belongs to the languages with weak language technology support. When asked what he sees as the next step in the development of resources and tools for the Serbian language, he mentioned the development of a comprehensive language model, which could then be applied to various tasks of language processing and generation.

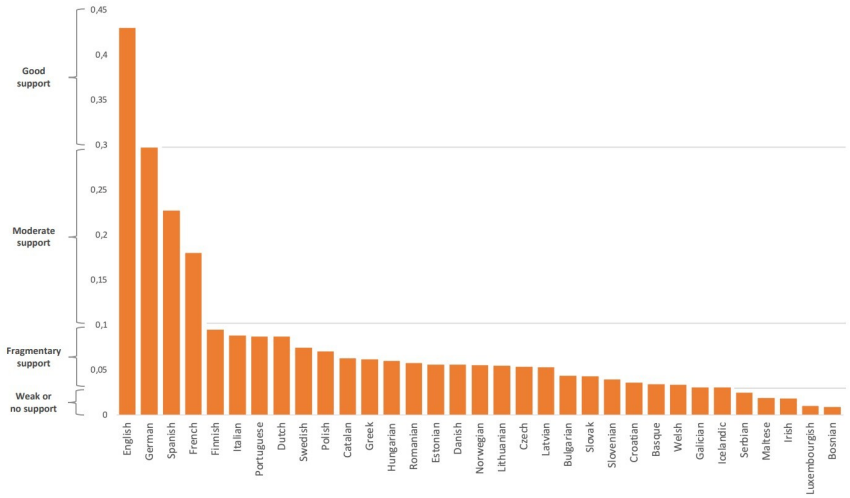

Figure 1. Language technology support for different European languages

The introductory part was followed by a practical and detailed demonstration of the platform use given by Penny Labropoulou, Senior Research Associate at the Institute for Language and Speech Processing of Greece. The presence of foreign lecturers allowed using the results of ELITR[4], a recently completed European project, for the needs of the workshop. ELITR participants developed a system for automatic translation and subtitling of

3. ELE report for Serbian.
4. European Live Translator project (ELITR).

meetings and conference presentations, as well as note-taking. Automatic captions of all workshop presentations and their translation into Serbian (or some other European) language could be followed in a separate window in real time.

After the introductory presentations came the examples from the domain of language technologies for the Serbian language. In the spirit of the project, the workshop was designed to bring together examples of successful applications of language technologies in the industry and the latest achievements in the scientific field. Thus, the first two presentations came from two large Serbian companies – the telecommunication service provider Yettel and the IT group ComTrade. Đorđe Hirš, Data Science Team Lead at Yettel Serbia, presented the method of unsupervised sentiment analysis that Yettel applies to the database of customer service voice complaints to improve their services. Dušan Jovičić, a programmer, presented the regional experience of the company ComTrade in developing conversational agents (chatbots).

Examples from the industry were followed by the presentation of resources and tools for the Serbian language developed in cooperation between the University of Belgrade (BU) and the JeRTeH Society. First, prof. Ranka Stanković spoke about the corpus of the modern Serbian language SrpKor, which at the end of 2021 got a new version – SrpKor2021. Compared to the previous version, this corpus is enriched with new texts from various domains, featuring ELTeC, a corpus of literary texts from the 19th and early 20th century, created as part of the COST action Distant Reading[5]. This corpus, composed of over 100 POS tagged literary works was the topic of the final workshop presentation. The ELTeC corpus was presented by prof. Cvetana Krstev and Milica Ikonić Nešić as a completely free resource that can be used for analyses and training models for the Serbian language. This corpus can currently be downloaded from the ELTeC GitHub. SrpKor2021, as well as the ELTeC corpus, can be searched through the NoSketch system, available to registered users through the JeRTeh website[6].

Biljana Rujević gave a practical demonstration of the *Leximirka* platform, developed for the unified supplementation, maintenance, and use of electronic dictionaries, corpora, and other lexical resources and tools created in cooperation with BU and JeRTech. This platform can also be accessed from the JeRTech website after registration. Mihailo Škorić presented BEaST, a newly developed POS tagger for the Serbian language that shows improved performance compared to the taggers it is based on: TreeTagger,

5. Distant Reading for European Literary History COST action.
6. JeRTeh website.

spaCy, and Stanza. This tagger, used to annotate the ELTeC text corpus, is available for download via GitHub. Branislava Šandrih Todorović spoke about the web portal *NER & Beyond*, developed within the COST action Distant Reading. The portal was designed to allow named entity recognition and tagging using various available tools but also mapping between the annotations produced by these tools. This platform is available through the JeRTeh website.

The ELG workshop was a real refreshment. On the one hand, it provided an opportunity to learn about the ELG project, as an important step towards unification and greater visibility of tools and resources for the Serbian language. On the other hand, the participants gained insight into some of the latest research results in the field of language technologies for the Serbian language and their applications. The dynamic structure within which the presentations of services, resources, and tools lasted 15 minutes each made it possible to get an overview of a large number of projects in a short time, with enough information about each of them. I look forward to the next opportunity to attend such a comprehensive and well-organized event, with a belief that every such effort contributes to improving the situation of language technologies for the Serbian language.

Distant Reading Training School – *Exploring ELTeC: Use-Cases for Information Extraction and Analysis*

Aleksandra Marković
aleksandra.markovic@isj.sanu.ac.rs
The Institute for Serbian Language SASA
Belgrade, Serbia

PAPER SUBMITTED: 28 May 2022
PAPER ACCEPTED: 8 June 2022

Training school *Exploring ELTeC: Use-cases for Information Extraction and Analysis* took place in Belgrade, from 22th till 24th of March 2022. This workshop, dedicated to ELTeC (abb. for European Literary Text Collection) was the final one within the COST Action 16204, under the title *Distant Reading for European Literary History*.

The training school was organized at the Faculty of Mining, University of Belgrade, and organizers were Prof. Dr Ranka Stanković (from the same faculty), Prof. Dr Cvetana Krstev (Faculty of Philology, University of Belgrade),[1] and Joanna Byszuk (Institute of Polish Language, Polish Academy of Sciences). The workshop was organized in a hybrid form, which means that attendees could choose whether they will take part in person or online. Among the attendees who participated in person were ones from Serbia, Slovenia, Romania, Belgium and Lithuania, and participants who were present remotely were from Austria, Britain, Portugal, Hungary. There were no formal knowledge requirements for the participation, but it was advised for the attendees to have at least basic computer skills. The target audience for this training school were researchers from the countries which took part in the Distant Reading project, interested in Digital Literary Studies, Corpus and Computational Linguistics, Literary Theory and their methodological uses across national traditions.

The training school offered hands-on approaches to information extraction and analysis of textual data, especially ELTeC corpora, developed within

1. Professors Krstev and Stanković were in charge for the production of the Serbian part of ELTeC. Those who are interested in the Serbian collection digitized within Distant Reading may find information about the work in the journal Infotheca, Vol 21, No 2, which is wholly dedicated to the Serbian novels collection. ELTeC can be reached at the Github of the Action

the aforementioned COST action. Among covered topics were different aspect of work with named and geographical entities: their recognition (NER), extraction (NEE), as well as their analysis; work with Wiki-ELTeC data, linking (historical) data with Nodegoat, the platform made for research in humanities),[2] semantic analysis with word embeddings and language models, and comparing corpora with stylometry.

Among the trainees there were many participants from the COST Action Distant Reading for European Literary History. Let's mention some of them: Christof Schöch (Professor of Digital Humanities at the University of Trier, Germany; the chair of this COST Action); Maciej Eder (the Director of the Institute of Polish Language, Polish Academy of Sciences); Diana Santos (Professor of Portuguese language, and Statistics for Humanities at the University of Oslo), Fotis Yannidis (Professor of Digital Humanities at the University of Würzburg in Germany); Dr Cvetana Krstev (Professor of Information Sciences at the Library and Information Science Department, Faculty of Philology, University of Belgrade, retired); Dr Ranka Stanković (Professor of Mathematics and Information sciences at the Chair of Applied Mathematics and Informatics, The Faculty of Mining and Geology, University of Belgrade).

The training school was organized in nine modules, and the attendees were expected to participate in all sessions.[3]

1. Christof Schöch held (online) an introductory lecture about the project and ELTeC (*What is ELTeC all about?*). This lecture was an introduction to the objectives of the COST Action *Distant Reading for European Literary History*, with a particular focus on the structure of the core deliverable of the project – the multilingual European Literary Text Collection (ELTeC).
2. Maciej Eder, Joanna Byszuk, and Artjoms Šeļa held a lecture (online) under the topic: *Exploring and comparing ELTeC corpora with stylometry*. The lecture was followed by the Stylo intro hands-on.
3. Diana Santos held an online session: *NER exploitation and analysis*; the lecture was followed by hands-on in R.
4. Benedikt Perak was talking about ELTeC Data Analysis, representation of the Geo-Entities and interlinking with knowledge bases; his lecture (held on site) was also followed by hands-on.

2. Nodegoat
3. The program of the training school.

5. Fotis Jannidis and Leonnard Konle held an online lecture: *Semantic analysis using word embeddings and language models*, followed by a practical part.
6. Ranka Stanković and Milica Ikonić Nešić held (on site) Wiki-ELTeC data session: *Wikidata introduction; Wiki-ELTeC schema* (all metadata from header plus main characters, their relations, places); *pipeline for Wiki-ELTeC data population*; *predefined SPARQL query exploration.* The talk was followed by hands-on: *population of Wikidata for other languages.*
7. Denis Maurel, Eric Laporte, and Cvetana Krstev held a lecture (online): *Unitex for processing of literary text: the case of NER automata. Enriching ELTEC texts by Named Entity Recognition using CasSys to parse texts with Unitex graph cascade of finite state transducers in different languages* (hands-on).
8. Jessie Labov, Pim van Bree, and Geert Kessels held an online lecture: *ELTec in Nodegoat: Introduction to the Nodegoat interface and how it works with this kind of data.*
9. Pim van Bree and Geert Kessels talked about *Using Nodegoat for working with the ELTeC data* (specifically the NER), demonstrating how to enrich it by linking it to open data sources (hands-on).

The workshop was intensive, informative and dynamic. All materials needed for work and hands-on were available at the GitHub.[4] The organization and lead through topics were great, topics were interesting and relevant, and practical work was conceived very well (even for those who lack higher level of computer skills, as in the case of the author of this review).

4. Workshop material

Author Guidelines

All *Infotheca* articles are published both in English and Serbian in the same issue. Authors should submit their articles in one of the languages; only after the notification of acceptance the translated article is expected (for Serbian authors; for all other authors translation from English to Serbian is provided by the journal). Except the printed edition, all articles are also published in the online edition in open access.

PAPER CATEGORIZATION

For documents accepted for publishing which are subject to review, the following categorization in the Journal applies:

1. Scientific papers:
   - Original scientific paper (containing previously unpublished results of authors' own research acquired using a scientific method);
   - Review paper (containing original, detailed and critical review of a research problem or a field in which authors' contribution can be demonstrated by self citation);
   - Preliminary communication (original scientific work in progress, shorter than a regular scientific paper);
   - Disquisition and reviews on a certain topic based on scientific argumentation.
2. Scientific articles presenting experiences useful for advancement of professional practice.
3. Informative articles can be:
   - Introductory notes and commentaries;
   - Book reviews, reviews of computer programs, data bases, standards etc.
   - Scientific event, jubilees.

Papers classified as scientific must receive at least two positive reviews. The opinions of the Editorial Committee do not have to correspond to those expressed in the published papers. Papers cannot be reprinted nor published under a similar title or in a changed form.

ELEMENTS OF MANUSCRIPTS

For scientific or professional papers the following data should be provided:

1. Papers should not normally exceed 15 A4 pages, Times New Roman 12pt. For longer articles the authors should contact the journal editors.
2. Names and surnames of all authors should be written in the sequence in which they will appear in a published paper.
3. After each author's full name, without titles and degrees, an e-mail address should be specified as well as the full and official name of his or her affiliation. (For large organizations full hierarchy of names should be specified, top down).
4. The submission date should be provided.
5. The authors should suggest the category of their paper but the Editor-in-Chief is responsible for the final categorization.
6. An informative abstract not normally EXCEEDING 200 WORDS that concisely outlines the substance of the paper, presents the goal of the work and applied methods and states its principal conclusion, should accompany the paper. The abstract should be supplied in both languages used for publication. In the abstract, authors should use the terms that, being standard, are often used for indexing and information retrieval.
7. Authors should supply at least 3 but not more than 10 keywords separated by commas that designate main concepts presented in the paper. The list of keywords should be supplied in both languages used for publication.
8. If paper derives from a Master's thesis or Doctoral dissertation authors should give the title of the thesis or dissertation, as well as a date of its submission and names of responsible institutions.
9. If the paper presents the results of authors' participation in some project or program, authors should acknowledge the institution that financed the project in a special section "Acknowledgment" at the end of the article, before the "Reference" section. The same section should contain acknowledgment to individuals who helped in the production of the paper.
10. If the paper was presented at a Conference but not published in its Proceedings, this should also be stated in a separate note.
11. Authors can use footnotes, while endnotes are prohibited; however, too long footnotes should be avoided. Authors can add appendices to their paper.
12. The referenced material should be listed in the section "References" at the end of the paper. In the reference list authors should include all information necessary for locating the referenced work. All items referenced

in the text should be listed here; nothing that was not referenced in the text should appear in this section.

EDITING CONVENTIONS FOR ACCEPTED PAPERS

1. Papers should be prepared and submitted using LaTeX(the journal style and all packages can be downloaded from the journal web site). Authors that are not familiar with LaTeXcan prepare their papers using Word, as .doc, .docx, .rtf or .txt documents. These authors should not use any special formatting – the final formatting and transformation to LaTeXwill be done by the Infotheca team.
2. The papers written in Serbian should use CYRILLIC alphabet because they will be printed in that script. The only exceptions are those parts of the text for which the use of the other script, such as Latin, is more appropriate. All scripts should be represented using Unicode encoding, UTF-8 representation.
3. Title of the paper should not be written in capital letters. The authors should keep the length of titles reasonable – preferably less than 90 characters. For all titles authors should provide a shorter title that will be used for page headers.
4. Italic type may be used to emphasize words in running text, while bold type or italic bold type can be used if necessary. Underlined text should be avoided. Please do not highlight whole sentences or paragraphs.
5. Paper can be divided in sections and subsections, but more than two levels of the section headings should be avoided. All sections and subsections will appropriately numbered. Appendices, if any, should come at the end of the paper and they will also be appropriately labeled. If using lists, do not use more than two levels of nesting.
6. All paragraphs should be separated by one empty line (one Enter).
7. Authors should avoid too wide tables keeping in mind that the journal is published on A5 paper and. All tables, illustrations, diagrams and photographs should not be wider than 72.5 mm (the width of one column) or (exceptionally) 150 mm (the width of the page). All illustrations should be prepared in some lossless format, for instance .png, .tif or .jpg and their resolution should be at least 300 dpi.
8. The authors are kindly requested to add (if possible) the link to the screen from which a screenshot was taken. When taking a screen shot of a part of some screen, authors are advised to use the Zoom possibility of the browser or other program. For diagrams that are produced with Excel, please provide the original .xls document.

9. All tables, illustrations, diagrams and photographs should be prepared as separate files, both in black-and-white for printing and in color for the on-line version. Captions that should be below tables, illustrations, diagrams or photographs should remain in the text. Each file should have the same name as the file containing the main text, followed by the type of material to which the ordinal number in the text is added. For instance, the file containing the fourth figure of the paper "Example" should be named Example_figure_4.
10. Please add additional document(s) that explain some specific aspects of formatting required for your paper, for instance, formulas prepared in LaTeXin a .pdf format.
11. URL addresses that appear in the paper should be placed in footnotes; the date when the site was visited should be given.

REFERENCES AND CITATION

1. Referenced material should be listed at the end of the text, within the unnumbered section References. The reference section should be complete; references should not be omitted. This section should not contain any bibliographic information not referenced in the main text. Referenced items should not be mentioned in footnotes.
2. Entries in the reference list should be ordered alphabetically by authors or editors names, or publishing organizations (when no authors are identified). If this list contains several entries by the same authors, these entries should be ordered chronologically.
3. For preparation of a reference list use Chicago Manual of Style reference list entry (www.chicagomanualofstyle.org).
4. Full names of journals, and not their short titles or acronyms, should be specified. Use the 10-point type for entries in the reference list.
5. All authors, whether they prepare their articles using LaTeXor Word, will prepare all the items from their References section using BibTeX templates that are given for all the examples at the Infotheca web site (http://infoteka.bg.ac.rs/index.php/sr/upu-s-v-z-u-r).