Info*theca* Journal of Digital Humanities

Impressum

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

Belgrade, Vol. 17, No. 1, Jun 2017

JOURNAL REDACTION:
Journal Infotheca
11000 Belgrade, Bulevar kralja Aleksandra 71

+381 11 3370–211
infotheca@unilib.rs

PRINTED BY:
Dunav
Belgrade

Journal is published twice a year

Table of Contents

Multi-dialectal Lexicon Building: Lessons Learned

UDC 811.14'06'374'282.2.
DOI 10.18485/infotheca.2017.17.1.1

ABSTRACT: In this paper, we discuss the lessons learned through the lifecycle of a dialectal electronic lexicon. Our approach is innovative because our lexicon is designed and built as a multi-dialectal (trilingual) dictionary (three dialects vs. one target language) instead of three monolingual dialectal dictionaries. Our system offers features that could not be possible with three monolingual dialectal dictionaries. Moreover, during the system's lifecycle we have got very specific demands for improvements (new requirements) that users were not able to express during the analysis phase. The lessons learned and the solutions invented for the system's ultimation (to respect the new requirements) can be helpful for other research or project with similar purposes.

KEYWORDS: Computational Dialectology, Dialectal Lexicography, Electronic Dictionaries, Lexical Resources, Modern Greek Dialects, Asia Minor Greek.

PAPER SUBMITTED: 08 February 2017
PAPER ACCEPTED: 28 April 2017

Nikitas N. Karanikolas
nnk@teiath.gr

Eleni Galiotou
egali@teiath.gr

Technological Educational Institute of Athens
Department of Informatics

1 Introduction

In a previous work (Karanikolas et. al., 2013), we have presented the design and implementation of a multimedia electronic dictionary of three Greek dialects in Asia Minor (Pontic, Cappadocian, Aivaliot). We had presented the linguistic and lexicographic approach adopted, as well as the principles for designing the macro/microstructure of the dictionary. We also had presented the conceptual model of the tri–dialectal dictionary and the equivalent relational schema. According to the above analysis a system has

been implemented that hosts lemmas and relevant lexicographic information from three Asia Minor Greek dialects.

However, during the lifecycle of the system, and because of the highly–qualified users, we have got very specific demands for improvements and we have caught the ultimate goal (an excellent system). In this paper, we report the lessons learned through this system's lifecycle and we present the improved design and the extended facilities of the 3–dialectal dictionary. We claim that our system can be used for other Greek dialects and that our extended design can be the base for multi–dialectal dictionaries for other languages. Our ultimate system can be used for multi–dialectal dictionaries of other languages so long as other virtual keyboards can be appended to it.

The paper is organized as follows. Section 2 presents the motivation for building the 3–dialectal lexicon and relevant work is presented in section 3. Sections 4 and 5 describe the initial requirements and the design according to the requirements set. Section 6 gives some details from the first implemented version of the system. Section 7 presents the demands for improvements and the relevant implementations. The result of the improvements is an excellent system and the design of this system is the topic of section 8 while conclusions are drawn in section 9.

2 Motivation

Pontic, Cappadocian and Aivaliot are three Greek dialects in Asia Minor which are not sufficiently documented and they are on the way to extinction. Until now, little interest has been shown in the dialects in question. The most interesting exception is the Papadopoulos' historical dictionary of Pontic (Papadopoulos, 1958). We can also find mentions to Cappadocian in some other works (Thomason, 2001; Thomason and Kaufman, 1988). There are also some glossaries for the Asia Minor Greek dialects containing words and idiomatic phrases accompanied by their meaning in Standard Modern Greek. However, in most of these glossaries, lemmas are stored in a very unsystematic way and crucial information, such as pronunciation or usages, is missing. Moreover, some verbs are listed in their past tense form while others appear in the present tense. Therefore, a sound linguistic analysis of Asia Minor Greek dialects is indispensable and gives insights as for the nature and mechanism of language change within the domain of dialectal variation.

This and other relevant social speculations (syllogisms) motivated us for the initiation of the AMiGre project, within the framework of THALIS pro-

gram. The project acronym (AMiGre) comes from the project's title: "Pontus, Cappadocia, Aivali: in search of Asia Minor Greek". One of the deliverables of the AMiGre project was the design and implementation of a multimedia tri–dialectal dictionary for three Greek dialects in Asia Minor (Pontic, Cappadocian, Aivaliot), which we discuss in this paper.

Dialectal dictionaries are usually treated as monolingual synchronic dictionaries. In our case (AMiGre), instead of creating three monolingual dialectal dictionaries, we have decided to treat and design a trilingual dictionary (three Asia Minor dialects vs. Standard Modern Greek). This is the most interesting technical motivation. It is also an interesting innovation because, it is permitting cross–reference links from lemma to lemma (of the same or different dialect) and equivalence links between meanings of lemmas from different dialects. These could not happen with three monolingual dialectal dictionaries.

3 Relevant Work

Electronic lexicography for Modern Greek was not concerned with the creation of dialectal dictionaries until very recently. The online dictionaries developed at the Portal for the Greek Language (Online, 2016) comprise the computerized versions of Georgacas' Greek–English Dictionary, Triandafyllides' Dictionary of Standard Modern Greek and Anastasiadi–Symeonidi's Reverse Dictionary. In addition, the Portal provides access to the computerised version of Kriaras' Concise Dictionary of Medieval Vulgar Greek Literature. The Institute for Language and Speech Processing has developed online bilingual dictionaries (Greek–English, Greek–German, Greek–Russian, Greek–Turkish, and Greek–Arabic). The dictionaries are under continuous development and enhancement and they are available from (ILSP, 2016). In addition, NLP tools for supporting lexicographic applications have been developed. Indicatively, in (Tsalidis et. al., 2010) infrastructure tools which are used for encoding morphological, syntactic and semantic information are reported as well as proofing tools such as a spelling checker, a hyphenator etc. As far as Greek dialects are concerned, the only computerized dictionary to our knowledge is the online lexical database of Cypriot Greek (Themistocleous, 2012). The online dictionary environment provides an enhanced searching mechanism as well as text to speech features for the pronunciation of Cypriot Greek words.

4 Initial Requirements

Dialectal dictionaries are usually treated as monolingual synchronic dictionaries (Béjoint, 2000; Geeraerts, 1989), due to limits in macrostructure (overall organizational scheme of lemmas) (Landau, 2001; Zgusta, 1971). Given that our purpose was to design and build an online dictionary, its macrostructure will not be restricted by physical constraints (limitations existing for print dictionaries), and could offer (virtually) "multiple macrostructures" mirroring the various searching options that we could build (Burke, 2003). Therefore, since there were no limits in macrostructure, we have decided to design and build a trilingual dictionary (Three Asia Minor dialects vs. Standard Modern Greek) (Xydopoulos and Ralli, 2012), instead of three monolingual dialectal dictionaries. The dictionary is named TDGDAM (Tri–Dialectal Greek Dictionary of Asia Minor) and it aims to be a linguistically–sound tri–dialectal dictionary in electronic form. One basic requirement of TDGDAM was that users should have access to a graphic (form based) representation of each lemma permitting them to handle pronunciation, meaning, usages and relations with other lemmas. The representation should be editable and for this to be possible, conventionally–adopted character sets should be used. Among other things, each lemma should contain the dialectal area and the source from which the lemma has been extracted. This type of dictionary constitutes an innovation not only for the Greek language and its dialects, but also for the international standards, as will be explained below.

Regarding its geographic and time scope, TDGDAM was designed to be a local/ microareal dialectal dictionary of non–synchronic nature that should include entries from different areas and time periods (Penhallurik, 2009). As it was decided from the beginning, the lemmas of TDGDAM should be drawn (directly or indirectly) from oral speech and written material of the particular dialectal varieties (Keymeulen, 2010).

Regarding TDGDAM's microstructure, our aim is to include formal information about pronunciation (phonetic form), grammar (categorial and morphological information), origin (etymology), meaning (synonymic and/or descriptive definitions), usage (thematic and register labels) and to provide linked multimedia resources (internal or external to TDGDAM) to enrich the semantics and pragmatics of lemmas (Barbato and Varvaro, 2004; Rys and Keymeulen, 2009; Xydopoulos and Ralli, 2012). To avoid different and arbitrary spelling codes for the same dialect (Durkin, 2010; Xydopoulos, 2012), headwords do not appear in a "semi–phonetic" transcription but in

(capitalized) orthographic form. In particular, the capitalized orthographic form departs from the spelling form in the standard dialect; it does not prescribe spelling rules in the dialect and allows for any alternative orthographic forms to appear in microstructure (Markus and Heuberger, 2007; Xydopoulos, 2012). Finally, authentic examples of use were considered as essential constituent information in entries which will appear in non–standard spelling, reflecting pronunciation as closely as possible with the use of diacritics, but avoiding a "semi–phonetic" transcription (Rys and Keymeulen, 2009).

Regarding the abilities for cross linking between items of the TDGDAM, we have defined 3 necessities: Cross–reference to other entries, related either through derivational processes or through semantic relations; Equivalence links between meanings of lemmas from different dialects; Synonymic/Antonymic relations.

The following 3 figures (figures 1, 2 and 3) present draft structural depictions of an equivalent number of lemmas that TDGDAM should contain. Based on these and other similar draft structural depictions we designed the TDGDAM system.

The terms synonymy (Synonym, 2016) and antonymy (Opposite, 2016) used previously and the terms homonymy (Homonym, 2016) and polysemy (Polysemy, 2016) that will be used later are very well defined. Their definitions are available on the internet.

5 Design

Based on the analysis presented in the requirements section and the draft structural depictions (see figures 1, 2 and 3) the following structure of lemmas is the result:

- Headword, dialect (dialectal region), morphological information/process and etymology are primary information with single values that together define and are dependent on the lemma.
- Each lemma can have many different realizations and each one of them is characterized by a slightly different phonetic realization dependent on the micro–dialectal region it originates from (the specific area within the wider dialectal region where the lemma's realization occurs).
- Each lemma can possibly have different meanings (i.e. polysemy), or be homonymous with other, semantically distinct, lemmas.

ΠΕΔΙΟ (FIELD)	ΛΗΜΜΑΤΙΚΗ ΠΛΗΡΟΦΟΡΙΑ (LEMMA VALUES)
1. ΛΕΞΗ-ΚΕΦΑΛΗ / HEADWORD	ΒΡΟΥΛΟ
2. ΛΕΞΙΚΗ ΚΑΤΗΓΟΡΙΑ (lexical category)	(Ο. ουδ.) (noun, neuter)
3. ΦΩΝΗΤΙΚΟΣ ΤΥΠΟΣ (phonetic type)	['vrulu]
4. ΑΡΧΕΙΟ ΗΧΟΥ ΠΡΟΦΟΡΑΣ (digital record)	WAV
5. ΕΝΑΛΛΑΚΤΙΚΟΙ ΤΥΠΟΙ (alternative types)	Βρόλους ['vrolus] (Παμφ. ΜΙΚΡΟΔΙΑΛΕΚΤΙΚΗ ΠΕΡΙΟΧΗ ΣΥΝΔΕΣΗ) (Pamphylia microdialectal region)
6. ΔΙΑΛΕΚΤΙΚΗ ΠΕΡΙΟΧΗ (dialectal region)	Αϊβαλί (Aivali)
7. ΜΙΚΡΟΔΙΑΛΕΚΤΙΚΗ ΠΕΡΙΟΧΗ (microdialect)	
8. ΜΟΡΦΟΛΟΓΙΚΗ ΔΙΕΡΓΑΣΙΑ (morphological process)	-
9. ΧΡΗΣΤΙΚΟ ΣΗΜΑΔΙ (usage)	1. ΦΥΤΟΛΟΓΙΑ (phytology)
10. ΟΡΙΣΜΟΣ (definition)	Βούρλο (bulrush - reedy plant)
11. ΑΡΧΕΙΟ ΕΠΕΞΗΓΗΜΑΤΙΚΗΣ ΕΙΚΟΝΑΣ	JPG
12. ΠΑΡΑΔΕΙΓΜΑ ΧΡΗΣΗΣ (usage example)	«Έκουψα καμπόσα βρούλα κι πέρασα αρμαθιά τα ψάρια πό πιασα σήμιρα» (a dialectal sentence using the lemma)
13. ΜΕΤΑΦΡΑΣΗ ΠΑΡΑΔΕΙΓΜΑΤΟΣ ΣΤΗΝ ΚΝΕ (modern Greek translation)	('Εκοψα μερικά βούρλα και .....) (the same sentence in Modern Greek)
14. ΘΗΣΑΥΡΟΣ (thesaurus)	
15. ΕΤΥΜΟΛΟΓΙΚΗ ΠΛΗΡΟΦΟΡΙΑ (etymology)	[ΕΤΥΜ ελνστ. βροῦλον] (originates from the Hellenistic βροῦλον)
9. ΧΡΗΣΤΙΚΟ ΣΗΜΑΔΙ	2. ΥΠΟΤΙΜΗΤΙΚΟ (pejorative)
10. ΟΡΙΣΜΟΣ	Ανόητος (fatuous)
11. ΑΡΧΕΙΟ ΕΠΕΞΗΓΗΜΑΤΙΚΗΣ ΕΙΚΟΝΑΣ	JPG
12. ΠΑΡΑΔΕΙΓΜΑ ΧΡΗΣΗΣ	«Ντιπ για ντιπ βρούλου τούτου του πιδί». (a dialectal sentence using the lemma)
13. ΜΕΤΑΦΡΑΣΗ ΠΑΡΑΔΕΙΓΜΑΤΟΣ ΣΤΗΝ ΚΝΕ	(Τελείως ανόητο αυτό το παιδί) (the same sentence in Modern Greek)
14. ΘΗΣΑΥΡΟΣ	---

Figure 1. Draft structural depiction of lemma ΒΡΟΥΛΟ

ΠΕΔΙΟ (FIELD)	ΛΗΜΜΑΤΙΚΗ ΠΛΗΡΟΦΟΡΙΑ
1. ΛΕΞΗ-ΚΕΦΑΛΗ / HEADWORD	ΛΙΩΣΤΡΑ
2. ΛΕΞΙΚΗ ΚΑΤΗΓΟΡΙΑ (lexical category)	Ο. Θηλ. (noun, feminine)
3. ΦΩΝΗΤΙΚΟΣ ΤΥΠΟΣ (phonetic type)	['ʎostra]
4. ΑΡΧΕΙΟ ΗΧΟΥ ΠΡΟΦΟΡΑΣ (digital record)	αρχείο WAV
5. ΕΝΑΛΛΑΚΤΙΚΟΙ ΤΥΠΟΙ (alternative types)	
6. ΔΙΑΛΕΚΤΙΚΗ ΠΕΡΙΟΧΗ (dialectal region)	Αϊβαλί (Aivali)
7. ΜΙΚΡΟΔΙΑΛΕΚΤΙΚΗ ΠΕΡΙΟΧΗ (microdialectal region)	
8. ΜΟΡΦΟΛΟΓΙΚΗ ΔΙΕΡΓΑΣΙΑ (morphological process)	Παραγωγή (derivation)
9. ΧΡΗΣΤΙΚΟ ΣΗΜΑΔΙ (usage)	1. ΥΠΟΤΙΜΗΤΙΚΟ (pejorative)
10. ΟΡΙΣΜΟΣ (definition)	γυναίκα που περιφέρεται εδώ κι εκεί (woman who strolls)
11. ΑΡΧΕΙΟ ΕΠΕΞΗΓΗΜΑΤΙΚΗΣ ΕΙΚΟΝΑΣ (picture)	-
12. ΠΑΡΑΔΕΙΓΜΑ ΧΡΗΣΗΣ (usage example)	«Ξιπόρτσι πάλ'-η λ'ώστρα» (dialectal phrase using the lemma)
13. ΜΕΤΑΦΡΑΣΗ ΠΑΡΑΔΕΙΓΜΑΤΟΣ ΣΤΗΝ ΚΝΕ (Modern Greek translation)	(Ξεπόρτισε πάλι η γυρίστρα) (the same in Modern Greek)
14. ΘΗΣΑΥΡΟΣ (thesaurus)	Συν: *αλλουγυρίστρα, σόρτα, τακιού* (Synonymy: *αλλουγυρίστρα, σόρτα, τακιού*)
15. ΕΤΥΜΟΛΟΓΙΚΗ ΠΛΗΡΟΦΟΡΙΑ (etymology)	[<*λιέμι* με -ωστρα ίσως από επιδρ. άλλων θηλυκών σε -ωστρα] (*λιέμι* with affix *ώστρα*)
16. ΔΙΑΠΑΡΑΠΟΜΠΕΣ (see also)	

Figure 2. Draft structural depiction of lemma ΛΙΩΣΤΡΑ

ΠΕΔΙΟ	ΛΗΜΜΑΤΙΚΗ ΠΛΗΡΟΦΟΡΙΑ
1. ΛΕΞΗ-ΚΕΦΑΛΗ / HEADWORD	ΑΛΛΟΥΓΥΡΙΣΤΡΑ
2. ΛΕΞΙΚΗ ΚΑΤΗΓΟΡΙΑ	(Ον. Θηλ.) (noun, feminine)
3. ΦΩΝΗΤΙΚΟΣ ΤΥΠΟΣ	[aluji'ristra]
4. ΑΡΧΕΙΟ ΗΧΟΥ ΠΡΟΦΟΡΑΣ	WAV
5. ΕΝΑΛΛΑΚΤΙΚΟΙ ΤΥΠΟΙ	---
6. ΔΙΑΛΕΚΤΙΚΗ ΠΕΡΙΟΧΗ	Αϊβαλί (Aivali)
7. ΜΙΚΡΟΔΙΑΛΕΚΤΙΚΗ ΠΕΡΙΟΧΗ	---
8. ΜΟΡΦΟΛΟΓΙΚΗ ΔΙΕΡΓΑΣΙΑ	Σύνθετο (composite)
9. ΧΡΗΣΤΙΚΟ ΣΗΜΑΔΙ	1. ΥΠΟΤΙΜΗΤΙΚΟ (pejorative)
10. ΟΡΙΣΜΟΣ	γυναίκα που περιφέρεται εδώ κι εκεί (woman who strolls)
11. ΑΡΧΕΙΟ ΕΠΕΞΗΓΗΜΑΤΙΚΗΣ ΕΙΚΟΝΑΣ	---
12. ΠΑΡΑΔΕΙΓΜΑ ΧΡΗΣΗΣ	"Ξιπόρτσι πάλ'ι η γ'αλλουγυρίστρα" (dialectal phrase using the lemma)
13. ΜΕΤΑΦΡΑΣΗ ΠΑΡΑΔΕΙΓΜΑΤΟΣ ΣΤΗΝ ΚΝΕ	(Πάλι βγήκε η αλλουγυρίστρα.) (the same in Modern Greek)
14. ΘΗΣΑΥΡΟΣ	ΣΥΝ *λυώστρα, σόρτα* (Synonymy: *λυώστρα, σόρτα*)
9. ΧΡΗΣΤΙΚΟ ΣΗΜΑΔΙ	2. ΙΑΤΡΙΚΗ (medicine)
10. ΟΡΙΣΜΟΣ	πόνος με πρήξιμο γύρω από το νύχι (pain & turgescence around nail)
11. ΑΡΧΕΙΟ ΕΠΕΞΗΓΗΜΑΤΙΚΗΣ ΕΙΚΟΝΑΣ	JPG
12. ΠΑΡΑΔΕΙΓΜΑ ΧΡΗΣΗΣ	"Έχου μνιαν αλλουνυρίστρα στου δαχτύλ'ιμ τσι πουν'εί" (dialectal phrase using the lemma)
13. ΜΕΤΑΦΡΑΣΗ ΠΑΡΑΔΕΙΓΜΑΤΟΣ ΣΤΗΝ ΚΝΕ	(Έχω μια αλλουγυρίστρα στο δάχτυλο και με πονάει.) (the same in Modern Greek)
14. ΘΗΣΑΥΡΟΣ	---
15. ΕΤΥΜΟΛΟΓΙΚΗ ΠΛΗΡΟΦΟΡΙΑ	[ΕΤΥΜ από το ρ. *αλλουγυρίζω*] (from the verb *αλλουγυρίζω*)
16. ΔΙΑΠΑΡΑΠΟΜΠΕΣ	---

Figure 3. Draft structural depiction of lemma ΑΛΛΟΥΓΥΡΙΣΤΡΑ

– For each meaning, different usage examples are essential.

Regarding the relations (between lemmas and meanings of lemmas) we concluded the following:

– Cross reference ("See also") links can be available for connecting lemmas that are semantically / pragmatically / morphologically / etymologically related to each other.
– Synonyms and Antonyms are two semantic relations that apply between lemmas. Both relations relate a lemma meaning with a lemma (the referenced one). Synonym and Antonym links are restricted between a lemma meaning and a lemma from the same dialect.
– There are meanings of different lemmas from different dialects that share the same definition. This relation is labeled "Other Dialect". In contrast with the rest of the relations, "Other Dialect" is a symmetrical relation.

The overall idea (lemma structure and relations) is strictly defined as it is depicted with the Entity Relation Diagram of figure 4.

The following four data dictionaries (tables 1, 2, 3, 4) explain the four sections (sub–schemas) of the overall ERD.

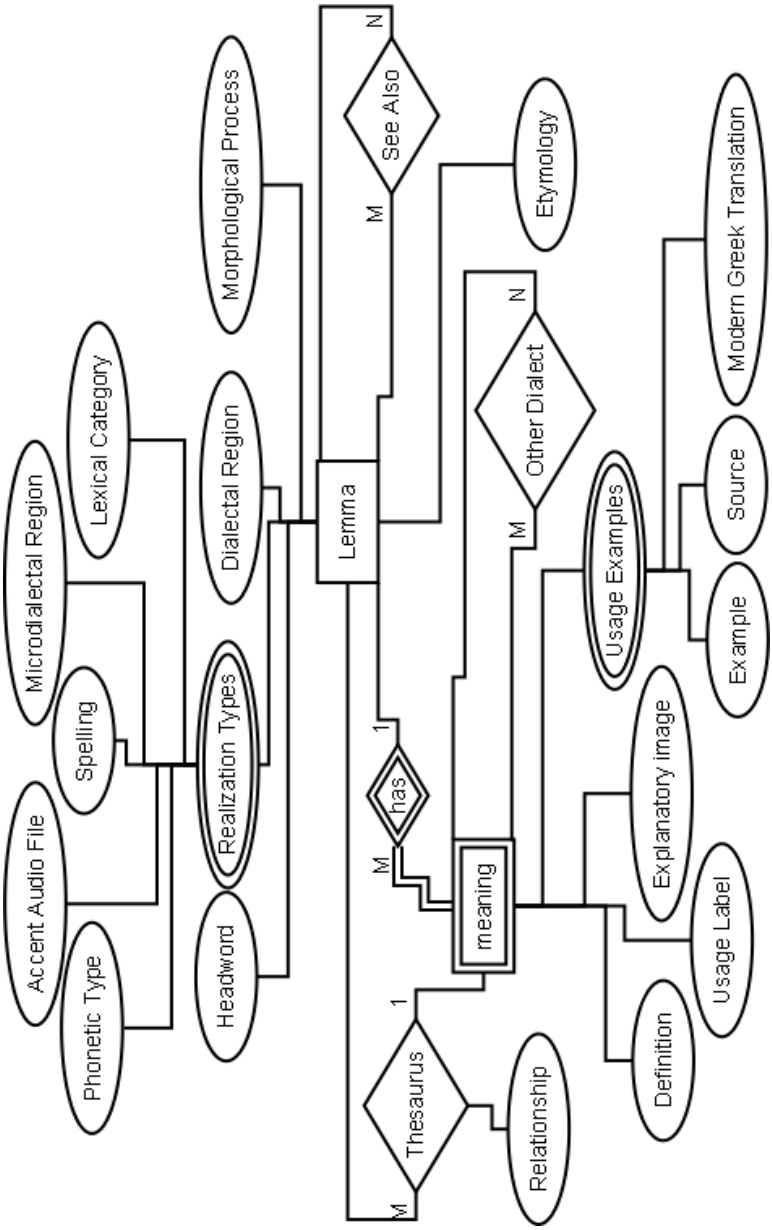
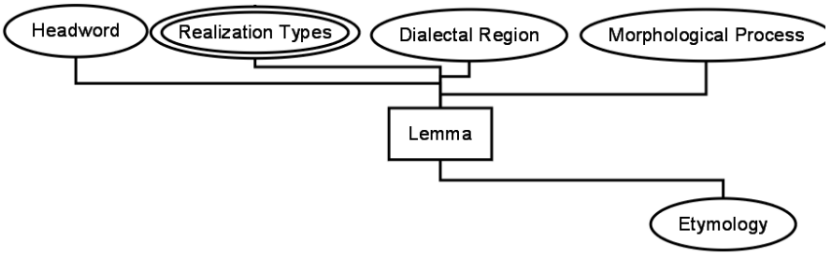


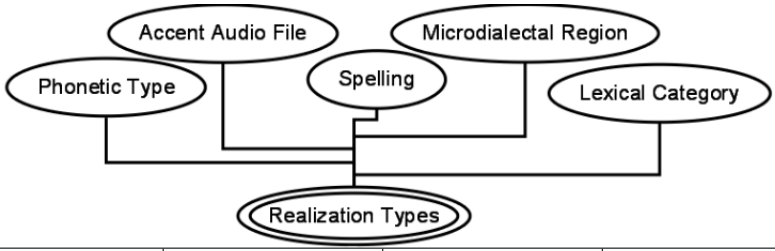
Figure 4. Entity Relation Diagram



Attribute	Definition	Data format	Example
Headword	The canonical form of the word	String containing only capital letters of the Greek alphabet	ΑΛΛΟΥΓΥΡΙΣΤΡΑ
Etymology	Basic information about the origin of the word.	String written in Greek with accents (polytonal)	Από το ρήμα αλλουγυρίζου (from the verb aluji'rizu)
Morphological Process	Different processes involved in word–formation.	A value from a predefined list of morphological processes	Σύνθετο (Compound noun)
Dialectal Region	The region/dialect in which the lemma is found	A value from a predefined list of Dialects	Αϊβαλί (AIVALI)

Table 1. Data dictionary for "Lemma"

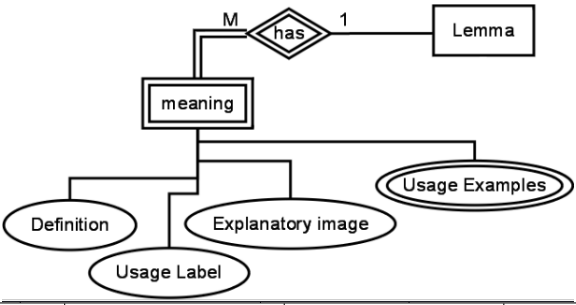
One possible implementation of the conceptual model (ERD) using a relational database is depicted in the relational schema of figure 5 which contains thirteen tables. However, only seven tables are important. The other six tables are lookup tables (listing the set of available values existing) related to some fields of the important tables. The important tables are highlighted (in figure 5) with thicker border and larger font in their title. Four out of seven important tables are the relational equivalents of the main conceptual entity ("Lemma"), the weak entity ("meaning") and the two multiple-valued com-



sub Attribute	Definition	Data format	Example
Phonetic Type	Phonetic transcription of (the examined) pronunciation of the word.	String containing letters of the International Phonetic Alphabet (IPA).	aluji'ristra
Accent Audio	Audio file of the authentic pronunciation of the word	String containing a file path	http://amigre.gr/xyzR1.wav
Spelling	Non–standard graphic representation of pronunciation according to the orthographic rules of Standard Greek, combined with diacritics to annotate any phonological alternations.	String containing the letters of the Greek alphabet and other diacritic symbols (accent, hyphens, parentheses and apostrophes)	αλλουγυρίστρα
Microdialectical Region	Name of a specific area within the wider dialectal region of the lemma in which the realization form is found	Value from a predefined list of microdialectical regions	
Lexical Category	Part of Speech & Gender	Value from a predefined list of lexical categories	Ουσιαστικό Θηλυκό (noun feminine)

Table 2. Data dictionary for "Realization Types"

posite attributes ("Realization Types" and "Usage Examples"). The remaining three important tables are the relational equivalents of the conceptual relations ("See Also", "Thesaurus" and "Other Dialect").

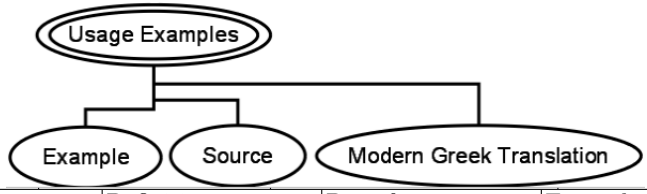


Attribute	Definition	Data format	Example
Definition	Short description of the meaning of a lemma	String in StandardModern Greek	Γυναίκα που περιφέρεται εδώ κι εκεί ("woman who goes around")
Explanatory Image File	Image illustrating the meaning of a lemma.	String containing a file path	http://amigre.gr/xyzM1.png
Usage Label	Formal indication of the context (stylistic/register/other) in which the lemma is used.	A value from a predefined list of domains	ΥΠΟΤΙΜΗΤΙΚΟ (pejorative)

Table 3. Data dictionary for "Meaning"

Only the table MeaningSets (the implementation of the conceptual relation "Other Dialect") needs more explanation. This relation is symmetrical by nature, i.e. whenever a meaning of a certain lemma from one dialect is declared as being the equivalent of the meaning of another lemma from a different dialect, then the reverse is implied. It is the structure of table MeaningSets and the application's logic that assures this symmetry. The other two relations ("See Also" and "Thesaurus") are not symmetrical by nature. This is

reflected in the relational schema (and the application logic). Consequently, the user must define the relation in both directions, in case an instance of them (the "See Also" or the "Thesaurus" relation) is symmetrical.



sub Attribute	Definition	Data format	Example
Usage example	Example (phrase or sentence) demonstrating the usage of the lemma under one specific meaning, in the original dialect	The whole example (the whole string) is written with the letters of the Greek alphabet and other diacritic symbols (accent, hyphens, parentheses and apostrophes)	Ξιπόρτσι πάλ'–η–γ'–αλλουγυρίστρα
Standard Modern Greek Translation	Translation of the usage example into Standard Modern Greek	String in Standard Modern Greek	Πάλι βγήκε η αλλουγυρίστρα
Source	Reference to the source from which the usage example was extracted	String (can be a book, a URL, etc)	

Table 4. Data dictionary for "Usage Examples"

The International Phonetic Alphabet (IPA) which is used in Table 2 – sub–schema for "Realization Types" – is an alphabetic system of phonetic notation based primarily on the Latin alphabet. It was devised by the International Phonetic Association as a standardized representation of the sounds of spoken language. The IPA is used by lexicographers, foreign language students and teachers, linguists, speech–language pathologists, singers, actors,

constructed language creators, and translators. Figures 6 and 7 present the most useful IPA charts.

Another approach to phonetic notation is SAMPA (Speech Assessment Methods Phonetic Alphabet) and it is a machine–readable phonetic alphabet. It was originally developed under the ESPRIT project 1541, SAM (Speech Assessment Methods) in 1987–89. It applied first to Danish, Dutch, English, French, German, and Italian (1989). Later, it applied to Norwegian and Swedish (1992). Subsequently it applied to Greek, Portuguese, and Spanish (1993). It has now been extended to Bulgarian, Estonian, Hungarian, Polish, and Romanian (1996).

6 Implementation

The GUI version of the system is based on two forms: "main form" and "meaning form". Figure 8 presents the main form for the lemma "ΑΛΛΟΥΓΥΡΙΣΤΡΑ". The main form is divided into 3 sections. The upper section provides information on the headword, etymology, morphological process and dialect. The middle section is a two–card panel. The first card in the panel is used for displaying and editing realizations, while the second one is used for providing the meanings list of lemmas. The lower section of form is a panel for hosting the "see also" reference list. A more detailed description of main form's middle section is provided in figure 9 which depicts the second card (meanings list) for the same lemma.

The "meaning form" of a lemma is invoked by an action button once the user selects an item from the "meanings list" of "main form". Figure 10 depicts the "meaning form" presenting certain meanings of the lemma "ΑΛΛΟΥΓΥΡΙΣΤΡΑ". The meaning form is divided into 3 sections. The upper section provides the definition of the meaning, optionally a picture and the usage label. The middle section of the form is a panel for hosting the "usage examples" list. The lower section of the form is a two-card panel. The first card in the panel is used for displaying and editing synonymic/antonymic relations (thesaurus), while the second one is used for handling the equivalents in other dialects.

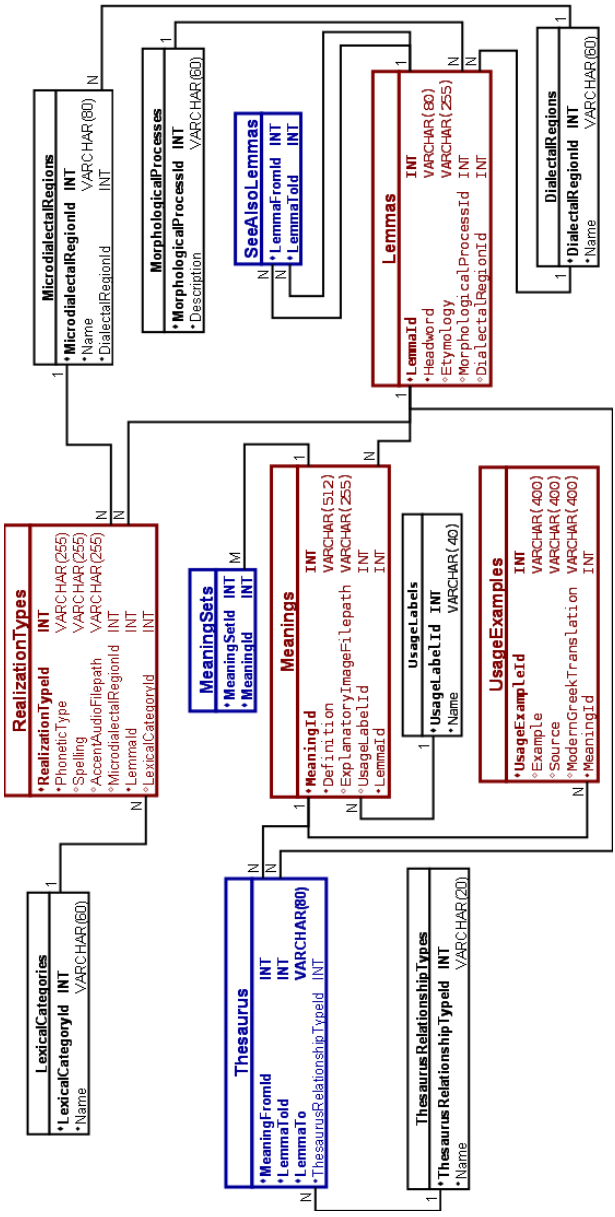


Figure 5. Logical Relational Schema

CONSONANTS (PULMONIC) © 2005 IPA

	Bilabial	Labiodental	Dental	Alveolar	Postalveolar	Retroflex	Palatal	Velar	Uvular	Pharyngeal	Glottal
Plosive	p b			t d		ʈ ɖ	c ɟ	k g	q ɢ		ʔ
Nasal	m	ɱ		n		ɳ	ɲ	ŋ	ɴ		
Trill	ʙ			r					ʀ		
Tap or Flap		ⱱ		ɾ		ɽ					
Fricative	ɸ β	f v	θ ð	s z	ʃ ʒ	ʂ ʐ	ç ʝ	x ɣ	χ ʁ	ħ ʕ	h ɦ
Lateral fricative				ɬ ɮ							
Approximant		ʋ		ɹ		ɻ	j	ɰ			
Lateral approximant				l		ɭ	ʎ	ʟ			

Where symbols appear in pairs, the one to the right represents a voiced consonant. Shaded areas denote articulations judged impossible.

Figure 6. International Phonetic Alphabet (rev. 2005) – Consonants (Pulmonic)

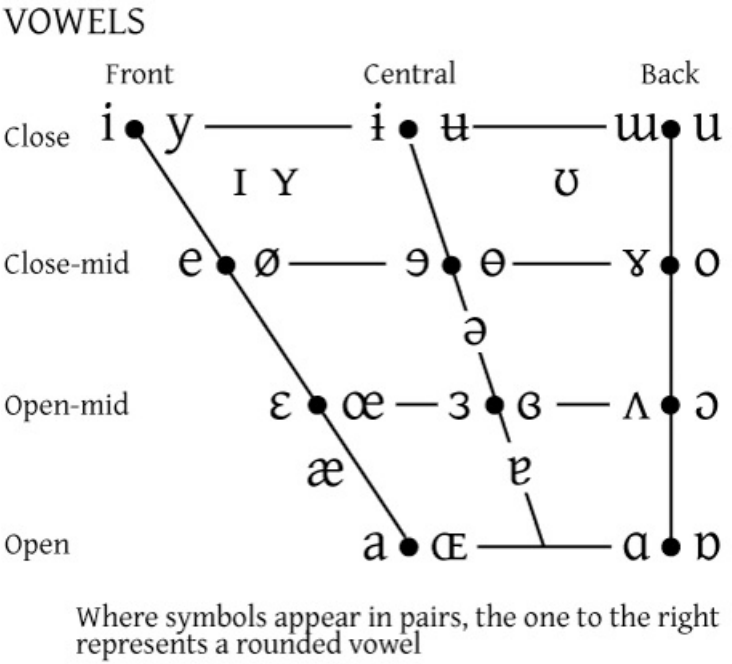


Figure 7. International Phonetic Alphabet – Vowels

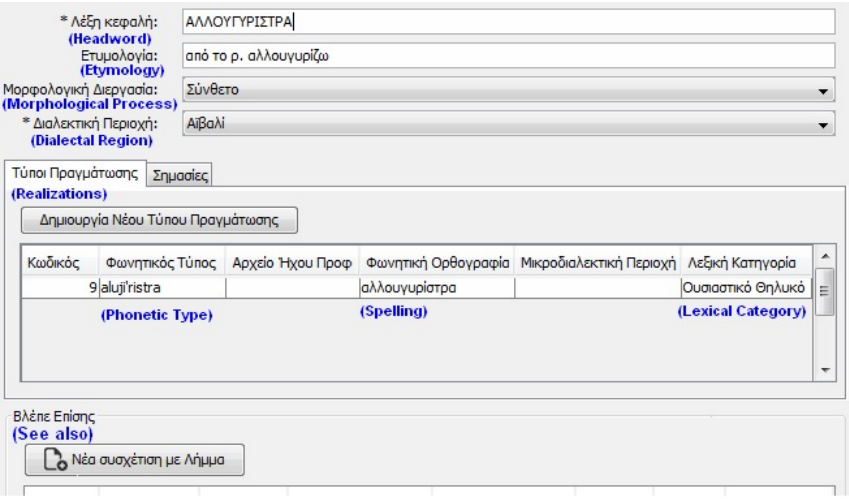


Figure 8. Main form of lemma "ΑΛΛΟΥΓΥΡΙΣΤΡΑ" – Realizations card in front

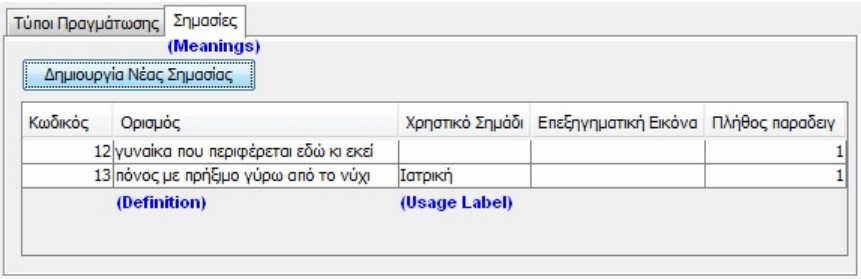


Figure 9. Main form of lemma "ΑΛΛΟΥΓΥΡΙΣΤΡΑ" – Meanings card

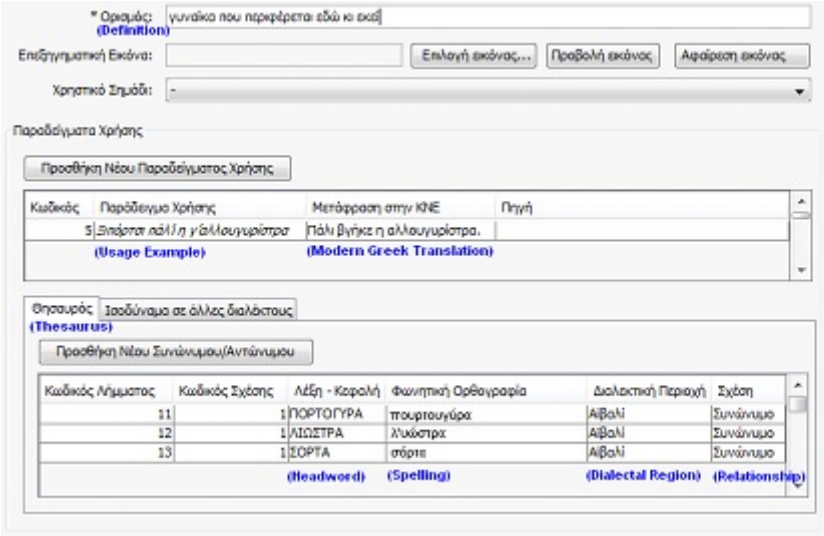


Figure 10. Meaning form of lemma "ΑΛΛΟΥΓΥΡΙΣΤΡΑ" – Some meaning with its thesaurus

According to Analysis (Requirements) and Design sections, the implemented system provides the user with the following character sets for editing the relevant fields:

- Etymology
  - Greek Polytonal
  - Loan characters from other alphabets in case of loan words (e.g. characters from the Turkish alphabet)
- Phonetic type
  - IPA
- Spelling (Phonetic Orthography)
  - Modern Greek
  - Accents
  - Hyphen, parentheses, apostrophe.

7 Ultimation

1. As it is well known, users express most of their arguments during the final stage of lifecycle of the initial development of the system (**acceptance,**

installation, deployment) and during the maintenance stage. There was no exception to this rule in the electronic lexicon. In our case users explained that the origin (etymology) attribute of a lemma, should be denoted with respect to the sources and the conventions of the originating language (from where the lemma comes). Therefore, in the case of a multi–dialectal dictionary, the etymology attribute can contain words from any of the languages that have influenced the dialect. Consequently, we came up with a solution which was to provide (in a latter development) visual keyboards for any affecting language. In the case of the 3 dialects of AMiGre the influencing alphabets are Greek, Ancient Greek and Turkish. Figure 11 presents a 3–card virtual keyboard with cards for lowercase ancient Greek, uppercase Ancient Greek and Turkish. These, together with Modern Greek (provided by the physical keyboard), permitted users to enter the required etymology of each lemma without any restriction.

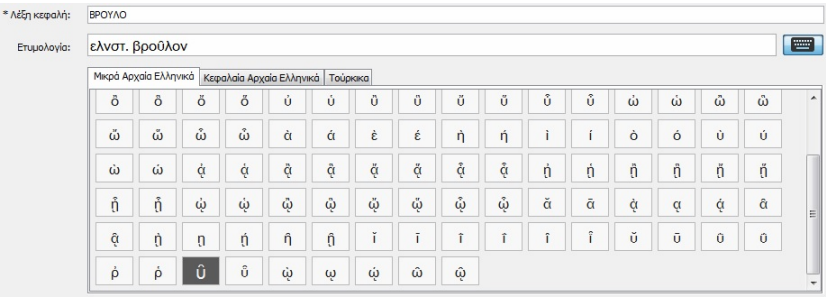


Figure 11. Virtual keyboard for Ancient Greek and Turkish

2. Regarding the phonetic type attribute and the intonation of vowels there are two options available. The first option is to place an accent before stressed syllable. The second option is to use the stressed version of the vowel. Both options are semantically equivalent. However, the first option is easier to use and implement in a system because there is no need to support both versions (with and without accent) of the IPA

symbols used for the vowels. Thus, in the initial development, our system permitted only the vertical accent before the syllable for denoting the intonated vowel. Therefore, in figure 8 the phonetic type is denoted with "aluji'ristra". After the initial development and during the maintenance stage a need to support accented versions of vowels came up. To cope with this demand, we have extended the system with virtual keyboards for easily inserting any IPA symbol, with and without accents. Figures 12, 13, 14 and 15 present the virtual keyboards used for the phonetic type attribute.

Figure 12. Alphabetic (IPA) symbols

The combinations of IPA symbols (fig. 12) with some of the Diacritic symbols (fig. 13) produce the accented IPA symbols. Figure 16 present another lemma that has phonetic type with accented IPA symbols.

Figure 13. Diacritic symbols

3. According to the linguistic analysis, the "spelling" attribute represents a non–standard graphic representation of pronunciation according to the orthographic rules of Modern Greek (target language), combined with diacritics to annotate any phonological alternations. Therefore, we concluded that the domain of values for the spelling attribute can be strings containing the letters of the Greek alphabet and definite diacritic symbols (accent, hyphens, parentheses and apostrophes). This conclusion was followed for the first implemented version of the system. Since the system got into the production stage, we became recipients of very specific demands for improvements in the spelling attribute. The users pointed out that some words, since they originate from other languages, have vowels that do not exist in the standard spelling system of the target language (Modern Greek). Since the studied dialects have a prominent number of loan words originated from other languages, this remark could not be ignored. So, we had to provide some way for the users of system to be able to understand how to pronounce the dialectal words, without engaging them to read phonetic (IPA) symbols. The simplest way was to allow the insertion of the grapheme representations (letters) used in the originating language of the loan words for the representation of vowels that do not exist in the target language (Modern Greek). The outcome was a small number of vowels with grapheme representations having umlauts (u, e and α, with umlaut). These letters were included in one virtual keyboard added for editing the spelling (phonetic transcription) attribute. Figure 17 represents this virtual keyboard.

Figure 14. Compound symbols

4. As we have already described, each lemma can have many different realizations and each one of them is characterized by a slightly different phonetic realization dependent on the micro–dialectal region it originates.

This principle drove our design and we built a system where each realization is characterized by one micro–dialectal area (sub–area of the wider dialectal region). However, in the production stage of the system, users pointed out that a single realization can exist in more than one micro–dialectal regions (of the dialectal region). To comply with this lately defined requirement we modified the data schema design and the application. The "Microdialectal region" attribute was changed to become multi–valued and the corresponding GUI item changed to hold a list of values (the domain of each value is the set of micro–dialectal regions existing for the dialect of lemma). Figure 18 presents the realizations of the Pontic lemma "ΟΜΜΑΤΟΤΖΑΤΖΙ" where the third realization (third line) has 3 micro–dialectal regions (Τραπεζούντα, Χαλδία, Σάντα).

Figure 15. Arrow symbols

5. Another feature of the system which was not defined in the analysis phase but emerged during the development phase was the content of a "see also" reference list item for a destination lemma having more than one meaning and/or having more than one realization type. The solution we decided to follow was to represent the definition of each meaning for the referenced lemma and the spelling of each realization for the referenced lemma in the "see also" list item. The lower section of figure 16 represents the empty reference list of a lemma but we can see the plural number used in titles of the relevant columns (Spellings and Meanings).

6. Another worth–mentioning feature of the system is that Synonymic/Antonymic links can refer to another lemma of the same dialect which may or may not be present in the system. Figure 19 represents a meaning of lemma "ΛΙΩΣΤΡΑ" with its "thesaurus" (table providing

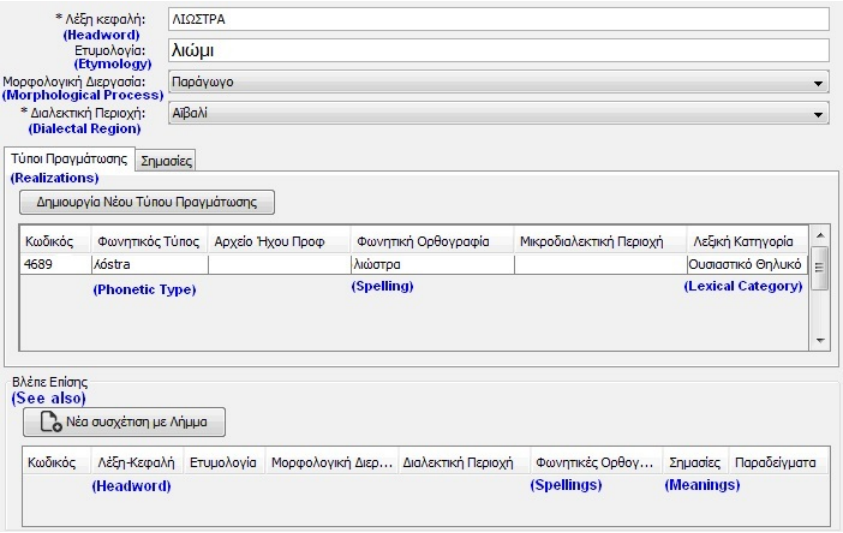


Figure 16. Main form of lemma "ΛΙΩΣΤΡΑ" with accented IPA symbols in phonetic type

the Synonymy/Antonymy links). In this figure, we can see 3 links (referring lemmas ΣΟΡΤΑ, ΤΑΚΙΟΥ and ΑΛΛΟΥΓΥΡΙΣΤΡΑ) but only the third one is present in the system. A comparison of figures 10 and 19 denotes that in the thesaurus table of the newer version (lower section of figure 19) we have replaced Spelling with Etymology. Note that we have also added a new column (Source type) in the usage examples table (middle section of figure 19).

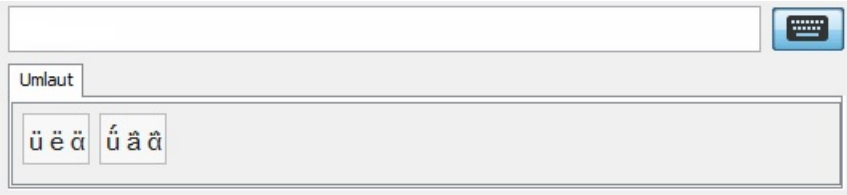


Figure 17. Virtual keyboard for graphemes representing additional vowels

7. So far, we haven't seen any example of equivalent in another dialect. Figure 20 is the main form of the Cappadocian lemma "ANTET" which has a single meaning. In figure 21 we provide the lower section of the meaning form displaying the "equivalent in other dialects" card. As depicted in figure 21 the only available meaning of the Cappadocian lemma "ANTET" has an equivalent meaning in the lemma Aivaliot lemma "ANTETI".

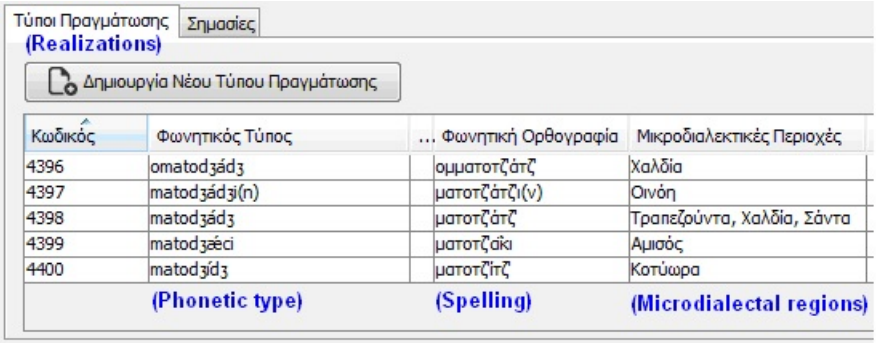


Figure 18. Realizations of Pontic lemma "OMMATOTZATZI"

8. During the production stage of the system's lifecycle we have noticed that users diverged from the regulations for writing the usage examples of lemmas. Usually users exploited the copy/paste feature of the operating system in order to enter characters not provided directly by the applications for the "usage example" attribute. For example, the value "Αςλαγεύω το μήλον" was entered in the usage example attribute of the lemma "ΑΣΛΑΕΥΩ". This value has Greek characters that are according to the regulations but also contains a Turkish character (the second character in the string value). The phrase "Αςλαγεύω το μήλον" as value in the usage example, together with the value "τουρκ. AŞlamak" (i.e. "from Turkish aŞlamak") in the etymology attribute of the same lemma can be an indication of how native speakers of the dialect could possibly write the dialectal word in their documents ("Αςλαγεύω"). However, this indication is hidden inside one of the meanings of a lemma. We suppose that it could be better to provide another attribute (named "indicative

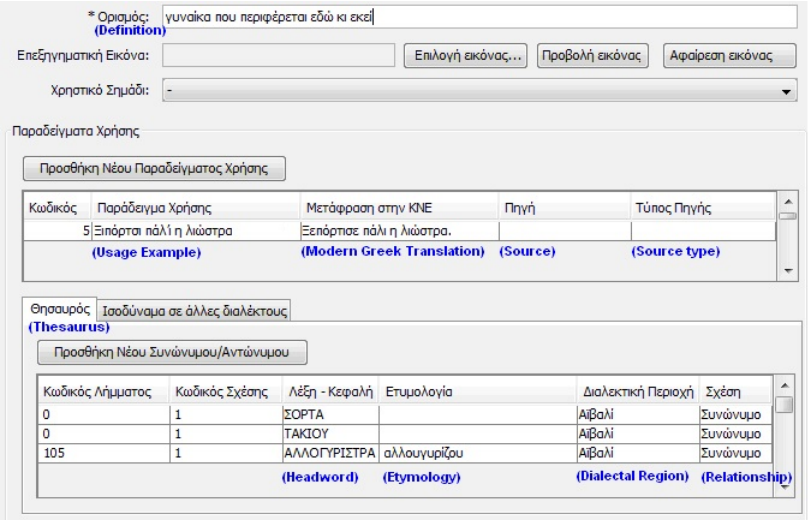


Figure 19. Meaning form of lemma "ΛΙΩΣΤΡΑ" – A meaning and its thesaurus

writing") in each realization of the lemma. In this way, the "indicative writing" would be directly available in the main form of lemmas and moreover it would be differentiated in each realization (micro–dialectal regions). This is the only feature that is not implemented in the system's ultimation because it is denoted very late, but we consider it very valuable for next multi–dialectal lexicons.

8 Extended – Improved design

The data schema (ERD) for supporting the ultimate system is given in figure 22.

9 Conclusions

TDGDAM's projected macrostructure includes ca. 2,500 entries from each of the three dialects of Asia Minor Greek (a total of ca.7,500 entries). These entries are drawn from collected vocabulary solely from the three dialects concerned and exclude all vocabulary found in Standard Greek (unless

differently used). Their listing is based on alphabetical, and not onomasiological, organization, accessed via dynamic searching options (Xydopoulos, 2012).

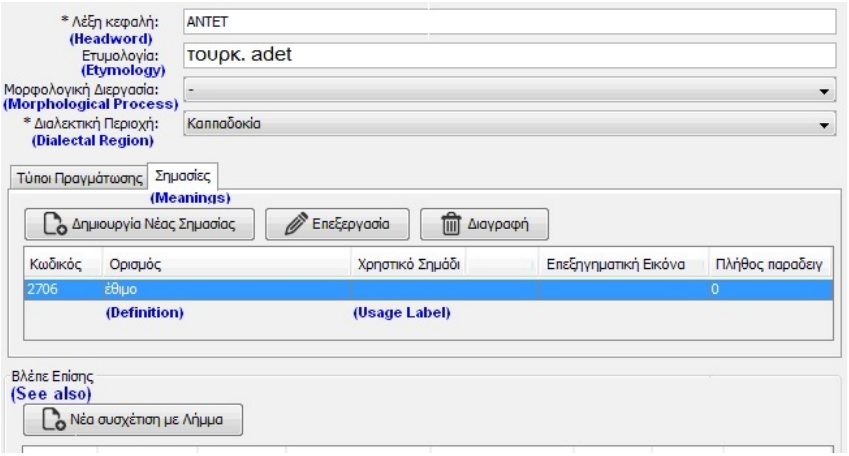


Figure 20. Main form of Cappadocian lemma "ANTET"

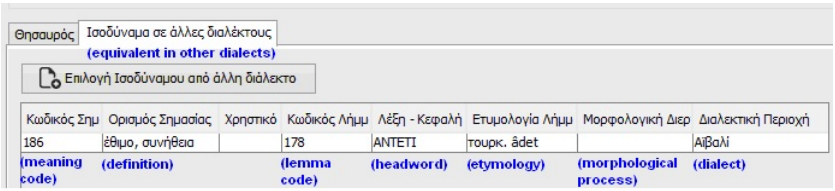


Figure 21. Meaning form of lemma "ANTET" – Equivalents in other dialects card in front

Acknowledgments

This research is co–financed by the European Union (European Social Fund – ESF) and Greek national funds through the Operational Program "Education and Life–long Learning" of the National Strategic Reference framework (NSRF) – Research Funding Program: "THALIS. Investing in knowledge society" through the European Social Fund. We thank Angela

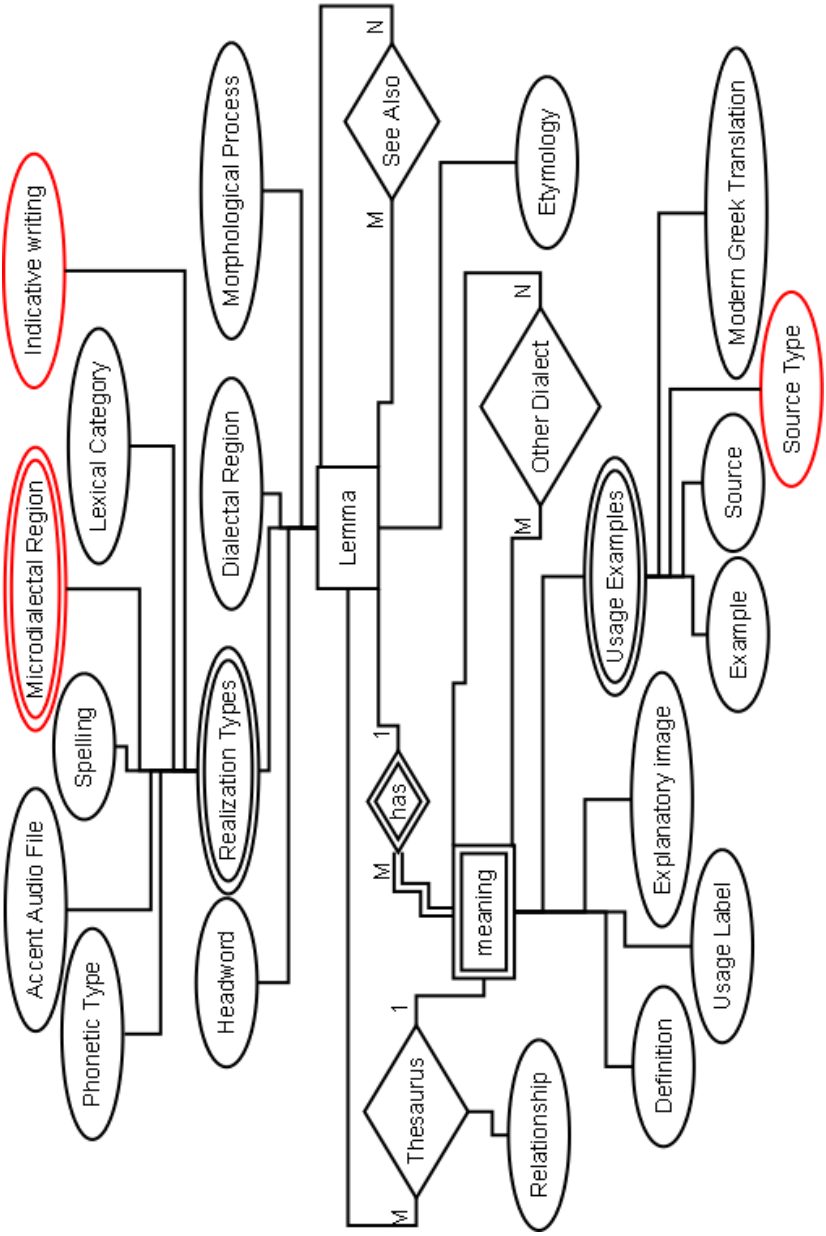


Figure 22. The improved ERD

Ralli, Professor of Linguistics at the Department of Philology of the University of Patras, who is the Coordinator of the whole AMiGre project. We also thank George J. Xydopoulos, Associate Professor of Linguistics at the Department of Philology of the University of Patras, who set the linguistic requirements for the dialectal electronic lexicon.

References

"The Dictionaries of ILSP". Institute for Language and Speech Processing. Accessed December 22, 2016. http://www.xanthi.ilsp.gr/dictionaries/

"Homonym". Wikipedia. Accessed December 22, 2016. https://en.wikipedia.org/wiki/Homonym

"Online Dictionaries". Portal for the Greek Language. Приступљено 22.12.2016. http://www.greek-language.gr/greekLang/modern_greek/tools/lexica/index.html

"Opposite (semantics)". Wikipedia. Accessed December 22, 2016. https://en.wikipedia.org/wiki/Opposite_(semantics)

"Polysemy". Wikipedia. Accessed December 22, 2016. https://en.wikipedia.org/wiki/Polysemy

"Synonym". Wikipedia. Accessed December 22, 2016. https://en.wikipedia.org/wiki/Synonym

Barbato, Marcello and Alberto Varvaro. "Dialect dictionaries". *International Journal of Lexicography* Vol. 17 no. 4 (2004): 429–439

Béjoint, Henri. *Modern lexicography: An introduction.* Oxford: Oxford University Press, 2000

Burke, Sean. M. "The Design of Online Lexicons". In *A practical guide to lexicography*, Piet van Sterkenburg, 240–249. Amsterdam: John Benjamins, 2003.

Durkin, Philip. "Assessing Non-standard Writing in Lexicography". In *Varieties of English in writing: The written word as linguistic evidence*, Raymond Hickey, 43–60. Amsterdam: John Benjamins, 2010.

Geeraerts, D. "Principles in Monolingual Lexicography". In *Wörterbücher / Dictionaries / Dictionnaires. Ein internationales Handbuch zur Lexikographie. An International Encyclopedia of Lexicography. Encyclopédie internationale de lexicographie*, Franz Josef Hausmann et al., 287–296. Berlin: Walter de Gruyter, 1989.

Karanikolas, Nikitas N., Eleni Galiotou, George J. Xydopoulos, Angela Ralli, Konstantinos Athanasakos et. al.. "Structuring a Multimedia Tri-dialectal

Dictionary". In *LNAI 8082: International Conference on Text, Speech and Dialogue*, Ivan Habernal and Vaclav Matousek, 509–518. Berlin Heidelberg: Springer, 2013.

Keymeulen, J. Van. "Pilot project: A Dictionary of the Dutch Dialects". Paper presented at the 14th EURALEX International Congress, Fryske Akademy, Ljouwert/Leeuwarden, 6–10 July 2010

Landau, Sidney. *The art and craft of lexicography (2nd edition)*. Cambridge: Cambridge University Press, 2001

Markus, Manfred and Reinhard Heuberger. "The architecture of Joseph Wright's English Dialect Dictionary: preparing the computerized version". *International Journal of Lexicography* Vol. 20 no. 4 (2007): 355–368

Papadopoulos, A. *Historical Dictionary of the Pontic Dialect*. Athens: Epitropi Pontiakon Meleton, 1958

Penhallurik, Rob. "Dialect Dictionaries". In *The Oxford History of English Lexicography Volume II*, A. P. Cowie, 290–313. Oxford: Oxford University Press, 2009.

Rys, Kathy and Jacques Van Keymeulen. "Intersystemic correspondence rules and headwords in Dutch dialect lexicography". *International Journal of Lexicography* Vol. 22 no. 2 (2009): 129–150.

Themistocleous, C. "Cypriot Greek Lexicography: An Online Lexical Database". Paper presented at the 15th EURALEX International Congress, University of Oslo, Oslo, 7-11 August 2012.

Thomason, Sarah G. *Language Contact. An Introduction*. Edinburgh: Edinburgh University Press, 2001.

Thomason, Sarah G. and Terrence Kaufman. *Language Contact, Creolization, and Genetic Linguistic*. Berkeley: University of California Press, 1988.

Tsalidis, Christos, Mavina Pantazara, Panagiotis Minos and Elena Mantzari. "NLP Tools for Lexicographic Applications in Modern Greek". In *eLexicography in the 21st Century: New Challenges, New Applications (Proceedings of eLex2009)*, S. Granger and M. Paquot, 457–462. Louvain-la-Neuve: UCL Presses, 2010.

Xydopoulos, George J. "Metalexikografikes paratiriseis sta leksika Benardi ke Syrkou (Metalexicographic comments to Benardi and Syrkou dialectal dictionaries)". *Patras Working Papers in Linguistics* Vol. 2 no. 1 (2012): 96–113

Xydopoulos, George. J. and Angela Ralli. "Greek dialects in Asia Minor: Setting lexicographic principles for a tridialectal dictionary". Paper presented at the 5th MGDLT Conference, Ghent, Belgium, September 2012

Zgusta, Ladislav. *Manual of lexicography*. The Hague: Mouton, 1971

Safeguarding and Promotion of Fragile Archival Material from the Lazić Family Private Library[1]

UDC 930.25:004.9
DOI 10.18485/infotheca.2017.17.1.2

ABSTRACT: This paper introduces possibilities of using new digital technologies in preservation and presentation of fragile archival material and other library material in the context of smart libraries and it presents a contribution to the draft proposal for standardization of digitization of such materials. Cutting-edge digital technologies and tools provide new possibilities for humanities researchers who can now conduct contemporary and transparent research of rare and fragile archival material. This paper will show that archival material, in the era of Industry 4.0 and the Internet of Things can be an extremely important part of the promotion of a culture in the context of smart libraries. However, the focus will be on standardization of digitization which is currently underway at the University Library.

KEYWORDS: smart libraries, new digital technologies, archival material, library material, digitization, standardization.

PAPER SUBMITTED: 15 July 2016
PAPER ACCEPTED: 4 November 2016

Vasilije Milnović
milnovic@unilib.rs

Milena Kostić
mkostic@unilib.rs

Matea Milošević
milosevic@unilib.rs

University Library Svetozar Marković"

1 Introduction

A smart library concept has been developing in the era of Industry 4.0 where the Internet of Things (IoT) has become all pervasive (Evans, 2011).

[1] The paper was written within the project entitled Safeguarding the fragile collection of the private archive of the Lazic family, coordinated by the University Library "Svetozar Markovic" and funded by the British Library Endangered Archives Programme

The role of smart libraries in such a context is to become information hubs where new technology and accompanying concepts are introduced and experimented with. Nowadays, libraries are centers where new technologies are studied and research conducted. The aim of such undertakings is development of more productive, concise, comprehensive and easier research activities. Libraries guide us to the digital world. They are places where new technologies are implemented (Min, 2012; Younis, 2012). Libraries are considered potential business incubators as these technologies can be applied in business and public sector. They embrace new technologies and overcome digital exile. The main goal of the project Safeguarding the fragile collection of the private archive of the Lazic family is to digitize and thus preserve for posterity extremely valuable private collections owned by the non-governmental organization "Adligat" (the Lazić Library). The collection that will be digitized and presented to the public consists of several sub-collections: law books, war publications, periodicals, calendars and archival material. The material has an added historical value in the context of marking the First World War centenary. Moreover, the material is invaluable to the academic libraries as it includes unique and rare publications some of which cannot be found elsewhere in libraries. For the first time the public will be able to see "Pregled listova", a confidential journal of the Serbian government in exile printed in Geneva, and an overview of news by the allies and enemies. The material is precious to various researchers from historians and sociologists, over anthropologists and philologists, to librarians and archivists. It is especially important to note that the whole project is a collaborative effort between an academic institution and non-governmental sector – the University Library "Svetozar Marković" and the non-governmental organization "Adligat". This form of cooperation is valuable as a contemporary form of project activities which will be rounded off with the participation of the private sector. Having gone beyond its primary functions, the library meets the needs of other types of organizations and professions, which is – bearing in mind the inevitable multidisciplinary approach – generally important for the advancement of a society, smart cities concept and smart libraries as the main information hubs of the future knowledge society.

2 Initiatives of the University Library towards the smart libraries concept

Bearing in mind the importance of open access approach for attaining the knowledge society and fulfilling the role of academic libraries in knowl-

edge dissemination, this extremely important and fragile material will be presented via the cutting-edge device Magic Box which is suitable for interactive display of such delicate and vulnerable material. The material can be searched through on a transparent touch screen while the physical publication can be seen behind it. In addition to digitized publications, photo galleries, 3D models and films can be displayed in this device. This device provides a unique experience regarding the presentation of rare and fragile publications whose availability is usually restrained. The University Library "Svetozar Marković" is the first institution in the East Europe and the third library in the world to own this device which will become a pulsating window into the world of interests of various experts in humanities.

Figure 1. Magic Box

.

According to the grant agreement signed between the University Library "Svetozar Marković" and the British National Library, all digitized material will be available at the British Library website[2] and in the special digital repository of the University Library whose construction is underway. Following the project plan, master files will be stored in tiff format and on a cloud platform – Therefore, owned by the University Library. A digital repository with searchable content is under construction where the documents will be available without contrastive background and calibration cards. Software

[2] http://eap.bl.uk/database/overview_project.a4d?projID=EAP833;r=9741#project_outcome

docWorks, which is the main model of a programme for organizing contents in Magic Box, will be used for the preparation of the material.

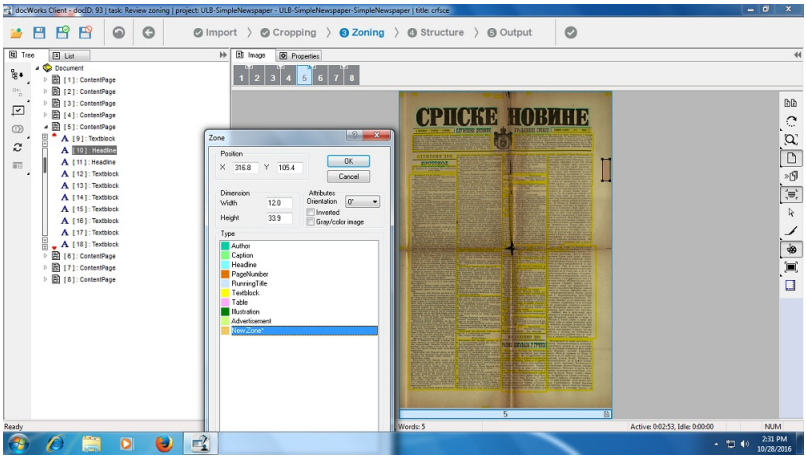


Figure 2. docWorks

.

Preparation of material in docWorks[3] consists of the following steps:

1. cropping page surfaces of some digital objects;
2. zoning objects by segmenting pages into blocks and columns with surfaces defined for OCR (Optical Character Recognition) and determining their type as regards the function in the object: titles, text, author, pictures etc.;
3. arranging the structure of the object (bullet, chapter, article) by connecting titles and contingent text;
4. correcting text and metadata
5. creating ready objects in the form of METS/ALTO files suitable for display in Magic Box and the repository.

All aforementioned steps imply an automatic analysis and then manual correction.

[3] http://content-conversion.com/#docworks-2

As PDF offers only limited search possibilities by keywords, the University Library has adopted METS/ALTO files. It should be pointed out that the University Library first obtained such files within the Europeana Newspapers project, and then it was the first library with the National Library of Serbia, to produce them without limit. The simplest definition of such files is that institutions can use them to adapt their digital objects so as to get searchable documents. This was also done with the digitized documents within the current project.

METS and ALTO standards were established for the easier description of digitization of printed material. The idea was to separate descriptive information from the content so that digital objects could be handled easily as when all the data are in one XML file (as is the case with TEI – Text Encoding Initiative format), the XML file is too large.

METS (Metadata Encoding and Transmission Standard) is an XML based open standard established by the Congress Library in Washington in 2001. It is used for permanent storing of files which describe digital objects, printed media (books, newspapers, journals), audio and video material etc. METS usually contains several types of metadata standards: descriptive, administrative, structural information, standards regarding physical and logical structure and links to other digital objects, pictures, audio-visual and textual files.

ALTO (Analyzed Layout and Text Object) is an XML based open standard also established by the Congress Library in Washington in 2001. It is used for digital description of the printed page layout so that the original page could be reconstructed. This file comprises content of an individual page of a digital document and can contain tags with more data about the very object. It describes styles, layout and the type of information blocks.

Digital objects structured in such a way will be much more operative and will provide a unique search – in the physical space when it comes to new technologies Magic Box and online when it comes to a specialized digital repository – with the results that will provide a detailed overview of collection contents to the user and fast and easy search by the keyword. In addition to the contents of the digitized object, ready metadata and expert literature accompanying the theme of the object will be provided for users. Thereby the book is not only digitized but also datafied. Books become data sets, i.e. text corpora, and words becomedata points. Hence, machines become readers.

3 The University Library contribution to the draft proposal of the digital repository of the Republic of Serbia

As the implementation of the project is still underway, project outcomes will be presented at the expert conference in which renowned experts in the field and potential partners from other fields will take part. Thereby wider public will be given an opportunity to participate actively in such activities. An added value of the project is the creation of a wider public discussion on the topic of smart libraries and digital citizenship. At the same time international projects are not only a means to increase funding of public cultural and scientific institutions, but to provide opportunities for professional development and advancement.

Up to now in Serbia digitization has been associated with scanning. Understandably, as a rule, international standards such as Technical Guidelines for Digitizing Cultural Heritage Materials[4] or standards recommended by UNESCO[5] are consulted when drafting general recommendations until a wholesome national standard is tailored. In coordination with the British Library, the University Library "Svetozar Markovic" has adopted new digitization concepts which contain new standards.

Several digital repositories have been established[6] within the projects of digitization of literary material in Serbia, which proved Serbia's readiness to take up a challenge set by an information technology revolution. Despite the fast development and success in the field, library staff who are skilled in the digitization of literary material have successfully responded to the pace set up by the most advanced centres in the field. The first drafts of standards, which should harmonize the quality of digital objects, have been sketched. Following successful digitization practice of literary material, the University Library strives to meet the digitization standards of the British Library, who is the world leader in the field.

[4] http://www.digitizationguidelines.gov/guidelines/FADGI_Still_Image_Tech_Guidelines_2016.pdf (accessed 27 Oct 2016)

[5] http://www.unesco.org/new/fileadmin/MULTIMEDIA/HQ/CI/CI/pdf/mow/digitization_guidelines_for_web.pdf (accessed 27 Oct 2016)

[6] These are some national digital repositories: Digital Repository of the National Library of Serbia (http://www.digitalna.nb.rs/(http://www.unilib.rs/istorijske-novine/pretraga), Digital Library of Matica Srpska (http://digital.bms.rs/ebiblioteka/), Digital Repository of the Belgrade City Library (http://dlibra.bgb.rs/dlibra)

The purpose of the Endangered Archives Programme (EAP) is safeguarding material by providing appropriate storing conditions and by digitization, which assures long-term preservation and wider access to the digital objects (End, End). To meet the programme demands, it is extremely important to adapt digital copies to archival standards. Some standards, which the British Library sets to its partners, are already on the list of successfully applied digitization practices of Serbian institutions. However, there are standards that pose a challenge which improves digitization practices and motivates institutions which cultivate and promote cultural heritage in accordance with the latest concepts such as Industry 4.0 and smart libraries.

The recommendations for creating digital copies of physical objects refer to the resolution, minimum being 300dpi (dots per inch), or format tiff (Tagged Image File Format). In addition to these widely known and broadly applied standards, by meeting the demands of the Endangered Archives Programme, we were introduced with the new practices and standards which helped us improve our own digitization guidelines.

Above all, work on this project, as a contribution to the draft proposal of digitization standards for old and rare books at the University Library, has brought about two new dimensions: colour and length.

As a geographical map is closely determined by dimensions and colour, a scan or a photograph (a picture of a digitized object)is more precisely determined with a ruler and a color calibration card. Thereby every scan or every photograph provides more precise data about the genuine physical characteristics of an object, by giving unquestionable information about its precise dimensions and colors.

Different sources of light have different temperatures. Photographs which are taken under different conditions do not portray precise colours of an object. To avoid this we use White Balance, which is a source based correction. Colours corrected in such a way change balance between red, green and blue curves (RGB curves), but not their shape nor position. Therefore, what is changed in the photograph is the light not colour shades. Moreover, when photographs are taken with different devices we do not get identical colours. That is why colour management was developed to make such conversions more subtle and to improve the quality of the photograph.

One of the main colour management tools is a calibration card or a colour scheme. Adjusting colours on the photograph to portray genuine colours of an object is a challenge in digital and analogue photography. A Swedish company from Gothenburg QPCards AB developed a cost-effective and efficient

solution which is based on anopen correction software and calibration cards which can be bought. There are several versions available.

QPcard is only one of the accepted models of calibration cards. These cards are industrially acquired, i.e. manufactured in factories, usually made of cardboard and cannot be printed through one's own efforts, especially if they should satisfy a particular standard. They usually contain a ruler and a color scheme.

Color correction in pictures with a QPcard is done by calibration software, QPcolorssoft 501, which can be downloaded from the manufacturer's website. This software and a QP card set up in the scanning surface create a reference profile. They should be set up indirectly to the camera sensor, neither at an angle nor in the shade during the scanning. As the white balance is fixed and a suitable colour profile with given parameters created, a reference correction profile is created and all other pictures taken with the QPcard can be calibrated.

Calibration consists of the following steps: the QPcard is selected, it is adjusted to the colours of the card so that every colour takes the right place in the pattern, then a specific colour profile and a reference calibration profile are created. When the profile is created all the pictures taken under the determined conditions and the same white balance can be corrected as a group.

If in addition to tiff, as a suitable format for pictures, 300dpi resolution, as the basic minimum resolution, sufficient and necessary for OCR, completely covered surface of the scanned object, from edge to edge, a colour scheme with a ruler was included. A digital object created in such a way would represent an almost ideal picture of the physical object.

Quality control and evaluation have to be carried out over the whole digitization process and the potential future standard. Within the current project, the University Library followed the EAP guidelines and took the following measures:

1. at the end of every workday it is necessary to carry out quality control of the scans;
2. scans are copied to the external hard drive and stored at separate locations (back up);
3. when the pictures are stored one needs to check if they are rotated properly so that the content can be read;
4. prior to permanent storing of the material MD5 checksum is applied to detect errors.

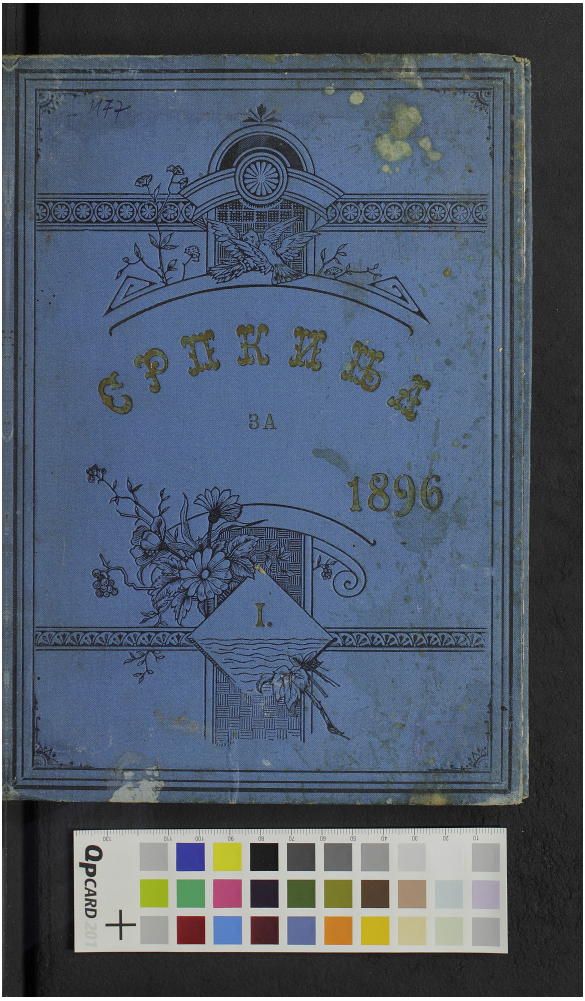


Figure 3. QPcard

.

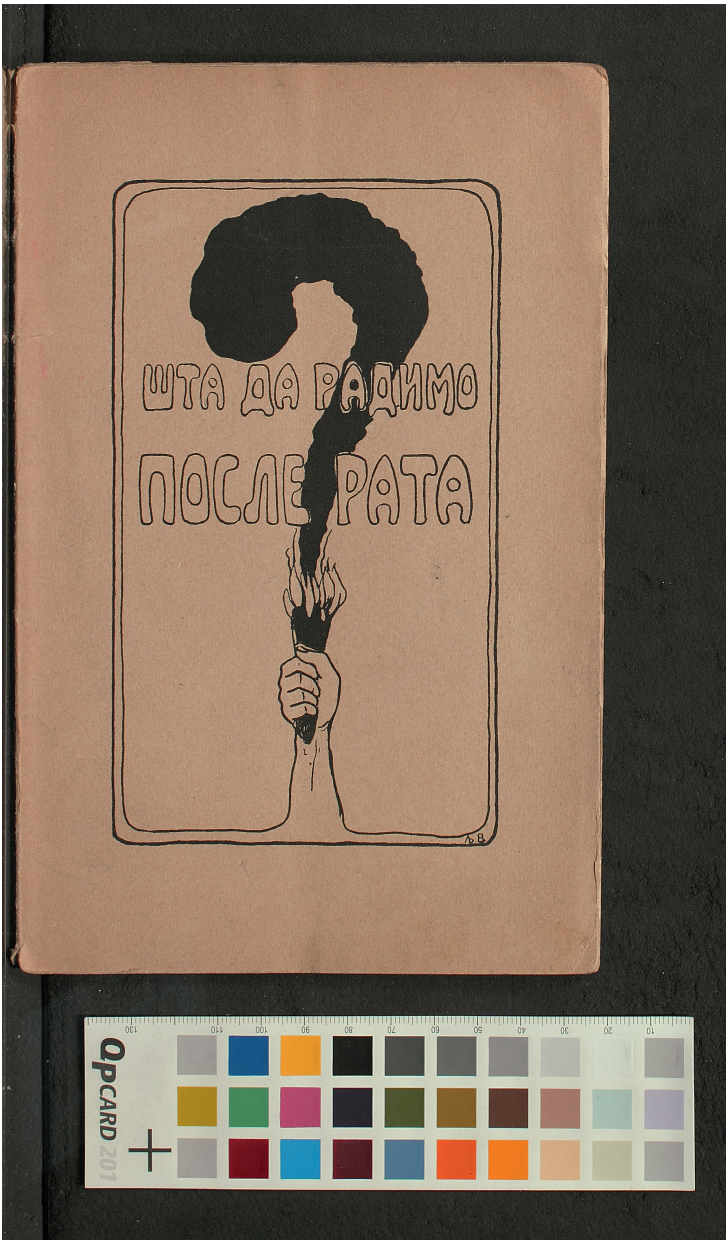


Figure 4. An example of colour correction with QPcolorssoft 501

The MD5 checksum for a file is a 128-bit value, something like a fingerprint of the file. There is a very small possibility of getting two identical checksums of two different files. This feature can be useful both for comparing the files and their integrity control.[7] To understand how this value works,one needs to imagine that there are two physically separated huge files for which it should be determined whether they are the same or different, but which cannot be joined or compared directly. With the MD5 checksumit is sufficient to calculate control sums for both files and then to compare them and determine whether the files are the same or different.

Application of best practice of world leaders in digitization can be a solid foundation for defining standards at the institution level but also the standards regarding digitization of literary material. Thereby, standardized scanning could serve as a basis for harmonizing quality of digitized material of different institutions. Harmonization of digital objects and institutional cooperation would unite somewhat dispersed energy of digitization centres in the country which would result in a unique and conjoint digital repository at the national level.

4 The Importance and Contents of the Digital Collection within the Endangered Archives project

There are several sub-collections in the collection of the Lazić Library which was digitized in the Endangered Archives project by the University Library. The digital collection, created in the project has 50,055 digitized pages.

4.1 Sub-collection "Law Books"

Law books, which have sparked great interest of the British Library award committee, can be divided into two groups: the ones published by Geca Kon and other law publications. Geca Kon was the most important Serbian and Yugoslav publisher at the beginning of the 20th century. Having glorified the victory of the Serbian army, he was imprisoned during the First World War. In the Second World War, being a Serbian Jew, he was held captive in the camp with his family where he died.

[7] http://www.fastsum.com/support/md5-checksum-utility-faq/md5-checksum.php

At certain periods his publishing activity was on a much larger scale compared to all other publishers in Serbia altogether. Many of the most important titles of Serbian literature, history and law were published at the time thanks to this renowned man. However, considering his participation in the First World War against the German and his Jewish origin, his publications were massively destroyed in both world wars. For example, his books were publicly burnt during the Austro-Hungarian military occupation. In the Second World War the German used him as an example of anti-Jewish propaganda especially during 1941. The books from his bookstore were confiscated and transferred to Vienna. During the War his publishing company was taken over by a fascist publishing company "Jugoistok", which, when communists came to power, became "Prosveta", the most important state publishing company between 1950 and 1990.Also, after the liberation, Geca Kon's publications were considered unpopular as they were published by the "class enemy". That is why some of his publications are a rarity nowadays, especially some law books which have never been digitized and whose physical condition is poor. Considered expert literature, these publications were not printed in large circulation. Due to the fact that their topics describe pre-war capitalist legal regulation, they were massively destroyed and represent a rarity nowadays. This collection stands out as a number of books came from the personal libraries of people or institutions which were important for the Serbian state and law. Notes on books, stamps and Ex Libris, usually the only trace of their existence, are in themselves invaluable and quite rare. Here we list only some of them:

1. *Political and Legal Discussions*, 460 pages, by Slobodan Jovanović published in 1910. This is an extremely rare document as the author was the president of the Press Bureau in the Serbian government in exile during the First World War, the president of the Government in exile during the Second World War, the president of the Serbian Royal Academy, rector of the University of Belgrade, Public Law Professor at the Faculty of Law and its dean. As works of a member of old establishment, his literature was destroyed after the Second World War. Back at the time, anyone who was reported to have kept his works in their house would risk police interrogation. He was exculpated in 2007. This copy belongs to the legacy of professor Petar Bingulac (1897–1990) which is now owned by the Lazić family.
2. *Bill of Sale, lectures at the Faculty of Law*, 149 pages by Živojin M. Perić published in 1920. This is an extremely rare publication by the

most renowned Serbian lawyer between the two world wars. This copy of the book has a stamp of the Supreme Court Library.
3. *Original Slavic Law before the 10th century*, 130 pages by Dr Karlo Kadlec is an extremely rare study on law.

The total number of publications in this sub-collection is 132. None of them has ever been digitized so their digitization is justified and will be welcomed by scientific, expert and general audience. The second part of the sub-collection of law books consists of 29 rare law books which were not published by Geca Kon. Some of them are mentioned here:

1. "Amendments of the regulations on disabled veterans and other regulations on the disabled" 130 pages published in Belgrade in 1938. This is an extremely rare publication which is especially important for the retrospective overview of the position and treatment of a great number of war veterans.
2. "Law on the Ministry of Foreign Affairs and diplomatic and consular offices of the Kingdom of Yugoslavia abroad" 32 pages, published in Belgrade in 1929 by the State Printing House of the Kingdom of Yugoslavia. This is a very rare publication for the official use, which was used by all Serbian consular offices in Europe, where renowned Serbian writers such as Ivo Andrić, Miloš Crnjanski and Jovan Dučić worked. This copy belongs to the legacy of professor Petar Bingulac (1897–1990).
3. "Collection of laws of the new age, proclamation of His Royal Highness the King as of January 6 1929", 67 pages, printed by Dr. Časlav M. Nikitović in 1929. This is one of the very rare publications as the laws in it remained in force even after the introduction of dictatorship. It belongs to the legacy of the judge Slobodan Ćirić.

4.2 Sub-collection "War Publications"

Serbian war publications (1914–1918) are very specific library and archival material considering the tragic events that happened to Serbian people at the time. After winning the battles at Cer and Kolubara, Serbia was attacked on three sides, so after the tough battles it was decided that the Serbian state, government, parliament, military and a large number of Serbian people should retreat to Greece. They retreated via Albanian mountains during winter, so there were less than 200 000 people who survived this tragic journey. Serbian state continued to exist in exile with its government

and the people, but without its territory, which is a specific case in history. One of the most important evidence of the continuity of the Serbian state is the war publishing activity. Therefore, all the publications printed in Corfu (where the seat of the Serbian government was), in Bizerte in Tunisia (where the great number of the wounded was) and in Thessaloniki, are treated as national heritage of great importance according to the Serbian law. This also applied to some of the rare publications of the Serbian emigration printed in Geneva, Nice, London and in the USA. The Lazić Library owns one of the most valuable collections of war publications, 156 of which are digitized within the Endangered Archives project. Some of them are:

1. "English-Serbian Dictionary" by Đorđe A. Petrović, 192 pages, published in Thessaloniki in 1918. This is an extremely rare publication used by the soldiers to communicate with the English medical staff.
2. "Law on the Amendments of the National Bank Law", 5 pages, published in Corfu in 1916. This is an extremely rare publication as there isn't a library in Serbia that possesses this document.
3. "Secret subversive organization, a report from the trial at the military court to the officers in Thessaloniki", 638 pages, printed in Thessaloniki in 1918. This is an extremely rare publication as the same officers WHO HAD previously organized the overthrow and assassination of the royal family Obrenović in 1903 were charged for an attempt to kill Serbian crown prince. This is the most important legal process held during the war.
4. "Serbian school day in France 13/26 March 1915", 242 pages, published in Niš in 1915. This is a very rare publication written during the retreat of the Serbian army. This publicationis considered especially valuable and rare. The role of the publication was to show that Serbian army and people were supported by the Allies.

4.3 Sub-collection "Periodicals"

1. "Pregled listova" is an extremely rare publication which will be presented to the public for the first time. This is a confidential informative journal for the members of the Serbian government in exile, printed in Geneva in low circulation (most probably less than 50 copies, 3 are only known to exist today). It is interesting that the journal was printed on different types of paper, on a hectograph, depending on the war atmosphere. "Pregled listova" is actually an overview of war publications, the selection

of the most important texts of the allies and enemies with the aim to inform the Serbian government about the media picture of Serbia and current war activities. Marked "confidential", it was available only to the highest state and war officials. The majority of copies were destroyed. If this periodical is digitized, that would be the first time that 4500 pages of this journal owned by the Lazić family became available to the general public.
2. "Misao" is an extremely rare Serbian war journal published in England. There were only four volumes, all owned by the Lazić family. They were edited by the then Serbian intellectuals in England. All four volumes are digitized.
3. "Krfske novine", 500 pages. Thisis a very important publication as Serbian literary works were published in it, poetry in particular, later considered a classic and the most important work created during the war. The publication comes from the legacy of Stevan Bešević who was one of the editors.

4.4 Sub-collection "Calendars"

According to the British Library Award Committee this is a very valuable collection. Calendars were once favourite periodicals, published once a year, on a low quality paper, in a form of a book covering different popular science topics and entertainment, to be read throughout the whole year. In addition to low quality paper, the calendars were not preserved but thrown away at the end of the year which is why they are extremely rare nowadays. Digitization of some calendars from the 19^{th} century is a real challenge considering the fragility of the paper. That is why the owners would not make them available in physical form to wider audience, not even under restricted conditions. This is the first time some of them will become available to the public.

1. "Vardar: a calendar for 1898", 62 pages, printed by the bookstore Lj. Jokisimović in 1897. This is an extremely rare calendar which cannot be found at the National Library of Serbia.
2. "Orao: an illustrated calendar for 1889 which has 365 days", 176 pages, printed in Novi Sad. This is a very rare publication as the texts of Serbian writers can be found there.
3. "Pastime calendar for 1855" by Svetozar Stojadinović published in Zemun, 67 pages, written before the reform of the Serbian alphabet in 1868.

It is extremely rare and it is considered national treasure according to the Serbian law.

4. "Srpkinja: an illustrated calendar for women from 1896", 168 pages, printed in Kikinda. This is one of the rarest publications from this collection written exclusively for women.

Overall, there are 32 publications in this collection.

4.5 Sub-collection "Archival Material"

Archival material about the First World War includes several letters, notes, postcards and other documents such as an original war poster - a call for help to Serbia in 1916, a photograph of a soldier on the day of Bulgarian capitulation, several short letters, extremely rare war postcards-photographs of Serbian Refugee Theatre from Bizerte (Tunisia), etc. From the perspective of the First World War centenary this material is invaluable. There is an extremely valuable collection of letters and telegrams addressed to the family of Field Marshal Živojin Mišić (predominantly to his wife Lujza) sent on the occasion of his death in 1921. This sub-collection includes the speech held on Živojin Mišić's funeral. The majority of these materials have never been published and the wider and scientific public is not familiar with their content. The total number of publications digitized within this collection is 105.

5 Conclusion

All things considered, it is clear that promotion and presentation of national heritage by using modern technologies and in the context of new concepts of smart libraries, the Internet of Things and Industry 4.0, cannot be carried out without a framework or standards at the national level. A contribution of the University Library to the draft proposal for standardization of digitization is only the first step towards reviving valuable archival and library material in Serbia. It emphasizes the importance of the heritage and being a part of an international project it presents the material comparatively and transparently in the framework of joint European heritage.

References

"Endangered Archives Programme. Guidelines for photographing and scanning archival material". Available at http://eap.bl.uk/downloads/guidelines_copying.pdf. Accessed 1.7.2016.

Evans, Dave. "The Internet of Things: How the Next Evolution of the Internet Is Changing Everything", 2011. Available at http://www.cisco.com/c/dam/en_us/about/ac79/docs/innov/IoT_IBSG_0411FINAL.pdf. Accessed 11.7.2016.

Min, Byung-Won. "Next generation library information service – smart library", *International Journal of Software Engineering and Its Applications* Vol. 6 no. 4 (2012): 171–194

Younis, Mohammed I. "SLMS: a smart library management system based on an RFID technology". *International Journal of Reasoning-based Intelligent Systems* Vol. 4 no. 4 (2012): 186–191

Open Educational Resources and Developing Countries: One Critical View

UDC 37.014(1-773)
DOI 10.18485/infotheca.2017.17.1.3

Marija Pantić
p.mmmmmm@gmail.com

University of Belgrade
Faculty of Philology

ABSTRACT: This paper examines the global socio-economic context of the emergence of open educational resources and their special role in the educational systems of the developing countries. The central place is reserved for the very idea of education, which is elaborated through a selection of criticisms and analyses of educational reforms and through questions of basic conditions, methods of teaching and expected results in education. Special attention is given to global poverty, that is, uneven global development, and in line with that to the increasingly aggressive reproduction of the existent socio-economic relations by means of uneven education. The consideration of these general conditions results in an inevitably sceptical and critical approach to the ideological, technological and pedagogical aspects of the initiatives that create and promote open educational resources.

KEYWORDS: open educational resources, developing countries, globalisation, inequality

PAPER SUBMITTED: 16 October 2016
PAPER ACCEPTED: 23 December 2016

1 What are open educational resources?

Open educational resources (OER) are any type of educational materials that are in the public domain or are released with an open licence, and that

can therefore be copied, used, adapted and re-shared freely and legally[1].
In 2001 the Massachusetts Institute of Technology announced that it would be making all of its university courses freely accessible for non-commercial use, which led to the emergence of open courses. The term "open educational resources" was formally adopted in 2002 at the Forum on the Impact of Open Courseware for Higher Education in Developing Countries, convened by the United Nations Educational, Scientific and Cultural Organization (UNESCO).

In 2005 the Centre for Educational Research and Innovation (CERI) of the Organisation for Economic Cooperation and Development (OECD) began a twenty-month study on the purpose, content, quality and funding of OER (Hylén, 2006, 1). The final report was published in 2007, and its conclusion is that OER can significantly affect curricula, pedagogy and assessment in traditional institutions; teachers, who instead of being "sages on the stage" have already become "guides on the side", will also lose their mediating role in providing learning materials. The need for assessment and recognition of competence acquired outside formal learning settings is likely to grow. Adapting to this, established institutions of higher education could have assessment as their primary goal instead of teaching (Center for Educational Research and Inovation, 2007, 125). The report presents these trends as inevitable and treats them pragmatically, not critically.

The 2012 Paris OER Declaration was adopted at the 2012 World OER Congress, held at the headquarters of UNESCO in Paris, and it recommends that governments, within their capabilities and powers:

1. foster awareness and use of OER;
2. facilitate enabling environments for use of information and communication technologies (ICT);
3. reinforce the development of strategies and policies in OER;
4. promote the understanding and use of open licensing frameworks;
5. support capacity building for the sustainable development of quality learning materials;
6. foster strategic alliances for open resources;
7. encourage the development and adaptation of OER in a variety of languages and cultural contexts;
8. encourage research on OER;

[1] UNESCO. "Open educational resources". Promotional video, 00:00-00:35 (3:05). Published 2012. Accessed 10 August 2016, http://www.unesco.org/archives/multimedia/?s=films_details&id_page=33&id_film=2573

9. facilitate finding, retrieving and sharing of OER;
10. encourage the open licensing of educational materials produced with public funds.

2 The purpose of education

The question of the purpose of education is often overlooked in discussions about the necessary reforms and available resources in education. Can education be expected to bring about the development of personal potential and social consciousness? What is the basis of an authentic social contribution, and how important is the authenticity of one's learning experience?

Creativity, individuality, social wealth and culture, as well as social awareness and responsibility, are central issues both in the contemporary critical pedagogy of capitalist societies and in socialist reforms implemented in the Soviet Union and elsewhere during the twentieth century.

Anton Makarenko, one of the founders of Soviet pedagogy, thought that the development of personality and character was the fundamental purpose of education, which therefore had to encompass all aspects of life. He pointed out that the full potential of individuals and collectives could be achieved only through respect and the understanding of individual and collective inclinations, abilities and needs. For that reason, fixed patterns could not be tolerated in education, and teaching resources had to be appropriate to the circumstances. Thanks to him, work, which had to be creative, collective and productive, was given an important place in education. Such work was to allow for personal expression, cooperation and the establishment of discipline, and the realisation of its social purpose – that is, its contribution to the economic strength of the society – would bring satisfaction, self-esteem and a sense of belonging (Makarenko, 1965).

The English professor and author Ken Robinson, an international adviser on education in the arts to governments, non-governmental organisations and educational and cultural institutions, notes four basic purposes of public education: economic development; the understanding of one's own identity and the identity of others, as a condition for overcoming conflicts of cultures, traditions and world views; the development of social awareness and responsibility; and the development of personal inclinations and abilities[2].

[2] Ken Robinson. "How to change education: From the ground up". YouTube video, 01:00-10:00 (24:02). A video recording of a lecture held at the Royal Society of

OER represent fixed patterns, and as means of teaching they are appropriate to the particular circumstances of the educational systems in which they are created, and thus to the individual and collective inclinations, abilities and needs that exist within these circumstances and that these educational systems recognise. Furthermore, OER contribute primarily to the economic development of developed countries, and accordingly to the expansion of already dominant cultures and world views.

3 Global education

Globalisation is not only an economic and political but also a cultural process, which has a decisive influence on education. Criticism of the globalisation of education cannot be developed independently of the general criticism of globalisation as a comprehensive process that involves and encourages competition between unequals. The greatest economic powers, and the greatest advocates of globalisation, argue that competition is the main driver of development. The application of this idea to education has far-reaching and tragic consequences. Students compete for a place in schools and universities; schools and universities compete for financial support and survival. Curricula and the organisation of teaching and learning activities in developing countries are being adapted to global developmental needs, which are determined by the developed countries. Standardised tests and school rankings should provide equal criteria for success, whereas unequal local conditions and different local needs are being crudely neglected.

The implementation of the Programme for International Student Assessment (PISA) began in 2000, and now half a million fifteen-year-olds from 65 countries and local administrations participate in it. The 2014 OECD report on PISA shows that the socio-economic context is the main determinant of the success for the participating countries, and emphasises the importance not just of the overall funding of education but also of equal distribution of available resources, which achieves the same quality of education for all children (OECD, 2014). Critics of this programme, such as participants in the TV debate "Is schooling for all a realistic goal?"[3] and the signatories of the

Arts in London on 1 July 2013. Uploaded by "The RSA" 18 July 2013, https://www.youtube.com/watch?v=BEsZOnyQzxQ

[3] Al Jazeera English. "Is schooling for all a realistic goal?". TV debate, 12:20-15:20 (24:24). Al Jazeera English, 2014. Accessed 12 May 2016, http://www.aljazeera.com/programmes/insidestory/2014/04/schooling-all-realistic-goal-2014410151530585416.html

letter to the main coordinator of the PISA programme, Andreas Schleicher (Meyer and Zahedi, 2014, i), point out that it does not take into account a whole range of skills that are essential for the development of children. According to them, it significantly narrows the concept of education, and success at this test does not reflect the real quality of a given educational system. Without disputing the importance of good results in mathematics and language, they refer to other important aspects of education, including creativity, self-confidence and the ability to interact. Additionally, the very practice of testing and ranking, although generally inevitable and present to some extent in all educational systems, represents a specific approach to education, which, depending on the stated objectives, can be more or less productive, but also extremely destructive. The enormous pressure on children of the countries successful at the PISA test causes serious damage to their health (Zhao, 2010), yet, in the global competition that the programme promotes, the educational systems of these countries are put forward as models of excellence (Zhao, 2014).

In 2001 the United States adopted the No Child Left Behind Act, which forced students, teachers and schools to compete against one another. Similar to that of PISA, the testing required by this act was of narrow focus and was not sufficiently adapted to the specific inclinations of children and to the specific circumstances of their upbringing, and it was conducted under a threat of closure for schools that failed to achieve an adequate success. Schools were required to guarantee that all children would acquire a prescribed minimum competence in reading, writing and mathematics, while programmes that did not directly contribute to this goal would lose support or be completely abolished (Leyva, 2009, 10). Although it is promoted as legislation aimed at helping disadvantaged children, many critics argue that the effects of its application are exactly the opposite. And this is what the children say about these tests:

"I just can't handle it. I just can't. Oh my, I'm so embarrassed"! "I have the whole world on my shoulders. This goes on my report card, then my college education, and that gets me a good job". "My hands were sweaty. I almost started crying and I had a funny feeling in my stomach. I hope I never feel that way again, especially [because of] some test [in which] I can't show the amazing work I did all year". "Why judge them only on one test? Why not homework? How about how neat they write, how much they've improved throughout the year..." "This test doesn't show the work you can do best,

like writing, my favorite. Every day I grow stronger, but do you see that? No..." "I feel that we shouldn't have to be nervous this young".[4]

Understanding the phenomenon of OER requires an examination not only of the technological circumstances of their conception, and the conditions and possibilities for their further development, but also of the socio-economic and ideological framework within which they are alternately promoted as a complement to the institutional programmes, a solution for the lack of institutional capacity and a radical challenge to the dominance of institutions (Knox, 2013, 4). Their emergence coincides with the first PISA test, with the No Child Left Behind Act in the United States and with the World Education Forum held in Dakar in 2000, which will be discussed in more detail below. All these initiatives are devoted to the increase of "human capital"[5][6](Kwon, 2009), instead of to the realisation of human potential, and to the education of a "globally competitive workforce" (Spellings, 2012), often to the detriment of the development of locally relevant competences.

In a capitalist society the educational system is subservient to the interests of capital, and its main purpose is the legitimisation and reproduction of existing social relations (Bukharin, 2001). The social value of education, as of any other social programme, loses importance, and one's education becomes a question of one's inclinations and personal benefit, which serves to explain the exclusion of the educational system from the domain of social organisation and responsibility. Everyone who expects to benefit from education is free to invest in it, in accordance with their expectations, but is also obliged to bear the risk of that investment. A preoccupation with their own ambitions and success leads students to competition at the expense of cooperation and a concern for the collective progress and learning environment (Leyva, 2009, 10)(Saunders, 2010, 23). OER rely on the self-initiative and

[4] Lerone D. Wilson. *No Child Left Behind*. A documentary film. Boondoggle Media, 2004. YouTube video, 00:00-01:45 (56:02). Uploaded by "Boondoggle" 6 April 2011. Accessed 15 June 2016, https://www.youtube.com/watch?v=yiGN7kVyeaM

[5] UN Department of Economic and Social Affairs (DESA). "Linking education to human capital development". Published 21 March 2011. Accessed 21 May 2016, http://www.un.org/en/development/desa/news/ecosoc/linking_education.html

[6] UN Deapartment of Economic and Social Affairs (DESA). "Human and natural capital as important as financial capital". Published 17 May 2012. Accessed 21 May 2016, http://www.un.org/en/development/desa/news/financing/new-paradigm-job-creation.html

self-directedness of their users, because they do not offer a dynamic learning environment that fosters the formation of a genuine collective. Despite this, they are promoted as the best educational resources: their creators are referred to as the best researchers and educators in their respective fields; their production and distribution involves the latest technologies, which are unreservedly presumed to be the best; and their users are undoubtedly those who want the best for themselves[7](Knox, 2013, 4).

The great interest shown in OER is a reflection of the importance and influence of the institutions that create them, but it also opens the way for the further expansion of their intellectual domination. OER are created mostly by elite institutions of developed countries while all other institutions are easily reduced to passive users of available resources, contributing little or nothing to the global exchange and flow of knowledge. The status quo - the economic and technological superiority of elite institutions - is affirmed and strengthened in this way. With appropriate accreditations, this approach to education could make many traditional institutions redundant. It is not difficult to imagine that, while creating a short-term illusion of flexibility and cost-reduction, this trend could lead in the near future to the monopolisation of global education.

OER are still poorly defined, and one must be wary of possible obstacles to their free use. The formats that require proprietary software for access and modification as well as technically limited platforms for content distribution fundamentally restrict the usability of OER. OER rarely justify their name, given the confusion that exists regarding the appropriate open licence: they are often understood to be free educational resources that do not allow modification, which means that they cannot be translated into other languages and adapted to local contexts (Hoel, 2014) without appropriate permissions issued on request, as is the case with most massive open online courses (MOOC).

4 Global inequality

Providing basic education for all the children of the world by 2015 was one of the Millennium Development Goals of the United Nations. According to recent data, 263 million children aged six to seventeen do not attend school,

[7] UNESCO. "Open educational resources". Promotional video, 00:43-01:03 (3:05). Published 2012. Accessed 10 August 2016, http://www.unesco.org/archives/multimedia/?s=films_details&id_page=33&id_film=2573

with 93.3 million of those living in sub-Saharan Africa and 100.8 million in south Asia[8]. According to a report published by the United Nations in 2014, 125 million children lack the basic skills of reading and writing after four years of primary education, which calls into question the quality of teaching and teachers (UNESCO, 2014, i). The World Education Forum in Dakar in 2000 set six goals that were to be achieved by 2015 (UNESCO, 2000, 15–17):

1. expanding and improving comprehensive early childhood care and education, especially for the most vulnerable and disadvantaged children;
2. ensuring that by 2015 all children, particularly girls, children in difficult circumstances and those belonging to ethnic minorities, have access to and complete free and compulsory primary education of good quality;
3. ensuring that the learning needs of all young people and adults are met through equitable access to appropriate learning and life skills programmes;
4. achieving a 50 per cent improvement in levels of adult literacy by 2015, especially for women, and equitable access to basic and continuing education for all adults;
5. eliminating gender disparities in primary and secondary education by 2005, and achieving gender equality in education by 2015, with a focus on ensuring girls' full and equal access to and achievement in basic education of good quality;
6. improving every aspect of the quality of education, and ensuring their excellence so that recognised and measurable learning outcomes are achieved by all, especially in literacy, numeracy and essential life skills.

UNESCO warns that these goals cannot be achieved without a significant change in the approach of states to the provision of education. Lack of political will at the national level is accompanied by a lack of political will on the part of the international community. When these goals were set it was agreed that every country with a plan would have the resources for the realisation of that plan, which has not happened in practice.[9]

[8] UNESCO Institute for Statistics. "263 million children and youth are out of school". Published 15 July 2016. Accessed 22 August 2016, http://uis.unesco.org/en/news/263-million-children-and-youth-are-out-school

[9] Al Jazeera English. "Is schooling for all a realistic goal?". TV debate, 02:55-03:15 (24:24). Al Jazeera English, 2014. Accessed 12 May 2016, http://www.aljazeera.com/programmes/insidestory/2014/04/schooling-all-realistic-goal-2014410151530585416.html

Globally, the number of out-of-school children in conflict zones has reached 61.9 million in 2016.[10] The number of Syrian children not attending school, whether they are in war-torn Syria or in exile, is at the moment 2 million and is growing daily[11]. In the schools of Gaza in September 2014, after two months of the bombing by Israel, collective therapy for students rather than lessons marked the delayed start of the school year (Weibel, 2014).

Wars prevent development, and the destruction of infrastructure can turn developed countries into developing ones (Zolnikov, 2013). Infrastructural problems make coping with diseases and natural disasters difficult, and corruption and organised crime obstruct efforts to solve infrastructural problems. [12] In the case of Haiti, the recovery from the 2010 earthquake and the cholera epidemic that followed was complicated by political instability, the irresponsibility of international organisations (Pilkington, 2016) and the poor coordination of numerous charitable organisations and initiatives (Knox, 2015). It is estimated that about half of Haitian children will never enter a classroom.[13] Schools in developing countries often have no electricity, drinkable water and basic sanitary conditions. Those schools that have some kind of electricity might not also have internet access. And even when they have internet access they lack qualified ICT teachers (Cave, 2013)(Mungai, 2011). Already lacking adequate resources and the necessary support, teachers in these countries are forced to work with a large number of students, which significantly limits their capacity for innovation in educational programmes. Overlooking these limitations may result in the

[10] UNESCO Institute for Statistics. "263 million children and youth are out of school". Published 15 July 2016. Accessed 22 August 2016, http://uis.unesco.org/en/news/263-million-children-and-youth-are-out-school

[11] Save the Children. "Children of Syria". Published 2016. Accessed 15 August 2016, http://www.savethechildren.org/site/c.8rKLIXMGIpI4E/b.7998857/k.D075/Syria.htm

[12] UN Interregional Crime and Justice Research Institute. "Organized crime and corruption". Accessed 27 July 2016, http://www.unicri.it/topics/organized_crime_corruption

[13] Al Jazeera English. "Is schooling for all a realistic goal?". TV debate, 19:55–20:00 (24:24). Al Jazeera English, 2014. Accessed 12 May 2016, http://www.aljazeera.com/programmes/insidestory/2014/04/schooling-all-realistic-goal-2014410151530585416.html

alienation of teachers and the collapse of an already fragile educational system.[14]

In such extreme circumstances, questions of innovation in education, in the form of modern methods and technologies and, in accordance with that, the appropriate training of teachers, seem absurd. However, there are schools at which computers arrive before there is a regular power supply: examples of this are the solar classrooms in Uganda[15] and the digital villages in South Africa (Cave, 2013). These are donor projects, but, given that infrastructural development in these countries largely depends on donations, the order of priorities is certainly confusing. Charities and non-governmental organisations implementing these projects, with the support of major corporations such as Intel and Microsoft, state that their goals are "lifting Africa out of the poverty trap by equipping the next generation to work in a global environment"[16], but also "encouraging and supporting the formation of Information Communication Technology businesses".[17]

Many children in developing countries are forced to help their parents feed their families, either by doing jobs that endanger their health or by helping with the housework and the upbringing of the younger children while their parents work.[18] To these children, even free education remains inaccessible. In developing countries, large ICT companies are simultaneously exploiting child labour (Russell, 2016) and sponsoring OER, by means of which they promote themselves.[19]

By using cheap labour, the ICT industry, like any other, achieves higher profits and lower costs of products and services, so the strong interest in

[14] Al Jazeera English. "Is schooling for all a realistic goal?". TV debate, 17:10–17:15 (24:24). Al Jazeera English, 2014. Accessed 12 May 2016, http://www.aljazeera.com/programmes/insidestory/2014/04/schooling-all-realistic-goal-2014410151530585416.html

[15] Maendeleo Foundation. "Mobile solar computer classrooms". Published 2015. Accessed 23 May 2016, http://maendeleofoundation.org/projects/mobile-solar-computer-classrooms

[16] Computers 4 Africa. "About Computers 4 Africa". Accessed 23 May 2016, http://www.computers4africa.org.uk/about/index.php

[17] Maendeleo Foundation. "About us". Accessed 23 May 2016, http://maendeleofoundation.org/our-story/

[18] UNICEF. "Child labour". UNICEF, 11 June 2015. Accessed 20 May 2016, https://www.unicef.org/protection/57929_child_labour.html

[19] OER Africa. "Who we are". Accessed 20 August 2016, http://www.oerafrica.org/about-us/who-we-are

training new generations of ICT workers in developing countries should not be at all surprising. The minimum wage in China is currently about nine times lower than the legally prescribed minimum in the US [20,21], in India 20 times lower, and in Uganda 725 times.[22] The way the ICT giants conduct their business in China and India well illustrates the possible path of development for countries that seek their chance for development in this sector[23](Chakrabortty, 2013)(Garside and Arthur, 2013).

5 Conclusion

OER are promoted as being especially useful for those who want to learn but are unable to engage in traditional models of learning. To such potential users they offer emancipation and the liberation from oppression and poor living conditions (Knox, 2013, 7), despite the fact that the extremely poor living conditions and oppression around the world are often caused by the sponsors of the institutions that create OER[24,25](Washburn, 2010, 5)(Palomino, 2013)(McClenaghan, 2015)(Burgis, 2015).

Unequal education is the result, not the cause, of the basic economic inequality. The claim that education offers a way out of poverty is omnipresent

[20] WageIndicator. "Minimum wages". WageIndicator, 2016. Accessed 23 July 2016, http://www.wageindicator.org/main/salary/minimum-wage

[21] United States Department of Labor. "Minimum wage". United States Department of Labor. Accessed 23 July 2016, https://www.dol.gov/general/topic/wages/minimumwage

[22] WageIndicator. "Minimum wages". WageIndicator, 2016. Accessed 23 July 2016, http://www.wageindicator.org/main/salary/minimum-wage

[23] Al Jazeera English. "India: High tech mirage". Documentary programme, 00:00-00:35 (24:59). Al Jazeera English, 2014. Accessed 25 May 2016, http://www.aljazeera.com/programmes/peopleandpower/2014/03/india-high-tech-mirage-201431985356939298.html

[24] Multinational Monitor. "BP: A legacy of apartheid, pollution and exploitation". Multinational Monitor, 1992. Accessed 15 June 2016, http://www.multinationalmonitor.org/hyper/issues/1992/11/mm1192_11.html

[25] Al Jazeera English. "The Secret of the Seven Sisters". Documentary series in four parts with an accompanying description. Al Jazeera English, 26 April 2013. Accessed 28 July 2016, http://www.aljazeera.com/programmes/specialseries/2013/04/201344105231487582.html. YouTube video, 3:10:02. Posted by "Rebaz Zedbagi" 25 May 2014. Accessed 28 July 2016, https://www.youtube.com/watch?v=XtYOjMmEMeg

both in the promotion of OER and in the more comprehensive international campaigns and appeals. To the poorest, however, it offers only a hope in utter hopelessness. A boy in Haiti explains his school attendance this way: "Mum and Dad send me to school to learn so when I grow up I am able to live a good life and have a nice car, like everyone else", while his teacher stresses: "Without school there is no life. If you become president, it's because you went to school"[26].

Education systems that lack financial resources cannot be helped by the additional funding of the best-funded educational systems. Materials for teaching and learning are only part of the complex process of education, and their availability only partially contributes to the accessibility of education. Any contribution to increasing access to education deserves proportionate praise and support, but campaigns, appeals and propaganda material calling for the support of OER equate accessibility with applicability, thus creating a wrong idea about the potential contribution of OER to improving global education (Bates, 2011)(Crissinger, 2015). OER do not offer but presume adequate solutions for understaffing, infrastructural and other problems of educational systems, in both developed and developing countries, and consequently direct attention away from these problems rather than towards their solution. The bigger the problems a country faces, the bigger the damage it suffers.

References

Bates, Tony. "OERs: The good, the bad and the ugly". *Online Learning and Distance Education Resources (blog)* (6 February 2011). Accessed 14 July 2016, http://www.tonybates.ca/2011/02/06/oers-the-good-the-bad-and-the-ugly

Nikolai Bukharin and Evgenii Preobrazhensky. *The ABC of Communism.* Marxists Internet Archive, 1965

Burgis, Tom. "Big Oil's sleazy Africa secrets: How American companies and super-rich exploit natural resources". *Salon*, 6 April 2015. Accessed 15 June 2016, http://www.salon.com/2015/04/06/big_oils_sleazy_africa_secrets_how_american_companies_and_super_rich_exploit_natural_resources

[26] Al Jazeera English. "Is schooling for all a realistic goal?". TV debate, 19:10-19:45 (24:24). Al Jazeera English, 2014. Accessed 12 May 2016, http://www.aljazeera.com/programmes/insidestory/2014/04/schooling-all-realistic-goal-2014410151530585416.html

Cave, Kathryn. "South Africa: Why have all the rural tech projects failed?". *IDG Connect*, 21 June 2013. Accessed 25 May 2016, http://www.idgconnect.com/blog-abstract/2292/south-africa-why-have-all-rural-tech-projects-failed

Center for Educational Research and Inovation (CERI). *Giving Knowledge for Free: The Emergence of Open Educational Resources*. Paris: OECD, 2007. Accessed 15 June 2016, https://www.oecd.org/edu/ceri/38654317.pdf

Chakrabortty, Aditya. "The woman who nearly died making your iPad". *The Guardian* no. 25(25 August 2013). Accessed 25 May 2016, https://www.theguardian.com/commentisfree/2013/aug/05/woman-nearly-died-making-ipad

Crissinger, Sarah. "A critical take on OER practices: Interrogating commercialization, colonialism and content". *In the Library with the Lead Pipe (blog)* (21 October 2015). Accessed 14 July 2016, http://www.inthelibrarywiththeleadpipe.org/2015/a-critical-take-on-oer-practices-interrogating-commercialization-colonialism-and-content

Garside, Juliette and Charles Arthur. "Workers' rights 'flouted' at Apple's iPhone factory in China". *The Guardian* no. 5 (5 Sptember 2013). Accessed 25 May 2016, https://www.theguardian.com/technology/2013/sep/05/workers-rights-flouted-apple-iphone-plant

Hoel, Tore. "Adopting OER for less used languages: We need hard talk on tools and infrastructure!". *Nordic OER*, 11 April 2014, archived by Internet Archive Way Back Machine 11 March 2015. Accessed 11 August 2016, https://web.archive.org/web/20150311234754/http://nordicoer.org/adopting-oer-less-used-languages-need-hard-talk-tools-infrastructure

Hylén, Jan. "Mapping users and producers of open educational resources". UNESCO, November 2006. Accessed 25 May 2016, http://www.unesco.org/iiep/virtualuniversity/media/forum/iiep_oecd_oer_forum_note1.pdf

Knox, Jeremy. "Five critiques of the open educational resources movement". *Teaching in Higher Education* Vol. 18, no. 8(2013): 821–832. Accessed 2 August 2016, http://www.academia.edu/2651447/Knox_J._2013_Five_Critiques_of_the_Open_Educational_Resources_Movement._Teaching_in_Higher_Education

Knox, Richard. "5 years after Haiti's earthquake, where did the $13.5 billion go?". *NPR*, 12 January 2015. Accessed 20 May 2016. http:

//www.npr.org/sections/goatsandsoda/2015/01/12/376138864/5-years-after-haiti-s-earthquake-why-aren-t-things-better

Kwon, Dae-Bong. "Human capital and its measurement". In *3rd OECD World Forum on "Statistics, Knowledge and Policy": Charting Progress, Building Visions, Improving Life. Busan, South Korea, 27-30 October 2009*, 2009. Accessed 1 August 2016, http://www.oecd.org/site/progresskorea/44109779.pdf

Leyva, Rodolfo. "No Child Left Behind: A neoliberal repackaging of social darwinism". *Journal for Critical Education Policy Studies* Vol. 7, no. 1(2009): 365–381. Accessed 25 July 2016, http://www.jceps.com/wp-content/uploads/PDFs/07-1-15.pdf

Makarenko, Anton S. *Problems of Soviet School Education*. Moscow: Progess Publishers, 1965

McClenaghan, Meave. "Investigation: Top universities take £134m from fossil fuel giants despite divestment drive". *Greenpeace Energydesk*, 23 October 2015. Accessed 15 June 2016, http://energydesk.greenpeace.org/2015/10/23/data-top-universities-take-134m-from-fossil-fuel-giants-despite-divestment-drive

Meyer, Heinz-Dieter and Katie Zahedi. "Open letter to Andreas Schleicher". *Global Policy Journal* (5 May 2014). Accessed 20 May 2016, http://www.globalpolicyjournal.com/blog/05/05/2014/open-letter-andreas-schleicher-oecd-paris

Mungai, Martin. "12 challenges facing computer education in Kenyan schools". *ICT Works*, 12 September 2011. Accessed 29 July 2016. http://www.ictworks.org/2011/09/12/12-challenges-facing-computer-education-kenyan-schools

OECD. "How is equality in resource allocation related to student performance?". *PISA in Focus* no. 44(2014). Accessed 17 May 2016, https://www.oecd.org/pisa/pisaproducts/pisainfocus/pisa-in-focus-n44-(eng)-final.pdf

Palomino, Joaquin. "The university of private enterprise", *East Bay Express* (10 April 2013). Accessed 15 June 2016, http://www.eastbayexpress.com/oakland/the-university-of-private-enterprise/Content?oid=3518686

Pilkington, Ed. "UN makes first public admission of blame for Haiti cholera outbreak". *The Guardian* no. 18(18 August 2016). Accessed 20 May 2016, https://www.theguardian.com/world/2016/aug/18/un-public-admission-haiti-cholera-outbreak

Russell, Jon. "Apple, Microsoft, Samsung and other tech firms implicated in child labour report". *TechCrunch* (19 January 2016). Accessed 27

May 2016, https://techcrunch.com/2016/01/19/apple-microsoft-samsung-and-other-tech-firms-implicated-in-child-labor-report

Saunders, Daniel B. "Neoliberal ideology and public higher education in the United States". *Journal for Critical Education Policy Studies* Vol. 8, no. 1(2010): 41–77. Accessed 25 July 2016, http://www.jceps.com/wp-content/uploads/PDFs/08-1-02.pdf

Spellings, Margaret. "Building a globally competitive workforce". *US Chamber of Commerce Foundation*, 31 May 2012. Accessed 4 August 2016, https://www.uschamberfoundation.org/newsletter-article/building-globally-competitive-workforce

UNESCO. "The Dakar Framework for Action: Education for All: Meeting Our Collective Commitments". In *World Education Forum, Dakar, Senegal, 26–28 April*. UNESCO, 2000. Accessed 15 May 2016, http://unesdoc.unesco.org/images/0012/001211/121147e.pdf

UNESCO. "Teaching and Learning: Achieving Quality for All (Education for All Global Monitoring Report, 2013/4)". Techrep. UNESCO, 2014. Accessed 15 May 2016, http://en.unesco.org/gem-report/report/2014/teaching-and-learning-achieving-quality-all

Washburn, Jennifer. "Big Oil Goes to College: An Analysis of 10 Research Collaboration Contracts between Leading Energy Companies and Major U.S. Universities". Techrep. Center for American Progress, 2010. Accessed 15 June 2016, https://www.americanprogress.org/wp-content/uploads/issues/2010/10/pdf/big_oil_lf.pdf

Weibel, Catherine. "In Gaza, children go back to school after a devastating summer". UNICEF, 2014. Accessed 10 June 2016, http://www.unicef.org/education/oPt_75921.html

Zhao, Yong. "A true wake-up call for Arne Duncan". *Education in the Age of Globalization (blog)* (10 December 2010). Accessed 15 May 2016, http://zhaolearning.com/2010/12/10/a-true-wake-up-call-for-arne-duncan-the-real-reason-behind-chinese-students-top-pisa-performance

Zhao, Yong. "How does PISA put the world at risk (part 1): Romanticizing misery". *Education in the Age of Globalization (blog)* (9 March 2014). Accessed 15 May 2016, http://zhaolearning.com/2014/03/09/how-does-pisa-put-the-world-at-risk-part-1-romanticizing-misery

Zolnikov, Tara Rava. "From developed to developing country: Water and war in Iraq". *Johns Hopkins Water Magazine* (31 May 2013). Accessed 5 July 2016, http://water.jhu.edu/index.php/magazine/from-developed-to-developing-country-water-and-war-in-iraq

Classification of Terms on a Positive-Negative Feelings Polarity Scale Based on Emoticons

UDC 811.163.41'322.2
DOI 10.18485/infotheca.2017.17.1.4

ABSTRACT: The goal of this paper is to draw attention to the possibility of using emoticon-riddled text on the web in language-neutral sentiment analysis. It introduces several innovations in the existing framework of research and tests their effectiveness. It also presents a software tool especially made for that purpose, explains how it builds a database with sentimental value of terms and offers the user manual. Finally, it presents a software tool that tests the new database and gives some examples of the analysis of the obtained results.

KEYWORDS: data mining, information extraction, emotions, text on the web.

PAPER SUBMITTED: 24 January 2017
PAPER ACCEPTED: 25 March 2017

Mihailo Škorić
miks@tesla.rcub.bg.ac.rs

University of Belgrade

1 Introduction

When creating natural language understanding software, there are two widely accepted approaches:

- Software that does not have a deep understanding of the meaning of written text, but only the grammar of the language that text is written on, which enables wider application.
- Software that has a deeper understanding of the meaning of the text, often limited to one or a small number of areas. This type of software is predominantly used for text classification.

Systems based on sentiment analysis assign sentiment values to text using multiple parameters, where a greater number of parameters means much greater complexity.

Under an assumption that reducing all possible parameters onto polarity, there isn't a significant information loss and that the machine can decide between multiple choices using a single parameter, task is reduced to determination of what is positive and what is negative.

With a goal of making a more cost-effective intelligent system, it would not be a good idea to ignore any available resources and that could potentially be useful. The main idea of this research is to use data mining methods for retrieval of metadata – in shape of determiners, that users of social networks inadvertently use in their messages (in the form of emoticons or language-universal phrases) and assigning values of sentiment polarity to terms in which those determiners are located. As the determiners are language-independent, the system would be language-independent as well. If it turns out to be valid, this method could allow machine learning the usage of huge corpus of texts that are pre-labeled with determiners.

1.1 Review of their former similar studies

In 2005 a series of experiments with the classification of mood for internet blog posts was published, and it served as the basis for many future studies (Mishne, 2005). These experiments used records of terms that appear in posts for which the authors themselves claimed were written in and reflected a certain mood. Term indexes were made and their frequency in posts that are equated with one of nine moods was calculated. Those nine moods were: amusement, fatigue, happiness, joy, boredom, a sense of succes, drowsiness, satisfaction and excitement. New posts were then tested on the frequency of indexed terms in order to determine the mood of the author who wrote them. Results were compared with human assessment of the same posts and it was concluded that the machine assesses the mood of the author only slightly worse than a human.

At University of Tokyo in 2009 a study was published, in which the nine moods were analysed in text using complex finite automata that recognizes the grammatical structure of the text (Neviarouskaya et. al., 2009). That same year, researchers at Stanford University in California unveiled the new web posts analysis system that uses algorithms that are trained to recognize emoticons and classify moods of *Twitter* posts as either positive or negative (Go and Bhayani, 2009). The goal was to create a system for the classification of posts, so that consumers can explore the attitude of previous customers before buying a product. Several different algorithms for machine learning

were trained with eight emoticons (five positive and three negative), and the results showed accuracy above 80% for guessing the mood of the posts.

Researchers at the Hebrew University in Jerusalem in 2010 carried out another similar research of sentiment expressed in *Twitter* posts, taking into account 15 emoticons, and 50 tags,[1] as addition, which is their original cotribution (Davidov et. al., 2010). The algorithms that were trained using tags, also succeeded in recognizing the mood of the post.

In the above mentioned studies emoticons in text are treated as character strings of explicit meaning. A different approach was proposed in 2010 by researchers at Hokkaido University in Sapporo. Dissatisfied with existting databases of emoticons and their values, they primarily dealt with how to determine their values more precisely. The idea was to treat emoticons as structures composed of separate elements that represent the eyes and the mouth. Composite parts were processed serarately to calculate their value. When the database was being made emoticons were classified by ten possible feelings: anger, resentment, excitement, fear, affection, happiness, relief, shame, sadness and surprise (Ptaszynski et. al., 2010). This study was later expanded, and in 2011 a paper was published on the research in which emoticons were defined as parts of natural language, so it was suggested that their research is should be included in natural language research (Ptaszynski et. al., 2011).

1.2 Basic information about the experiment

Goal of the experiment is to test a new approach to text extraction using emoticon extraction in a new way, by combination of the following three ideas:

- Emoticons that will be used in the experiment will not be assigned discrete values such as positive or negative, but values on a scale from most positive to the most negative. This will be also reflect on the terms whose values will also be assigned on this scale.
- Only universal and language-neutral determiner strings will be used. Goal is to create a fully language-independent system that would greatly broaden the possible corpus.

[1] Users of *Twitter* platform have an option to additionally mark their posts with tags so that posts that talk about a certain topic can be found more easily using tag serach.

– Instead of separate messages, such as those on *Twitter* platform, messages that are a part of a conversation will be used. The goal of this is to test the ability of the determiner to influence not only the message in which it is located, but also the messages in the immediate vicinity.

An additional outcome of this experiment are two software tools, specifically designed to perform the experiment. These applications should enable future researchers to test similar ideas, and a detailed explanation of the software could help in the development of entirely new, better and more complete tool.

The main idea of this experiment is to prove that it is possible to:

– build an inverted index of terms in a language-neutral way using a corpus of texts that contain known determiners.
– automatically assign values to terms on positive-negative scale using those determiners, so that specific values reflect the attitude of the people using specific terms.

The experiment consists of two parts:

– creation of a database containing an inverted index of terms and their values, using software tool specifically made for this experiment.
– testing values of the terms from the database by comparing them with human assessments and by using the software testing tool that is specifically designed for databases that are the product of the first part of the experiment.

2 Preparation for the database creation

Database which is a product of this experiment brings terms (words that appear in conversations) and values they reflect their polarity (if any) in numerical form together in one place. The values are calculated by taking into account the proximity of the determiners to the observed term and the value of those determiners. The determiners can be either emoticons or phrases that appear in the conversation, which by nature are not of universal meaning and reflect a positive or negative attitude, replacing facial expressions and/or intonation in the written text.

The intensity values of a determiner directly affects the intensity value which he transfer onto the term. Also, the closer the determiner is to the term, the more its value affects the value of the term.

For the database formation to be successful and its final outcome satisfactory, three prerequisites should be met:

- collected corpus must be organized in a certain way;
- collected texts and messages must contain determiners that would help assign a value to a nearby term;
- determiners must have a predetermined value.

In the following chapters it will be explained in detail how the databases for this study were created, from the collection of the corpus to the export of completed database, which can then be used in several ways.

2.1 Collecting textual corpus

The basic idea was for the database to be based on a corpus of texts containing determiners which express positive or negative attitudes of interlocutors. Messages used for this experiment were gathered from chat histories files of *Facebook* users, for which a suitable preprocessing XLS transformation was prepared (explained in chapeter 3.1). Several volunteers downloaded these files from the official web page[2] and forwarded them to the author of this work in order for them to be included in the corpus used for the research.

Six ZIP files (from six users) sizes from 1.85MB to 167.09MB were collected. Together, those files contain 3,884 conversations of 2,019 different users and 1,843,826 messages that those users exchanged. The content of these messages is used in making the database for this research. With this, the first two aforementioned prerequisites were fulfilled.

2.2 Assigning values to the determiners

The values of determiners were obtained by a method of survey which was conducted on the internet platform *Google forms*, which people were able to access through hyperlinks published on social networks or passed from the other participants. Respondents were tasked with asigning values between 0 and 10 to a set of chosen emoticons and phrases, where 0 represented the highest intensity of negative mood, and 10 the highest intensity of positive mood. They were told to consider before rating how a determiner reflects

[2] Social network *Facebook* allows all its users to download a single ZIP file that contains all of their current multimedia files and chat history via settings on personal profile tab.

their feelings or mood when they send it or how they perceive it when they receive it in a message. Respondents also had the opportunity to propose additional determiners that they consider to be relevant and suggest its value.

During a period of 30 days, 389 survey participants assessed the 19 submitted and proposed additional 22 determiners, 9 of which were accepted[3]. The results differed from one determiner to another – somewhere the results were more balanced, while somewhere they were more speckled (Figure 1). After the survey ended, the value of each determiner was calculated as the arithmetic mean of a set of values that the respondents submitted (or proposed for 9 subsequently added determiners).

In order for values to better reflect what they represent – a point on a positive negative scale – they were mapped from set $(0, 10)$ onto set $(-1, 1)$, which was done by applying the formula:

$$x = (x - 5)/5 \tag{1}$$

Thus, the value of which corresponds to a negative mood of highest intensity became -1, while the value corresponding to the highest intensity of the positive mood became 1. Also a universal correction factor that multiplies all values was calculated. This factor was equal to the ration of the highest value 1 and tha value of a determiner with highest positive value 0.81, so that determiner *<3*, which, according to participants of the survey reflects a positive mood of the highest intensity, was assigned a maximum value of 1. Using this all determiners were assigned proper values and the last prerequisite for database forming was fulfilled.

3 Database construction

Software, developed specifically for this research was used to create the database. It was written in C# programming language and it can be ran on *Windows* platform, using any version of the operating system that runs to 64 bit. The interface is user-friendly and entirely in Serbian. The goal in designing this software was that any researcher who speaks Serbian can use it independently, create a new database or use it for a new research.

[3] A total of 143 users actually suggested a determiner, and the necessary number of proposers of a new determiner was 48, or more than one-third. If enough users suggested a determiner its value was calculated as the mean of all the proposed values and it was involved in the study.

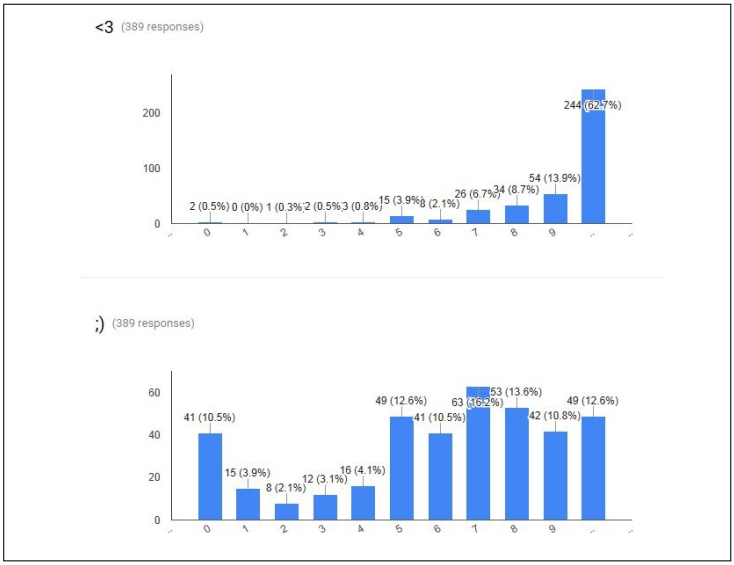


Figure 1. An example of the different level of agreement of users considering the value of determiner *<3* (higher agreement – over 60% of respondents agree on one answer) and determiner *;)* (lower agreement – five most commonly chosen responses in range of just 5.7%).

Software's task is to read the input file (in proper form), and trough seven steps, transform it into a database that contains terms found and their value on a positive-negative scale depending on the determiner that are in the vicinity of the respective term. In the following chapters the all steps of the database creation will be explained.

3.1 Preprocessing

Preprocessing is performed on the file that the user selects by pressing a button *Pronađi datoteku* (Open file). That opens a *Windows Explorer* catalogue in which user selects a file from the local computer in the usual way. The tool currently supports only files that contain chat history of *Facebook* social network, and the only file format available for selection is *htm*. The

determiner	meaning	value	determiner	meaning	value
:))	smile	0.56	:)	smile	0.26
:D	grin	0.91	:P	tongue out	0.35
:p	tongue out	0.24	xD	grin	0.59
:o	wonder	-0.22	:O	wonder	-0.29
:((	sad face	-0.86	:(	sad face	-0.66
:/	peeve	-0.48	:\	peeve	-0.48
:**	kisses	0.88	:*	kiss	0.80
hahaha	laugh	0.72	haha	laugh	0.15
<3	heart	1.00	;)	wink	0.26
:'(	cry	-0.74	lol	laugh	0.04
^^	joy	0.95	-.-	speechless	-0.45
:3	cat	0.94	*.*	glint	0.95
:S	peeve	-0.05	:'D	fall about	0.95
o.o	disbelief	-0.10	...	speechless	-0.18

Table 1. The final list of determiners and their assigned values.

user continues by pressing the button *Preprocesiraj i transformiši korak po korak*[4] (Figure 2).

Because the software uses XSL transformations, the basic precondition is for the input to be a well formed XML file. User picked file (Figure 3) is first purged of all the characters that may prevent it from being well-formed. It's done using the *Regex.Replace* function and regular expression *[\u0000-\u001F]*, which finds the first 32 characters from *ASCII* set and then replaces them with an empty string. To further ensure that the document is well-formed character *&* is replaced with a whitespace and a new root element is introduced.

Preprocessing is finished with an XSL transformation that transform a document into one that can be additionally processed (Figure 4). All nodes whose children do not contain user-exchanged messages are removed. Only node kept is * * (Figure 3), which contains the messages. These nodes are constructed, a new temporary file with a current document is saved and the preprocessing ends.

[4] Alternativly, he can use *Izvrši kopletnu transformaciju prema podrazumevanim podešavanjima* (Execute complete transformation using default settings) button, and start a complete transformation using default settings, or the last settings used in the step by step transformation.

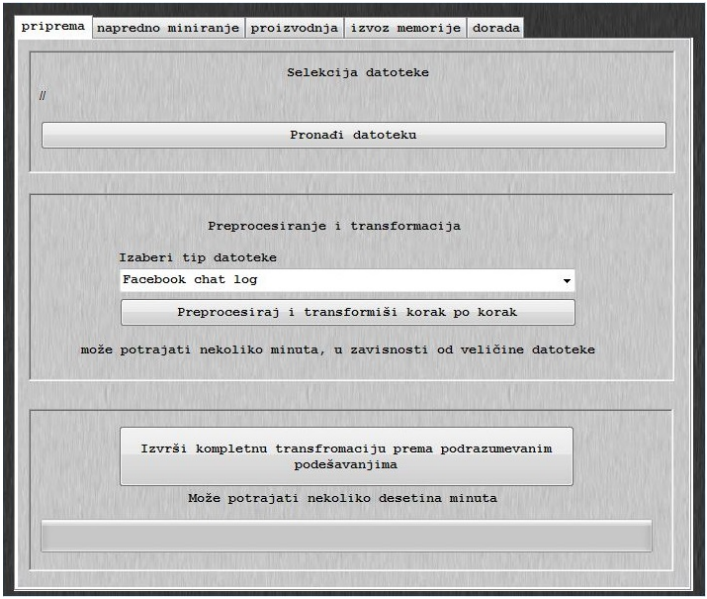


Figure 2. The appearance of user interface software for text mining.

3.2 The extraction of values from the document

Determiners extraction is performed in two steps. The first part of the extraction consists of picking determiners that will be searched for and the values that they stand for. Determiners and values can be set one by one, or this step can be skipped. In case this step is skipped, the software will automatically load the default determiners and their values (Table 1).

All default determiners and their values listed in the advanced text mining catalog (Figure 5). The values can be changed manually via the text field next to each term. If any determiner is undesirable, its value should be replaced with "/" character and it will not be used in text mining. In case the user wants to test a new determiner, he can use *dodaj emotikon* (add emoticon) button at the bottom of the page, after he fills the necessary fields – term string ant it's value. If the entered value of a determiner is not in the range of -1 to 1, entry will not be accepted and the user will receive a message about a failed new determiner addition. The maximum number of determiners that can be used is 36.

```
▼<html>
  ▼<head>
     <meta http-equiv="Content-Type" content="text/html; charset=UTF-8"/>
     <title>Korisnik 1 - Messages</title>
     <link rel="stylesheet" href="../html/style.css" type="text/css"/>
   </head>
  ▼<body>
    ▼<div class="nav">
       <img src="../photos/profile.jpg"/>
     ▶ <ul>...</ul>
     </div>
    ▼<div class="contents">
       <h1>Korisnik 1</h1>
      ▼<div>
       ▼<div class="thread">
          Korisnik 1, Korisnik 2
        ▼<div class="message">
         ▼<div class="message_header">
            <span class="user">Korisnik 1</span>
            <span class="meta">Tuesday, February 9, 2016 at 2:44pm UTC+01</span>
          </div>
         </div>
         <p>Tekst poruke</p>
        ▼<div class="message">
         ▼<div class="message_header">
            <span class="user">Korisnik 2</span>
            <span class="meta">Tuesday, February 9, 2016 at 2:39pm UTC+01</span>
          </div>
         </div>
         <p>Tekst poruke</p>
```

Figure 3. A fragment of the input file.

```
▼<root>
  ▼<conversation>
    ▼<messages>
      ▼<message>
          <sender>Korisnik 1</sender>
          <text>Tekst poruke</text>
       </message>
      ▼<message>
          <sender>Korisnik 2</sender>
          <text>Tekst poruke</text>
       </message>
```

Figure 4. Fragment of the input file after preprocessing.

Advanced users can modify text file *data/emoticons.txt* which contains all the default determiners and their values. At any time, determines can

Figure 5. Settings example for extracting value from the text.

be restored to default settings of the above file, or new settings can dubbed over existing ones. Both options are activated by pressing the appropriate buttons located on the bottom of the screen (Figure 5). In both cases, due to possible irreversible data loss, the user will need to confirm the process in an additional dialog box that will appear on the screen.

Once the user is satisfied with the settings he can switch to the second part of the extraction, which is done by pressing *ekstraktuj vrednosti iz teksta* (extract values from the text) button on the production tab (Figure 7). Before pressing this button the user can also, optionally, check the field *Koristi predodređene regularne izraze za poboljšanu pretragu* (Use predetermined regular expressions for improved search), which should lead to a greater number of determiner extractions in the text.

Execution begins with optional fields check. If the field is checked, document goes through several *Regex.Replace* functions that search for known deviations of several determiners. but first, automatic mapping of characters in XML are reversed, as this part of the transformation doesn't require

well-formed XML document. Entity reference *lt;*(character *&* that marks the begining of the entity reference was replaced with a whitespace during preprocessing step) is replaced with *<* character, so that the emoticon *<3* cound be found in the text. Entity reference *039* is replaced with *'* character, so that emoticons such as *:'(* could be found. String *:-* is replaced with character *:* , so that both characters with and without *nose* such as *:-D* or *:-(* could be mined. All the Cyrillic characters used in default determiners are transliterated into Latin in order to find the emoticons in the messages of users who use the Cyrillic alphabet (*:Д*, *xД*). All variants of emoticons *xD*, *:S* i *o.o* written in either upper or lowercase are converted into its basic format. Hyperlink beginnings *http://* and *https://* are replaced with a neutral character string *yy*, so that they wiil not be mistaken for *:/* emoticon. Finally, regular expressions *[a|h|A|H][a|h|A|H][h|H][a|A][h|H][a|h|A|H]* and *[h|H][a|A][h|H][a|A]* are used to find as many different examples of expression of laughter, and the expressions found are replaced with *hahaha* and *haha* respectively.

If this option is not checked, this step will be skipped and only determiners in its default form, which is listed on the advanced mining tab, will be found and processed.

The mere extraction involves loading selected terms and their values from advanced mining tab in two arrays, and then then processing them. Each of the elements of the first array, expressions array, pass through a *for each* loop in which each of its appearances in the text is replaced with *<emot value='x'/>* node, representing an empty XML node in which x is the value of the current determiner which is loaded from the second array, values array. If the value found is */* instruction will not be executed, and if the value of an expression is not between -1 and 1 the program will report an error and stop the execution.

After each of the determiners is replaced with a node whose attribute is its value (Figure 6), XML document becomes well-formed again and is ready for further processing. The user will receive a message that the extraction is completed, and he can move onto the next step.

3.3 Assigning values to segments of text

In the experiment segment of text or a sentence is equated with a sole message sent from one user to another. The basic idea is that the values of determiners found in the message reflect the value of the message, and the surrounding messages in some special cases. Depending on the message

```
▼<message>
   <sender>Korisnik 1</sender>
   <text>tekst :)</text>
 </message>

▼<message>
   <sender>Korisnik 1</sender>
   <text>teskt <emot value="0.26"></text>
 </message>
```

Figure 6. Message appearance before and after extraction of determiners.

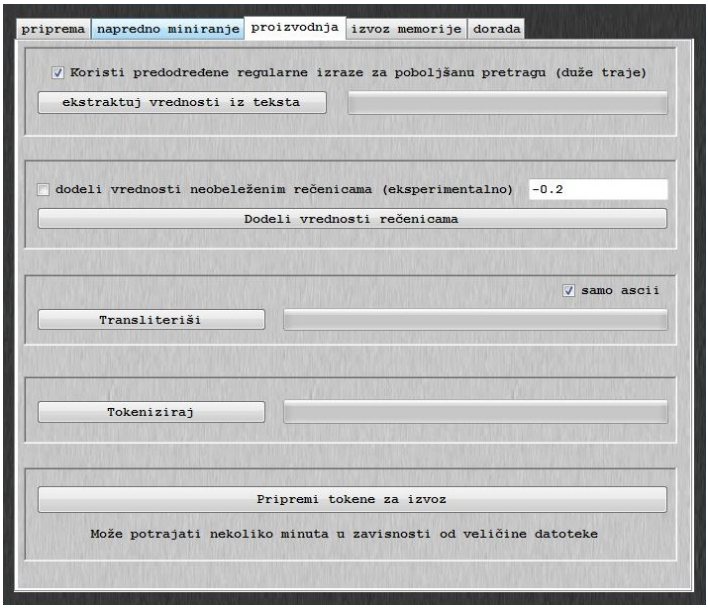


Figure 7. Step by step database production tab.

text content (beside the determiner[5]) and the content of the previous and following messages, there are three types of segments values allocation:

[5] In this case, all messages that contain less than 4 characters are treated as empty.

1. if the message contains text beside the determiner and the following message also contains text – only determiners found in the message affect its value Examples:
   *A: Happy birthday :**
   B: Thank you very much
   Determiner *:** will affect the segment value, while the message of person B has no effect on it, because it does not contain a determiner. *A: Happy birthday :**
   B: Thanks :)
   Determiner *:** will affect the segment value, while the message of person B has no effect on it, because the determiner in that message refers to the text in the same message

2. if the message contains text, but not a determiner, and the following message contains determiner but not text – determiner will refer to the previous message. Example:
   A: I missed the bus
   B: :(
   Determiner *:(* from the message of person B will refer to the message of person A, becouse the second message does not contain text, and its determiner must refer to previous message.

3. if the message contains both the determiner and the text, and the following message contains determiner but not text – determiners from both messages will refer to the message that contains text. Example:
   A: I missed the bus -.-
   B: :(
   Both determiners -.- and *:(* will refer to the message of person A, becouse the second message does not contain text, and its determiner must refer to previous message.

There is also a forth possibility – assign values to messages no determiners affect. This feature is based on the assumption that the absence of determiners also affects the sentiment (by intuition, negative one). The disadvantage of this option is that we can not be sure that the absence of determiners really bears (negative) meaning, and the assessment of value is also difficult task. An additional problem is that the message may contain a determiner that was not used in the text mining tab. Therefore, this possibility is optional, and the value it assigns is not predefined. If the user wants to assign

a certain value to emph unlabeled messages, he can do it by checking *dodeli vrednost neobeleženim rečenicama* button and by filling respective text box with value between −1 and 1 (figure 7).

```
▼<root>
  ▼<conversation>
    ▼<messages>
      ▼<message>
          <sender>Korisnik 1</sender>
          <text>tekst poruke 1 <emot value="0.26"><emot value="-0.26"></text>
       </message>
      ▼<message>
          <sender>Korisnik 2</sender>
          <text><emot value="0.15"></text>
       </message>
```

```
▼<root>
   <emotext value="0.15">tekst poruke 1</emotext>
```

Figure 8. The appearance of the document segment before and after the value was assigned – both messages contain determiners, but the second one does not contain text (case 3)

The mere assignment of values is done via XSLT transformation. First, for each *labeled* message the arithmetic mean value of all the determiners that refer to it is calculated, and this value is written into its new attribute node *value*, while the earlier made determiner nodes are removed. Then, if the optional box was checked, previously assigned messages are assigned value from the optional text box. The result of the transformation is an XML document containing root element and in it nodes of all messages that have an assigned value. Upon completion of the transformation the user will get the message that the assignment is completed and green light to move on.

3.4 Transliteration and reduction

These two steps are done together by pressing *transliteriši* (transliterate) button of the production tab. The purpose of this step is to get a clean out text for the creation of the database. In this case, the text is considered to be

any string that contains only alphanumerics. Optional checkbox *samo ascii*, serves to further limit the collection only to alphanumeric characters from the ASCII character set.

The reason this is optional is that it introduces nearly as many problems as it resolves. Assuming that not all social networks users use the full capacity of Unicode character set, but also *cropped Latin*[6], a problem occures where strings *istraživanje* (*research*) and *istazivanje* (*research*) will not be recognized as same, but different words. Converting all characters into ASCII set resolves this problem but introduces a new problem where words that would differ in Unicode set would be recognized as same – for example *španac* (*spainiard*) and *spanać* (*spinach*), would be recogniyed as same. This option is therefore best to be used depending on the situation.

Before proceeding further processing, the current XML document goes through XSL transformation in which all uppercase letters are changed to lowercase using *translate* function, so that words would be recognized no matter the capitalization of letters.

The second step of the text clean-up, is cleaning all the characters that are not letters or numbers, eliminating the possibility that the word contains any punctuation or other non-alphanumeric character. Again by using the *translate* XSL function all unwanted characters (all characters except for the small Latin letters and digits) are converted into whitespace. This process can result in incorrect words if the user accidentally entered an undesirable character during typing (*Mar?ia*) or if he deliberately used a special character (*M@ria*). In the first case, the string will become emph Mar ia, and in the second *M ria*.

3.5 Tokenization

User starts this step by pressing the *tokenizacija* (tokenize) button of the production tab (Figure 7). Tokenization is done by using whitespace as a delimiter, so a token is defined as any character string between the beginning of the message and the first whitespace, any character string between two adjacent whitespaces, or any charachter string between the last whitespace and the end of the message. It is carried out on an XML document level, with each token getting its own XML node. The first part of tokenization is done using *Regex.Replace* function in three steps:

[6] Latin letters without diacritics, such as: *c*, *z*, *s* instead of *č*, *ž*, *š*.

- First, the beginning of each message in an XML document is found. It is done by searching for the string *">*, which occurs only at the end of an opening tag of elements that contain an attribute, in this case each element that contains a single message. As the root element does not contain an attribute, it will not be found. After each of the found expressions string *<token>*, which marks the beginning of the token, is added.
- Then the end of each message is found, by searching the string *</emotext>* – which is the ending tag of each message node. After each of the found expressions string *</token>* is added, to get string *</token></emotext>*. With this step, each message in the XML document becomes one token: *<emotext value="x"><token> ... </token></emotext>* .
- The last step is to divide the interior of each message into tokens, using whitspaces. To avoid unwanted replacement within the opening tag *<emotext value="x">*, for whitespace between the name of the element and the name attribute, each *emotext value* string is replaced with *emotext_value*, and then, each whitespace in the document is replaced with *</token><token>* string. Character _ becomes whitesapce again and document structure becomes: *<emotext value="x"><token> ... </token><token> ... </token></emotext>* .

The second part of tokenization is performed using four additional XLS transformations that filter selected tokens and give them value:

- The first transformation removes all messages containing more than forty tokens, in order to avoid awarding the value to all words in, for example, a text pasted into chat, which is not an active messages in the conversation.
- Second transformation deletes all tokens containing less than two characters (assuming they do not carry any meaning, because they either represent a non-functional word or were created by breaking large character strings such as hyperlinks), and all the tokens that contain more than twenty characters (assuming that the vast majority of them is random).

After the second transformation, another operation is needed, introducing of an additional token without value to the last place in the document. This is done by inserting *<token>ERR0001</token>* string before the closing tag or root node in order to get *<token>ERR0001</token>*. The meaning of this operation will be explained in the following section.

– The third transformation is used to make tokens independent. Initially every token is assigned witha a value of a parent message which is allocated in a new attribute for each token, and then all messages tags are removed, leaving only the root node and in it token nodes with associated values. It also sortitra all tokens in alphabetical order in relation to their text content, and assignes each token with a new attribute *no*, whose value becomes the ordinal number of the token in the document, defined by *position* XSL function.

Figure 9. The appearance of a document fragment before and after tokenization.

– Fourth transformation does not contribute to the structure of the document, but only the speed of its processing. Name of each element *token* is shortened to *t*, and each attribute *value* to *v*, which greatly reduces the size of the file and accelerates the further processing of the document.

3.6 Creating index of tokens with values

This step is performed by pressing the *pripremi tokene za izvoz* (prepere tokens for export) button in the bottom area of prduction tab (Figure 7). Execution of this command is possible only after all the preceding steps of preparation were completed.

For index to serve its purpose, all terms in it receive two attributes: the mean of all values that refer to this token in the used corpus and the number its repetitions. Creation of the index is done by using six steps XLS transformation of an XML document that contains tokens:

– In the alphabetically arranged documet the first node, x, is found, and the first following node with different content, y, is found.
– Their ordinals are fetched from the *no* attributes. For x values is assigned to variable n, and for y to varable m.

– Text of the x node is inserted into a new XML document for inverted index, in element t (token).
– New element t is assigned with attribute c (count), which indicates the number of repetitions of this token in the database, and its value is equal to the difference of the ordinal numbers from the two nodes.

$$c = m - n \tag{2}$$

– New element t is assigned with attribute v (value), indicating the value of the token on a positive-negative scale, which is equal to the arithmetic mean value of attributes emphv for all tokens containing the same exact text.

$$v_x = \frac{\sum_{i=n}^{n+c-1} v_i}{c_x} \tag{3}$$

– Node y becomes the first node in the document and the procedure starts again.

This procedure assigns attributes to all the tokens except for the last one, *ERR0001* token, which was created during tokenization of the document and is used for marking the end of the document, and for calculating the attribute c for the previous sibling token.

t (term)	v (value)	no (ordinal number)
zdravko	-0.1	859231
zdravko	0.3	859232
zdravlje	0.26	859233

t (term)	v (ordinal number)	c (repetitions number)
zdravko	0.2	2
zdravlje	0.26	1

Table 2. Appearance of database fragment before and after the final transformation.

Having successfully carried out the conversion of a set of tokens with the values to the final inverted index of terms with its values (Table 2), the user will get the message about the successfull completion, and the index will be stored in a temporary file until his final export.

4 Database update and export

Pressing the *Izvezi u novu datoteku* (Export to new file) button of the export tab (Figure 10) opens *Windows Explorer* file saving catalogue. The format in which the file will be saved is XML and the user chooses the name and location. Content is copied from the temporary file created in the previous step, and it represents the final database.

In the event that the user has already processed some files, he also has *Dopuni postojeću bazu* (Update existing base) available (Figure 10). It allows the data to be drawn from the new corpus and update previously exported databases. In this case, pressing the appropriate button opens the *Windows Explorer* open file catalogue. The user needs to mark the file he wants to update, and updated database is created in four steps:

Figure 10. Appearance of export/update databases tab.

– The selected file is loaded and added to content of previously created index stored in a temporary file.

- String *</root><root>* is found and removed. This results in a document with one, instead of two root elements.
- XSL transformation is used re-sort all tokens in alphabetical order.
- As in this case there can't be more than two tokens with the same text, their unification is simpler. If token does not have a pair, it is copied, and if it does it gets new attribute values. Value attribute is calculated as the quotient of the sums of products of attributes v and c of both elements and the total number of their repetitions (2.4). Repetitions attribute is equals the total number of their repetitions (2.5).

$$v = \frac{v_1 * c_1 + v_2 * c_2}{c_1 + c_2} \tag{4}$$

$$c = c_1 + c_2 \tag{5}$$

  The new elements with the attributes are kept while the old ones deleted from the index, and thus only one copy of each token with a specific attribute remains.

The new document replaces the one that was marked in the open file catalog, so this is an option that should be used carefully. Upon completion, the user will receive a message about successfull completion.

If the user wants to manually review the newly created database, and does not feel comfortable in an XML environment, he can export it into CSV (comma separated values) format that can be viewed using *Microsoft Excel*, *Open Office Calc* or similar software for working with tables. He does this by pressing the last, fourth button on the export/update tab.

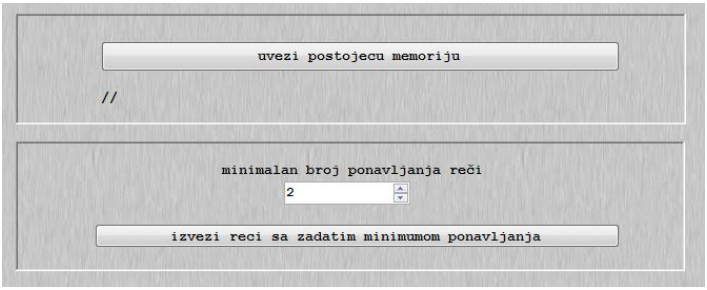


Figure 11. Appearance of database finishing tab.

The database can be plucked based on repetition number of the terms inside of it. This option is located on the fifth software tab, experimental database finishing tab. The user must first load the database by pressing *uvezi postojeću memoriju* (import existing database) button, and then chooses how many times the term bar needs to be repeated to be taken into account (Figure 11).

If the user, for example, choses number 2, by pressing the *izvezi reči sa zadatim minimumom ponavljanja* (export terms with specified minimum repetition) button, a new database will be formed without all the terms repeated less than 2 times (once), and will be saved in the desired place. User should use this option if he considers it necessary that the term appears a number of times before it is considered representative.

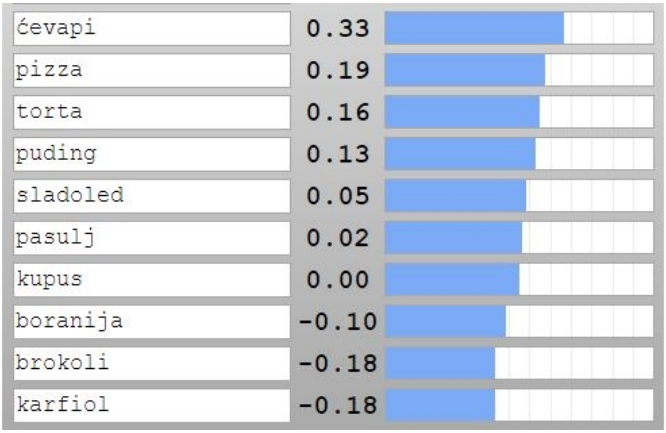


Figure 12. Example of value comparison of terms from the database using a web application – fast food, desserts, vegetables.

Once the user has finished creating the desired database and successfully exported them, they can be tested manually with the help of search option built into any software for working with text or XML documents. Also, the base in any of its forms can tested using software specially-made for this experiment (Figures 12 and 13).

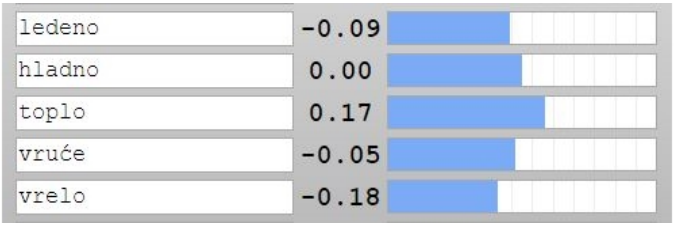


Figure 13. Example of value comparison of terms from the database using a web application – temperatures.

5 Conclusion

5.1 Review of the experiment outcome

Despite relatively small sample of 1,843,826 messages, from roughly estimated 500 billion messages belonging to the users of *Facebook* social network in Serbia, the experiment met its objectives. Based on the extracted only twenty-eight determiner strings and with the help of seven regular expressions that find their variations, and without prior knowledge of the language or its grammar, software formed a database of terms and their values on a positive-negative scale.

The system has been tested by several independent evaluators and based on their responses it can be concluded that the values are non-random and representative, hich implies that this method of terms classification is possible for any natural language using determiners, but it is still necessary to carry out a systematic evaluation.

The algorithms used are still far from perfect, and in there is plenty of room for improvement and progress considering them. Documented process of the extraction and transformation should contribute to future studies of the same or similar ideas. The next step can be extracting an expanded set of determiners, operation over an extended set of texts or, ideally, an expanded set of parameters that should be evaluated. Corpus expanded, for example, to other natural languages would greatly contribute to the volume of terms in the index, while the expanded set of parameters could add a new depth of understanding of the text by both the machines and the people.

5.2 Possible applications

This method of determiner extraction and generally similar researches can find a wide variety of applications divided into two groups.

Social and demographic research:

- Marketing research: exploration of current vs alternative approach to the marketing of products and services. This is the most common use of similar studies primarily for financial reasons, because using these companies can save money or get a new influx of goods.
- The public opinion research: what people like, what they do not like, what are their opinions about things or ideas that text refers to. It can be used in different ways in social and demographic research to quickly and efficiently collect large amounts of information.

Developing intelligent systems that work with information:

- Information retrieval: retrieval of specific information in the text, as well as finding information that can not be precisely defined. Classification of texts according to the mood that is expressed in them to help find the necessary information.
- Natural language understanding and analysis: understanding of written text and text queries, analysis of moods in the text, processing of digital linguistic resources such as automatic parallelization and automation of any operation that requires a deep understanding of the written text.
- Artificial intelligence: automated conversation in natural language and work with clients.

5.3 Possible deficiencies and improvements

- The problem of possible random responses during the survey:
  Create a system that rejects the member who answered at random, for example, using an introduction of additional questions that requiring specific responses. In that way careless users who do not read the questions are more likely to be identified.
- Systematic evaluation of the results:
  Conduct a detailed systematic evaluation in order to determine the credibility of the results.

References

Davidov, Dmitry, Oren Tsur, and Ari Rappoport. Enhanced sentiment learning using twitter hashtags and smileys. In *COLING '10 Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, 241–249. Association for Computational Linguistics, 2010.

Go, Alec, Lei Bhayani, and Richa nad Huang. Twitter sentiment classification using distant supervision. Processing, 2009.

Mishne, Gilad. Experiments with mood classification in blog posts. In *Proceedings of ACM SIGIR 2005 workshop on stylistic analysis of text for information access*, Vol. 19, 321–327. Citeseer, 2005.

Neviarouskaya, Alena, Helmut Prendinger, and Mitsuru Ishizuka. Compositionality principle in recognition of fine-grained emotions from text. In *Proceedings of the Third International ICWSM Conferenc*, 278–281. The AAAI Press, 2009.

Ptaszynski, Michael, Rafal Rzepka, Kenji Araki, and Yoshio Momouchi. Research on emoticons: Review of the field and proposal of research framework. In *The Seventeenth Annual Meeting of The Association for Natural Language Processing*, 1159–1162. The Association for Natural Language Processing, 2011.

Ptaszynski, Michael, Pawel Dybala, Radosalw Komuda, Rafal Rzepka, and Kenji Araki. Development of Emoticon Database for Affect Analysis in Japanese. In *Proceedings of the 4th International Symposium on Global COE Program of the knowledge Federation*, 203–204, 2010.

Authority Control in Serbia

UDC 025.342:006.44(497.11)
UDC 025.342:004.65(497.11)
DOI 10.18485/infotheca.2017.17.1.5

Ana Savić
ana.savic@nb.rs
National Library of Serbia

ABSTRACT: This paper describes the process of constructing Serbian authority database of personal names, its value for cataloguing, universal bibliographic control development and Serbian library community. It continues with a discussion about authority control in the world in the 21st century and describes authority control as a tool for library contribution in Linked Open Data agenda.

KEYWORDS: authority control, Serbian authority database of personal names, Linked Open Data

PAPER SUBMITTED: 13 September 2016
PAPER ACCEPTED: 15 December 2016

1 Introduction

During its long tradition, libraries and allied institutions were primarily oriented to collecting and cataloguing the results of intellectual or artistic work of individuals and institutions. The unambiguous determination of individuals and institutions associated with the described work and a description of the publication has been the focus of cataloguing. The aim was to locate the resource and to provide the access to the described resource (Angjeli et al., 2014). Identification of persons associated with resources in catalogues came from its focus on the choice of its name and form that will be the basis of the uniform heading together with a unique formulation of the uniform heading and aspirations to establish a syndetic structure[1] of catalogue. The Linked Open Data concept and the Semantic Web (top-down model in which the bibliographic records of electronic catalogues are based

[1] The syndetic structure comprises the system of "see" and "see also" cross references to other indexing terms.

on the MARC format), which fitted and suited the needs of 1990s users, were no longer sufficient to ensure the existence of the library catalogues which needed to become a part of the Web 2.0 reality. "The need to improve interoperability within the World Wide Web gave rise to the development of the Semantic Web, which in turn led to the appearance of many new ways to control and standardize the description of documents, solve problems surrounding diverse indexing systems, and improve the interoperability of records (SKOS, SIOC, Dublin Core, FOAF, etc.)(Леива-Медерос и др, 2015)." One of the ways of reshaping the catalogue in this respect is an intensive work on the creation and development of authority files and linking them with the bibliographic databases. Authority files projects have begun being developed within the universal bibliographic control in the early 1970s of the 20th century, when a recommendation that the national bibliographic agency "should maintain an authority control system for national names, personal and corporate, and uniform titles, in accordance with international guidelines ..."(Guidelines, 1979) was adopted. Accordingly, unambiguous identification of persons connected with the resource being described on the national level or the national personal name authority files for the newly created environment of the Semantic Web has become an extremely valuable tool to connect and exchange the information, in the first step of creating international files (for example, the VIAF[2] ISNI[3]), then a reliable source for use beyond the library community.

2 The implementation of authority control in the Cooperative Online Bibliographic System and Services (COBISS)

Cooperative Online Bibliographic System and Services COBISS are being developed since 1987 when the shared cataloging system in the former Yugoslavia was established. The Institute of Information Science (IZUM) from Maribor was responsible for the development of organizational solutions and software. In 1991 IZUM promoted the COBISS system as an upgrade of the shared cataloguing system. Due to social and political changes, COBISS system continued to develop as a Slovenian library and information system. When the libraries of emerging countries reestablished cooperation on the

[2] VIAF (Virtual International Authority File), https://viaf.org/
[3] ISNI (International Standard Name Identifier), http://www.isni.org/

territory of the former Yugoslavia in 2003, an agreement was signed on establishing of the COBISS.Net network. National databases of five countries (Slovenia, Serbia, Bosnia and Herzegovina, Macedonia and Montenegro) were connected.

During the development of the system as a national Slovenian system, where the focus was to connect local databases in union COBIB database, consistency, uniformity and quality of the local bibliographic databases and union bibliographic databases were held "by authority control for authors' personal names, by duplicates control, by COMARC software controls, by record editing, by global code lists for all standardized data (e.g. countries, languages, UDC), by local code lists to provide uniformity of data within a library (e.g. locations, internal designations), by automatic counters (e.g. accession numbers, numbering in call numbers), by unique identification control of serials, etc."[4] However, the authority control project has not evolved alongside library and information system.

The first analyzes were carried out in 1994 and the first prototype version of the authority file was made in 1996 when the cooperative bibliographic database contained approximately 1 million bibliographic records. Assessing the quality of authority file as unsatisfactory, IZUM continued to upgrade the authority file prototype and made three versions. CONOR V 3.0 was promoted in 1998 and represented a solid basis for further development of the authority control project. It was followed by further consideration of the editing records, work organization, preparing manuals for authority data, proposed ways of linking bibliographic and authority records and proposals to upgrade the software. In collaboration with the National and University Library (NUK), the testing of V 4.0 draft authority file CONOR began in Ljubljana, year 2000. The editing and entering of new records onto the next version started in 2001 and by 2003, the authority file was included in the shared cataloguing system in Slovenia. This version was based on a number of top quality bibliographic records that were the basic for the software to create authority records. Firstly, the authority records for personal names for Slovenian authors were created. The plan was that the entire machine created authority records should be edited before including authority file in the shared cataloguing system.

[4] Narodna biblioteka Srbije, „*Platforma COBISS – Kooperativni online bibliografski sistem i servisi*" http://www.cobiss.net/platforma_cobiss-SR.htm (Accessed 22. 07. 2016)

The implementation of authority control implied upgrading software for sharing cataloguing. In January 2003, the software COBISS2/Cataloguing V6.0 for cataloguing with authority control has already been developed and training of cataloguers started. Software testing continued and a new edition COMARC/A based on the 1998 edition was issued. It was planned that phase two of the project should also include corporate names.

Today, the personal name authority file of authors and corporate bodies in Slovenia contains over 900,000 authority records (Seljak, 2004).[5]

The implementation of authority control in the shared cataloguing system in Slovenia was a significant step towards linking data and providing access to data, enriching the quality of national library information system and strengthening partnership with similar international systems. The project has shown that the implementation of authority control in bibliographic databases originally designed without authority control was an extremely complex task. It would have been ideal if the development of both, the projects of shared cataloguing and authority control, could have been arranged in parallel. This concept is more economical, requires less time and is a great professional challenge for cataloguers.

3 Authority control in Serbia

Based on the proposal project of Ljiljana Kovačević and Gordana Popović-Mazić and on the initiative of a Virtual Library of Serbia (VBS), in January 2006 a document "*Guidelines for the Preparation of Bibliographic Records for the Program for the Creation of Personal names Authority Records*" had been prepared. Naturally, once the proposal was adopted, the realization of the project was started in collaboration with IZUM, while the National Library of Serbia, the Matica Srpska Library and the University Library "Svetozar Marković" were responsible for the project with the participation of all libraries in Serbia who maintain and develop their catalogues under the COBISS platform. It should be noted that the COBISS system does not gather Serbian library network as a whole and the implementation of the authority control is not related directly and absolutely to all local catalogues, but in strategic and technological terms this solution was optimal. The COBISS system is the only one who possessed the potential to be

[5] http://cobiss4.izum.si/scripts/cobiss?ukaz=BASP&bno=509&id=1955089254138144

the base of the construction and expansion of a modern national authority control in Serbia, which is assumed to be likely to give results.

The emphasis was on name heading since it was decided to create the personal name authority file first, with further expansion to the scaling of all other access points. In late 2008 the VBS team for the catalogue editing was considering an issue of so-called "parallel headings" and took the decision to implement this solution in the Serbian authority file. The first phase of the implementation of authority control in Serbia was related to the preparation and marking of bibliographic records, and making set of records that will serve as a basis from which the authority records will be extracted. In order for this phase of the project to be put in place, cataloguers of the National Library of Serbia, the Matica Srpska Library and the University Library "Svetozar Marković" were engaged. "Parallel headings" are the specifics of Serbian personal names authority file and it refers to a parallel use of the Cyrillic and Latin script in the creation of the uniform heading for the personal name. The national libraries of other countries are in a similar situation, for example, the National Library and Archives of Canada, which has a so-called bilingual cataloguing policy. "Parallel headings" are created for persons whose names appear in the etymological form, both in English and French.

The focus was on the primary authorship or co-authorship i.e. data entered fields 700 and 701 COMARC/B format. Preference is given to Serbian authors and foreign authors whose works have been translated into Serbian. Selected bibliographic records are supposed to be accurate and uniform heading created in accordance with applicable cataloging rules.[6] Slovenian model was used as a base, however Serbia needed to develop its own model. This was also necessary in view of the mentioned specificity which is reflected in the parallel headings. The legacy of the card alphabetical library catalogues in Serbia is reflected in the shared electronic catalogue COBIB.SR and the use the Cyrillic and Latin script in the selection and creation of the personal names uniform heading. Therefore, it was essential to establish the principle of "parallel heading" which encompassed the Cyrillic and Latin form of the name of a person,

[6] *Rules and Guide for the Creation of the Alphabet Catalog Part 1,2* is still valid cataloguing code for cataloguing in Serbia and adopted as the national cataloguing code. With certain modifications concerning the rules for the creation of personal names uniform heading of authors whose etymological form in foreign Cyrillic, standardized forms of names of persons is formed in accordance with the first part of the mentioned cataloguing code.

and also pairing phonetic and etymological form of the name of the author.

Example 1

```
200 1<7>cb  Cyrillic - Serbian <a>Младићевић<b>Жељко<f> 1977-
200 1<7>ba  Latin <a>Mladićević<b>Željko<f> 1977-
```

Example 2

```
 200 1 <7> ca - Cyrillic script unspecified <a>Алексеев<b>
 ГлебВасильевич <f> 1892-1938

200 1 <7>ba - Latin<a> Alekseev <b>GlebVasil´evič<f> 1892-1938

400 1 <7>cb - Cyrillic - Serbian <9>srp - Serbian <a>Алексејев<b>
ГљебВасиљевич<f> 1892-1938
```

Example 3

```
200 1 <7>cb - Cyrillic - Serbian <a>Томас<b>Скарлет<f> 1972-
200 1 <7>ba - Latin<a> Thomas <b> Scarlett <f> 1972-
```

Example 1 illustrates the "parallel heading" formed for Serbian authors. It consists of the Cyrillic and the Latin form of the personal name. For those authors whose work is created in a language other than Serbian, and etymological forms of their names are written in Cyrillic/Latin other than Serbian alphabet, "parallel headings" make etymological form of the Cyrillic and transliterates it into Latin forms (if the etymological form of the name is in Cyrillic), however not Serbian (Example 2) or phonetically, the Serbian Cyrillic and Latin etymological form (Example 3).

This phase of the project took longer than planned; it was completed in 2010 and resulted in a corpus of over 54,000 bibliographic records. We used those records as basics for initial personal names authority file CONOR.SR which was installed in 2013.

As a model for the implementation of authority control we used the Slovenian project and the experience of Slovenian colleagues, with modifications. The first step was to install the software COBISS3/Cataloguing and the transition to a new technological platform. There was organized training for chief librarians in editing authority records and establishing the form and structure of parallel headings. Within COBISS3, during 2013, all assumptions have been satisfied for the start of editing of initially automatically created authority records.

Based on the experiences of Slovenian colleagues, we decided to edit all authority records before linking authority file and bibliographic database. Slovenian colleagues were automatically connected to authority file and bibliographic database without editing authority records, which was neither efficient nor economical and it did not make cataloguing process better or easier. This situation imposes the need for simultaneous editing of the authority records and all the bibliographic records linked with them.

On the operative level, a team of six librarians (two librarians from the National Library of Serbia, two from Matica Srpska Library and two from the University Library "Svetozar Marković") was formed, which is the highest level in the organization of the further realization of the project. Their main tasks were: creating a work plan, establishing the principles of senior librarians' team, planning of duties and writing guidelines for records editing.

The editing of the initial personal name authority file began in 2015 with only 6 chief catalogers engaged. As the highly complex task was not progressing as planned, the team size gradually increased, firstly the Matica Srpska Library engaged more cataloguers, then the University Library "Svetozar Marković" and finally the National Library of Serbia. There are 23 catalogers currently engaged and 8533 authority records[7] have been edited. It is planned to expand the teams and, over time, work efficiency will be increased. *Guidelines for the Preparatory Authority File CONOR.SR Editing* are completed.

The aims of the personal name authority file project is primarily ensuring quality control and standardize the information which would facilitate the work of cataloguers and classifiers, and enable users to search and use the resources efficiently. This can only be achieved by consistent use of preferred forms of terms, concepts and entities. The implementation of authority control increases the presence of a library on a global network and a more efficient use of their resources. In addition, the aim was also to further international exchange of information and improve cooperation with other libraries and allied institutions. It should be mentioned that in Serbia there is still a number of libraries that will indirectly benefit from the results of creating authority file. Many libraries maintain their electronic catalogues within the systems that do not support bibliographic data sharing with libraries that work within COBISS platform. Authority files should also be the source of lexical and bio-bibliographical data, and as such could have a widespread use. Direct exchange of data is the widest framework of authority

[7] Data relating to 1. 08. 2016.

control, but we should not negate the importance of using authority files as a data source.

Personal name authority file CONOR is based on international standards[8] and the tendency is to expand the project to other access points of bibliographic records, while the long-term goal is, of course, participation in a virtual international authority file.

4 Authority control in 21st century

The growing trend of opening data, providing access to data and their use and exchange refers to many national and international systems and libraries are invited to engage in this agenda. The first step is the transition to Linked Open Data communication. Broadly, libraries are adopting new ways of achieving one of its primary goals – dissemination of knowledge. This was possible to achieve only with the new data structure, adapted to modern ways of communication. And this was not a small step. According to D. Hillmann et al., the standards related to the cataloguing have been designed for users and referred to the printed text documents or presented in a document format on the Internet (Hillmann et al., 2010). The focus was on libraries adjustment towards interaction between computer-computer, but in a contrasting way to the one we knew. This also includes interaction between the library community and non-library community, where data exchange was becoming necessary. The Semantic Web and the Linked Open Data has become a new reality. Libraries have something to offer to the new platform. Controlled terms lists, classification schemes, thesauri, subject headings and authority files, as library artefacts, are suitable for structuring so as to be adequate for the information exchange also beyond the library community.

By adopting the RDF model (Resource Description Framework), the basic framework for defining, use and exchange of metadata that increased interoperability of data, libraries take a more active role. They are offering their resources to the Web users and take part in the Semantic Web. It

[8] The most important documents are: *GARE – Guidelines for Authority and Reference Entries*, edited by IFLA on 1984; *GSARE – Guidelines for Subject Authority and Reference Entries*, edited by IFLA on 1993 and *UNIMARC/Authorities* also edited by IFLA on 1991. Later, tree more documents were published: *FRBR – Functional Requirements for Bibliographic Records* in 1998; *FRAD–Functional Requirements for Authority Data* (2009) and *FRANAR – Functional Requirement and Numbering of Authority Records* (2009).

was the only choice. Otherwise libraries would be ignored in the world of information. According to G. Byrne and L. Goddard, by developing RDA (Resource Description and Access), FRBR (Functional Requirements for Bibliographic Records) and FRAD (Functional Requirements for Authority Data), the libraries have embraced the change (Byrne and Goddard, 2010). We agree with the same authors that the RDA has also raised a number of questions and has distanced librarians towards the new concept, especially among cataloguers. The authority files were the first serious advantage for participating of libraries in the integration of information. It was the way to discover library collections in machine-readable form. Online public catalogues have been satisfying this need; however, the new requirement is to make the information available and present it in a machine-readable form. The library information should be present on the Internet in a standardized form that can be enriched and easily connected with other information. In this context, the international standard ISNI uniquely and unambiguously identifies the name of the creator and the corporate bodies since 2006, and there are currently 9.12 million names. ISNI is connected with the VIAF. VIAF is one of the first attempts of linking authority records of different libraries using advanced technologies and the aim was to create an online public service and to improve authority control at national and international level. Then, the British Library released the British National Bibliography in 2011 as Linked Open Data open-related data. It included local URI in one field of authority record, and similarly did The Library of Congress, as well as The French National Library, which assigns permanent authority records identifier derived from the identification number of the record. Such projects have also been launched by Spanish, German and Hungarian libraries. These are all examples of an increasing number of libraries present on the Internet and library value for its traditional users and the Web community. Authority files have the characteristics of bio-bibliographic lexicons. They provide information about the name of the person associated with some intellectual or artistic content, including alternative forms of names and forms of names in other languages, data on the year of birth and/or death, other biographical data and extract from a bibliography, which is illustrated by Example 4.

Example 4

```
001 <a>c - corrected or revised record  <b>x -
authority entry record <c>a - personal name entry
```

```
100 <b>a - established  <c>srp - Serbian
<g>cb - Cyrillic - Serbian

101 <a>srp - Serbian

102 <a>srb - Serbia  <b>cs - Central Serbia  <a> bih -
Bosnia and Herzegovina  <b>rs - Republika Srpska Република Српска

106 <a>0 - may be used as subject access point 120
<b>a - differentiated personal name <a>a - female

152 <a>PPIAK - Pravilnik i priručnik za izradbu abecednih kataloga
19011 <a> 1951 <b> 01 <c> 31
200  1 <7>cb - Cyrillic - Serbian <a> Радуловић <b>
Здравка <f> 1951- <r> 04130
200  1 <7>ba - Latin <a>Radulović <b>Zdravka <f>1951- <r>04130

340 <a> Bibliographer adviser. She graduated from the
Department for Comparative Literature and Library Science
at Faculty of Philosophy, University of Sarajevo.  In the
same Department, she defended her master's thesis 1998.
She worked at the Department for Library Science at
Faculty of Philosophy, University of Sarajevo,
then in Bibliographic Department of Central National Library
of Montenegro „Djurdje Crnojevic“. Since 1997 she has worked
in National Library of Serbia and in period 2005-2015 she
was head of Bibliographic Department. Also, she was a laureate
of award "Dušan Panković“ for outstanding achievements
in the field of bibliography “.

810 <a> Библиографија часописа "Библиотекар" :
(1948-1997) / Здравка Радуловић, Долорес Калођера-Петровић

810 <a>Evropske integracije [Elektronski izvor] :
bibliografija monografskih publikacija i članaka u
serijskim publikacijama : 1995-2008 /
[izrada bibliografije Zdravka Radulović]. - 2009
```

```
810 <a>Библиографија часописа "Дабро-босански Источник" :
(1887-1911) / Здравка Радуловић. - 2010
```

The importance of authority files reflects in improvement of library activities and their impact on the entire information community. Authority files improve library services, expand the library target groups and thus the library takes an active role in the information environment. The second aspect of importance goes beyond the borders of a service and refers to the general exchange of information and modern communication model.

5 Conclusion

The Linked Open Data concept, applied to the library as a public institution, means that the data they create and store are publicly available, information use is free and can be reused. Realization of this concept has not been possible by maintaining bibliographic databases. They are relational databases, and it was necessary to increase the interoperability of data and make them available via the Web. This is achieved by including authority files into existing electronic catalogues. The data structured that way have become part of the Linked Open Data concept. That concept has strengthened the library presence on the Internet, expanded library users group and it has been the answer to the modern age which is an "era of global librarianship, based on cooperation and exchange of resources and information and partnership at all levels." (Билбија, 2015). Pieces of information are combined and the dissemination of information in a new way builds a partnership with users and provides better library services. The basic principle of universal bibliographic control is the joint use of information. Modern technologies offer new ways of achieving this principle. The authority files ensure relevance of the data globally, internationally and outside the library community. Also, the exchange of data between technologies is enabled. This is the goal of the Serbian community library, too. The overall objective is the accumulation of knowledge and its systematization and open, shared use of all human knowledge.

References

Angjeli, Anila and Andrew Mac Ewan and Vincent Boulet. "ISNI and VIAF – Transforming ways of trustfully consolidating identities". Paper presented at: *IFLA WLIC 2014 – Lyon – Libraries, Citizens, Societies: Confluence for Knowledge in Session 86 – Cataloguing with Bibliography, Classification & Indexing and UNIMARC strategic Programme* (2014). In: *IFLA WLIC 2014, 16-22 August 2014, Lyon, France*. Accessed 22.7.2016, library.ifla.org/985/1/086-angjeli-en.pdf.

Byrne, Gillian and Lisa Goddard. "The strongest link: Libraries and linked data". *D-Lib Magazine* Vol. 16 no. 11/12 (2010). Accessed 7.8.2016, http://www.dlib.org/dlib/november10/byrne/11byrne.html.

Hillmann, Diane, Karen Coyle, Jon Phipps and Gordon Dunsire. "RDA vocabularies: process, outcome, use". *D-Lib magazine* Vol. 16 no. 1/2 (2010)

Seljak, Martaetal. "Vzpostavitev normativne kontrole v knjižničnem informacijskem sistemu COBISS.SI, Slovenija". *Organizacija znanja* Vol. 9 no. 2 (2004): 1–16. doi:10.3359/oz0402037.

UBC, IFLA International Office for and UNESCO. "Guidelines for the National Bibliographic Agency and the National Bibliography". United Nations Educational, Scientific and Cultural Organization, 1979. Accessed 22.7.2016, nesdoc.unesco.org/images/0004/000486/048658eo.pdf.

Билбија, Биљана. *Основе библиотекарства 2.* измијењено и допуњено изд. Бањалука: Народна и универзитетска библиотека Републике Српске, 2015.

Леива-Медерос, Амед и др. "AUTHORIS – инструмент за нормативну контролу на семантичком вебу". *Гласник Народне библиотеке Србије 2014/2015* (2015): 35–53

Exhibition "Digital Cyrillics"

Milena Kostić
mkostic@unilib.rs

Vesna Vuksan
vuksan@unilib.rs

Zoran Bajin
bajin@unilib.bg.ac.rs

University Library "Svetozar Marković"

PAPER SUBMITTED: 15 July 2016
PAPER ACCEPTED:10 December 2016

On the occasion of celebrating 90th anniversary and marking the Cyril and Methodius Saints Day also known as the Day of Slavic Literacy and Culture, the University Library "Svetozar Marković" organized "Digital Cyrillics" exhibition accompanied by a virtual and web exhibition which lasted from May 24 to July 1 2016. The authors of the exhibition were librarians from the University Library "Svetozar Marković" Dr. Zoran Bajin, Vesna Vuksan and Milena Kostić. At the grand opening at the Carnegie Hall of the Library on May 24th 2016, the audience was addressed by Prof. Dr. Nada Kovačevic, Vice-Rector of the University of Belgrade, Prof. Dr. Aleksandar Jerkov, Head of the University Library "Svetozar Marković", Mrs. Tamara Butigan Vučaj, Deputy Head of the National Library of Serbia, Dr. Adam Sofronijević, Deputy Head of the University Library "Svetozar Marković" and via Skype by Prof. Dr. Günter Mühlberger from the University of Innsbruck and Mr. Claus Gravenhorst, Director of Strategic Initiatives at CCS Content Conversion Specialists GmbH[1] in Hamburg. The recording of the opening of the exhibition is available on the following link: http://media.rcub.bg.ac.rs/?p=5259.

1 Exhibition at the Library

On the exhibition at the Library the selection of books about the work and influence of Cyril and Methodius, the translations of works of Slavic literatures into Serbian, published from the mid 19th century to the beginning of the First World War were presented. These translations have revived

[1] http://content-conversion.com/

cultural links between Serbs and other Slavic people and languages, whose unity was strengthened by Cyril and Methodius 1000 years ago. Digital card catalog was also presented at the exhibition. This catalog is available on the University Library website: http://www.unilib.rs/pretraga/katalog/.

2 Virtual Exhibition

Trying to keep up with the new technological trends in librarianship, and bearing in mind the protection of rare books, the University Library bought Magic Box with the support of the Ministry of Culture and Information. This is a cabinet display which provides virtual, yet very real experience of leafing through the digitized and protected library collections. Magic Box, located on the ground floor, was presented to the audience for the first time at the exhibition opening. The Gospel, which is an exquisite representative of old manuscripts written in the script created for the Slavic people by Cyril and Methodius, was displayed in Magic Box on the day of the opening.

The digitized Gospel can be leafed through on the interactive display and also representative copies from the collections of the University Library: old Cyrillic manuscripts, historical newspapers, photographs and 3D objects.

The presentation prepared by Mr. Claus Gravenhorst from CCS Content Conversion Specialists GmbH in Hamburg, which manufactures Magic Box is available on the following link http://www.unilib.rs/pretraga/katalog/. DocWorks software used for improving work with digital documents by creating METS/ALTO files is described in it. The University Library acquired the software in 2016 within the project "New Digitization Horizons in Serbia" supported by the Ministry of Culture and Information. Magic Box, a virtual cabinet display which displays METS/ALTO files, is also depicted in the presentation.

3 Web Exhibition

The web exhibition gives access to the digitized Cyrillic manuscripts and books about work and importance of Cyril and Methodius and translations of works of Slavic literatures into Serbian published between the mid 19th century and the beginning of the First World War. The collection is available on the University Library website on the following link: http://www.unilib.rs/sadrzaji/virtuelne-izlozbe/. The description of the digitized materials follows.

3.1 Cyrillic Manuscripts

The University Library "Svetozar Markovic" exhibited seven representative manuscripts from its Cyrillic manuscripts collection in virtual space. Being of unique artistic value, the Gospel is exhibited also in the digital device. Time and place of the origin of the Gospel, which stands out with the calligraphic skills of the scriptor and illustrations of four gospelers, have not been determined. There is a testimony on the manuscript covers that in 1685 the book was owned by a priestmonk Pahomije from the Crna Reka Monastery. In addition to the Gospel, the following manuscripts were also exhibited virtually: Panegyric of the scriptor Venijamin from the Monastery Holy Trinity near Pljevlja from 1595; Monthly Church Book for January with the Service to Saint Sava written by Teodosije at the end of the 16th or beginning of the th century; Psalter written in 1652 with appendices from the th century with an inscription of Dositej Obradovic; "Služabnik" from the 3rd quarter of the 16th century which contains a Short Service to the Saint Simeon; Transcript of Dušan's Code from Baranja created in the first half of the 16th century with a short version of Syntagma Canonum by Matthew Blastares and Codex Justinianus; and a collection of ancient narrations accompanied by Indian deductive stories – Ćorović 31.

3.2 Books on Cyril and Methodius

Books about life and work of Cyril and Methodius were physically exhibited at the Library. Three titles were selected for digitization: Serbian translation of the book about Cyril and Methodius by Theophylact of Ohrid published in Budim in 1823; A thousand-year jubilee of St. Cyril and Methodius with Dimitrije Matić's speech and a short Methodius' biography by Janko Šafarik published in Belgrade in 1863; and an illustrated Geschichte der Heiligen Slaven-Apostel Cyrill und Method by a Czech priest and historian Jan Bily published in Prague in 1863.

3.3 Translations of works of Slavic literatures

The majority of publications displayed in the physical space were translations of works of Slavic literatures created between the middle of the 19th century and the beginning of the First World War. Seven publications from this collection were selected for digitization. Translations of Russian literature: The Captain's Daughter by Alexander Pushkin published in

1849, The Overcoat by Nikolai Gogol published in Mostar in 1902 and Khadzhi-Murat, the last fiction work by Leo Tolstoy whose translation into Serbian appeared on the eve of the war. Translations of Polish literature: Cossack Revenge by Mihail Tchaikovsky published in Belgrade in 1854 and Tales by Henryk Sienkiewicz published in Novi Sad in 1887. Translations of Czech literature: The Village Novel by Karolina Světlá published in Belgrade in 1879. Translations of Bulgarian literature: Baja–Ganje by Aleks Konstantinov from 1907.

The event was covered in the media. There was news on TV, radio and web pages. Overall there were 26 broadcasts. The majority of web portals published the following news[2] written by the National Television of Serbia. Some news focused on the description of the exhibition "Digital Cyrillics" while some focused on the description of Magic Box.

[2] http://www.rts.rs/page/stories/ci/story/124/drustvo/2328542/vremeplov-kroz-istoriju-uz-digitalnu-vitrinu.html

Review of the 2nd KEYSTONE Training Summer School on Keyword Search in Big Linked Data

Marija Radojičić
University of Belgrade
Faculty of Mining and Geology

Ranka Stanković
University of Belgrade
Faculty of Mining and Geology

Sebastijan Kaplar
University of Novi Sad
Faculty of Technical Sciences

PAPER SUBMITTED: 9 April 2017.
PAPER ACCEPTED: 8 May 2017.

Within the framework of COST action[1] IC1302 with acronym KEYSTONE (semantic KEYword-based Search on sTructured data sOurcEs)[2], the 2nd KEYSTONE Training School named "Keyword search in Big Linked Data" was held from 18th to 22th of July 2016 in Santiago de Compostela in Spain. The aim of the school was to present current topics in the fast-growing area related to Keyword Search in Big Linked Data. The School was intended mainly for graduate and post-graduate students at the beginning of their academic careers, and was organized by the Centro Singular de Investigación en Tecnoloxías da Información (CiTIUS)[3] of the Universidade de Santiago de Compostela (USC).[4]

The basic idea behind the summer school was the analysis and management of large linked data that especially included topics like: Big Data, Linked Data, natural language processing (NLP), the Semantic Web and information retrieval (IR). Eight prominent speakers from these areas have

[1] European Coopeartion in Science and Technology. www.cost.eu/COST_Actions
[2] http://www.cost.eu/COST_Actions/ict/IC1302
[3] Centro de Investigación en Tecnoloxías da Información (CiTIUS), https://citius.usc.es/
[4] Universidade de Santiago de Compostela, http://www.usc.es/

prepared interesting materials, which are still publicly available.[5] The summer school was attended by 38 participants from thirteen countries, i.e. from three continents – Europe, America and Africa. The summer school was organized so that the participants had the opportunity to hear lectures by eminent lecturers from different European universities during the morning, while within the afternoon workshop they had the opportunity to try out what they have learned, through practical examples. Also, participants had the opportunity to hear the experiences of representatives of Hewlett Packard as well as the company E˜Xenia on "Keyword search in Big Linked Data".

Laura Po, lecturer at the Department of Engineering of the University of Modena and the Emilia Region[6] spoke about research, visualization and query formulation over large data sets of linked data. The participants were first introduced to the concepts of linked data and open data, how they are published, and their role in the Semantic Web, using very vivid examples. After that she talked about research and query formulation over large sets of linked data. Special attention was paid to the SPARQL[7] language used for query search over open linked data. The concept of the Datahub has been introduced, as well as metadata search for available data sets, methods and formats for dataset downloading (requiring), and the use of SPARQL endpoints, when they are available for a specific collection. The importance of visualization of linked open data was highlighted, with demonstration of the LODeX tool with specific query examples and the resulting interactive graphs.

During next session, the summer school students were introduced to the basic concepts and techniques in information retrieval (IR). Through a comprehensive and interesting lecture by Mikhail Lupu from the Technical University in Vienna[8], participants had the opportunity to learn about diverse evaluation techniques as well as the key criteria for evaluation in IR. Special attention was given to the application of statistical methods for verification of results. This issue was tackled by Sergei Zer from the University of Southampton[9] with his lecture "Collective Intelligence: Crowdsourcing Ground Truth Data for Large Scale Evaluation and Information Retrieval", which captured the attention of the auditorium from the first to the last

[5] https://eventos.citius.usc.es/keystone.school/slides.html
[6] Università degli Studi di Modena e Reggio Emilia (University of Modena and Reggio Emilia), http://www.unimore.it/
[7] SPARQL Protocol and RDF Query Language
[8] Technische Universität Wien (TUW), www.tuwien.ac.at
[9] University of Southampton, www.southampton.ac.uk

minute, involving participants in the discussion and showing them interesting examples.

Substantial lecture of Genevieve Varga-Solar, researcher from the French National Centre for Scientific Research,[10] completed the picture of the achievements so far related to storage and processing of Big linked data. She paid special attention to trends in the analysis of Big data, data mining and the data science, as well as to the specifics of distributed modelling, warehousing, predictive analytics, clustering, treatment and research in stream processing & mining and declarative languages.

Elena Demidova from the University of Southampton introduced participants of the summer school to interactive search over large structured data sets based on keywords. Search of structured data using keywords, as opposed to the classic query language (SQL), requires prior preparation of data (indexing) and translating user information needs into a convenient form. The good side of this approach is the possibility of querying even when the database schema is unknown, with a drawback, namely imprecision of the interpretation of user information needs, because the initial request expressed by keywords translates into a (most probable) structured query. Query formulation and structuring for large databases, for example Freebase, with over 22 million entities, 350 million facts, 7,500 relational tables and about 100 domains, requires an effective and scalable interactive query construction, which was presented both theoretically and through examples. During the summer school professor of University of Belgrade, Faculty of Mining and Geology, Ranka Stanković, gave a lecture on language resources. She paid special attention to query expansion and semantic annotation, as the means for improvement of search results, in terms of increasing the response without loss of precision. During the lecture, participants were introduced to various tools for language processing. Within the workshop, Professor Stanković presented some of the resources developed for Serbian and demonstrated their application.

Mauro Dragoni from the Bruno Kessler Foundation,[11] on the last day scheduled for lectures, emphasized the importance of looking at a document from different perspectives, and presented several case studies that contributed to a deeper understanding of the concepts presented.

[10] Centre national de la recherche scientifique, French Council of Scientific Research (CNRS), http://www.cnrs.fr

[11] Fondazione Bruno Kessler, www.fbk.eu

The summer school ended with perhaps the most interesting session, where participants organized in groups of 3 to 4 participated in the Hackathon competition. The task given to the participants required both programming skills and application of tools and knowledge presented during the summer school. Within the competition, two well-known data sets of linked data were used for solving the task: DBPedia and GeoNames. DBPedia is the central repository of linked data extracted from Wikipedia, which describes more than 4.5 million entities classified in a consistent ontology, of which approximately 1,445,000 entities about people, 735,000 related to cities, 411,000 for artworks, and 241,000 about organizations. There are versions for 125 different languages, but the ontology for English is the most used and the largest. In addition to links to images and external sites, categories from Wikipedia and YAGO ontology are associated to entities, which allows users to create and submit SPARQL queries over data derived from Wikipedia. Many data sets and DBPedia ontologies can be downloaded or directly accessed online using SPARQL http://dbpedia.org/sparql endpoint, as well as searched by keywords using DBPedia Lookup[12] tool, or downloaded and used as local linked data.

GeoNames,[13] the geographical database, contains over 10 million geographical names and 9 million unique geographical entities, classified into nine classes, and annotated with 645 markers. These classes and markers are described by GeoNames ontology[14] and predefined codes.[15] Each element has a GeoNames URI and a corresponding RDF document with XML data. For example, the element for Belgrade (Beograd) has URI http://www.geonames.org/792680/ and the corresponding document is http://www.geonames.org/792680/about.rdf, in which, besides its name in English, Belgrade is described in additional 67 languages. GeoNames elements are linked to each other using the three types of geographical relations: subordinate, in terms of administrative sub-unities, adjacent, as for example a neighbouring country, and spatially close, as for example settlements that are located at a small distance. Thus, all regions in Montenegro can be accessed when the suffix contains.rdf is added to its base URI http://www.geonames.org/3194884.

The data from the GeoNames database can be downloaded and used locally, but they can also be accessed online via its root hierarchy node http:

[12] http://wiki.dbpedia.org/projects/dbpedia-lookup
[13] http://www.geonames.org/
[14] http://www.geonames.org/ontology/documentation.html
[15] http://www.geonames.org/export/codes.html

//sws.geonames.org/6295630/about.rdf, and then by following *contains* relations, or by using web services based on keywords[16]. Finally, there are links from GeoNames elements to Wikipedia and DBPedia.

Participants were given the task to use DBPedia and GeoNames to create a web application for: 1) finding the administrative areas affected by a specified hurricane, 2) finding all the hurricanes that hit specified administrative areas, and 3) assessing how certain authorities were prepared to cope with a hurricane.

1. In the first scenario, the user has to specify keywords that identify the hurricane, which include its name (Katrina, Emily, etc.) and optionally year (2005, 2006, etc.). The text on the hurricane needs to be located in DBPedia, and then extracted, while the search should include the abstract and the text describing the affected area. From snippets of text, the system should extract the names of administrative areas using GeoNames database and Stanford NER[17] tool for Named Entity Recognition.
2. In the second scenario, the user defines the administrative area by keywords, after which the GeoNames database is used to find settlements in the given area, and finally DBPedia searched in order to identify the hurricanes that hit those settlements, i.e. area.
3. The third part comprised of determination of a numerical indicator that assesses the preparation level of the administrative area for a hurricane using data from DBPedia (casualties, damage, wind velocity, duration, type, and the like) and the number of inhabitants in the affected area found in the previous query. The task also included ranking based on the proposed indicator.

One of the strategies of the Serbian team for solving the given problem was the creation of an application in *Python* programming language, which uses all available open resources in order to access the given data and process them. The idea was that once the user selects a hurricane of interest, the system sends a SPARQL query to the open DBPedia API to extract texts from the abstracts and from the description of the affected area. After that, the affected areas were extracted from the results of the previous query, and a new query was sent to the GeoNames database, which retrieved information on all affected settlements for the given area. Information such as population, number of casualties, material damage, wind speed,

[16] http://www.geonames.org/export/geonames-search.html

[17] http://nlp.stanford.edu/software/CRF\begingroup\let\relax\relax\endgroup[Pleaseinsert\PrerenderUnicode{ʁʃθ}intopreamble]NER.shtml

and the like, were extracted from resources obtained, and formed the basis for determination of numerical indicators that showed the preparedness of a particular administrative area for the hurricane. These numerical indicators are meant for creating a mathematical model and its visual representation and interpretation.

On the last day of summer school, marked by a genuinely competitive atmosphere and devoted work of all participants, the teams presented the developed solutions. The most successful solution was developed by students of master and doctoral studies in Zaragoza, and as the best team, they won a valuable prize: the scholarships for stay and research at the University of Santiago de Compostela. The system evaluation implied the assessment of the precision and recall of the system, speed of execution, as well as the effectiveness of the solution. After the closing ceremony, participants had the opportunity to enjoy the picturesque city of Santiago de Compostela and the charms of Galicia.

All materials from the summer school are available at https://eventos.citius.usc.es/keystone.school/index.html.

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).