

14th International Conference on Current Research Information Systems, CRIS2018
**FAIRness at University of Novi Sad - Discoverability of PhD
research results for Non-Serbian scientific community –**

Dragan Ivanović^{a,*}, Lidija Ivanović^a, Colin Layfield^b

^aUniversity of Novi Sad, Serbia

^bUniversity of Malta, Malta

Abstract

The paper describes features implemented at University of Novi Sad digital library of PhD dissertations which improves the discoverability of the dissertations written in Serbian language for the non-Serbian scientific community. The Serbian Ministry of Science prescribed a rule that each dissertation, before it is defended, has to be verified by publishing its main results in at least one paper published in a journal belonging to the Web of Science JCR list. This paper is usually published in the English language. An approach, which enables establishing the connection between those journal papers and PhD dissertations, has been implemented at the University of Novi Sad. This approach makes accessible research presented in Serbian PhD dissertations for the English scientific community through associated journal papers. Moreover, this paper describes two approaches for enabling search on research presented in Serbian PhD dissertations for ex-Yugoslav scientific community. The first approach is based on common preprocessing tool. This approach has been implemented and integrated into PHD UNS digital library. It is based on similar morphology of ex-Yugoslav languages and small parallel vocabulary of Serbo-Croatian terms. Taking into account that development and maintenance of a complete parallel vocabulary is time-prone, we started the another approach based on the generation of a multi-lingual semantic space utilising latent semantic analysis which also seeks to exploit the similarities between the ex-Yugoslav languages.

© 2019 The Authors. Published by Elsevier B.V.

This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/3.0/>)

Peer-review under responsibility of the scientific committee of the 14th International Conference on Current Research Information Systems, CRIS2018.

Keywords: FAIR principles, Multilingual search, non-English publications, PHD UNS, linking of publications

1. Introduction

The University of Novi Sad has more than 50,000 students and 5,000 employees and as such it is one of the largest educational and research centres in Central Europe. The University of Novi Sad offers 300 accredited study programmes on all study levels, undergraduate, masters, specialist and doctoral studies, which are carried out at the Faculties and University Centres for Interdisciplinary and Multidisciplinary Studies.

About 5,000 PhD dissertations have been defended at University of Novi Sad since its foundation, and more than 1,000 in the last four years. More than 95% of those dissertations have been written in Serbian. Furthermore, each

*Corresponding author. Tel.: +381 21 485 2447

E-mail address: dragan.ivanovic@uns.ac.rs

PhD dissertation typically has a key words list in Serbian and English. Key words documentation is prescribed by the University of Novi Sad board and contains the following metadata: title, author, commission members, institution, abstract, advisor, publication year, publisher, etc.

Metadata and data should be findable and understandable to humans and machines according to the **FAIR** principles¹. Authors of scientific publications, including PhD dissertations, usually want to attract interest to their work to enhance their reputation both nationally and internationally. Although metadata about dissertations is expressed in Serbian and English improves the discoverability of Serbian PhD dissertations, non-Serbian researchers still can't read and search by full-text PhD the dissertations which are written in the Serbian language. This issue is the main topic of this paper.

The rest of the paper is structured as follows. The second and third chapters briefly describe the FAIR principles and the DOSIRD UNS project (digital library and CRIS UNS system), respectively. The fourth and fifth chapters describe our approaches which make Serbian PhD dissertations more findable and accessible for English-speaking and ex-Yugoslav scientific community, respectively. At the end of the paper, some conclusions and further development directions are being discussed.

2. FAIR principles

The FAIR Data Principles are a set of guiding principles to aid in making data findable, accessible, interoperable and reusable². In order to be findable, data and supplementary materials should have sufficiently rich metadata and a unique and persistent identifier. Moreover, metadata and data should be understandable to humans and machines in order to be accessible. Also, data should be deposited in a trusted repository. Furthermore, metadata should be represented by using a formal, accessible, shared, and broadly applicable language for knowledge representation in order to improve interoperability. At the end, data and its metadata should have a clear usage licenses and provide accurate information on provenance.

Although FAIR principles have been implemented at the University of Novi Sad digital library, the topic of this paper is a specific aspect of FAIR principles - making findable and accessible research information stored in PHD UNS dissertations to the non-Serbian scientific community.

3. DOSIRD UNS

The main goal of the DOSIRD UNS project (<http://dosird.uns.ac.rs/>) is developing software infrastructure for the research domains of the University of Novi Sad. The project has been started in 2009 and its main result is CRIS UNS³ (Current Research Information System of the University of Novi Sad). CRIS UNS is a CERIF compatible CRIS system⁴. PHD UNS is the digital library of PhD dissertations defended at University of Novi Sad. It is also a result of the DOSIRD UNS project. Moreover, PHD UNS is integrated with the CRIS UNS system⁵ and it is based on a CERIF compatible data model^{6,7}.

The PHD UNS digital library preserves:

- Metadata prescribed by the University for all dissertations defended at the University;
- PDF files representing PhD dissertation for dissertations defended in the last ten years;
- Open-access license signed by a PhD student for dissertations defended in the last four years.

A web page for searching dissertations has been implemented (www.cris.uns.ac.rs/searchDissertations.jsf). Dissertations are searchable by metadata in both languages (Serbian and English) and by full-text (extracted from PDF files). Search results can be represented in visual (in the form of words cloud) and textual mode. Textual mode includes representation of results in four formats: Harvard citation style, Dublin Core, MARC 21, and ETD-MS. If a PDF file is attached to dissertation record and the dissertation is published under an open-access licence, a link for downloading the dissertation is made available. Before a download is performed, a web page with information about the signed open-access license is shown to the user. A PhD author can select and sign one of six Creative Commons licences. Each download is stored in the download logs and is used as an input to an automatic recommendation sys-

tem⁸. Metadata about all dissertations published under an open access licence is available for exporting via OAI-PMH protocol to service providers (DART-Europe, OpenAIRE+, NDLTD, etc.).

All previously stated features implemented in PHD UNS improves FAIRness of research information available in PhD dissertations, but there is still a language barrier to take into account in that most of the dissertations are written in the Serbian language. As already stated, authors of scientific publications, including PhD dissertations, usually want to attract interest to their work to enhance their reputation both nationally and internationally.

4. English scientific community

English has become the global language in the economic, scientific and cultural world largely dominated by Anglo-American countries. However, even if English is the dominant language in science, it is certainly not the native language for the majority of scientists and there is much to be discovered in academic material not presented in English. More than 95% of PhD dissertations of the republic of Serbia have been written in Serbian. However, the Serbian Ministry of Science prescribed a rule that each dissertation, before it is defended, has to be verified by publishing its main results in at least one paper published in a journal belonging to the Web of Science JCR list. The objective of this is to ensure results published in the Serbian language PhD dissertation are also visible to the global science community through associated journal papers which are usually published in English. The same approach can be adopted for all non-English digital libraries of research publications all over the world.

An approach which enables establishing a connection between those journal papers and PhD dissertations has been implemented at the University of Novi Sad PHD UNS repository and CRIS UNS system by developing the following features:

- The connection between a PhD dissertation and journal paper, which presents the main results of the PhD research, can be established by the author of the dissertation or by the librarian of the institution where the dissertation has been defended (see Figure 1).
- If a PhD dissertation is a result of local searching of the PHD UNS digital library, a reference to the associated journal paper written in English is provided (see Figure 2).
- If a PhD dissertation has been exported to ETD repositories networks, such as DARTEurope (www.dart-europe.eu/full.php?id=1058428), a link (identifier) to the PhD dissertation stored in PHD UNS opens web page will provide a reference to the associated journal papers written in English (see Figure 3). Connected journal papers can be searched and downloaded using other scientific output repositories: Scopus, Google Scholar, etc. Taking into account that this is outside of our system, we can only log in PHD UNS system the information that connected journal papers have been displayed to the user. We can't analyse how often this feature is used or how many times users downloaded the connected journal papers.

A similar approach can be adopted for establishing the connection between datasets and PhD dissertations taking into account that the CRIS UNS system can also store datasets using the *cfResultProduct* entity of the CERIF data model. However, at this moment CRIS UNS system doesn't preserve datasets.

5. Ex-Yugoslav scientific community

Since the beginning of national conflicts and Yugoslav wars (in the 90s), all ex-Yugoslav republics became independent countries. The official language of Yugoslavia was the Serbo-Croatian language. Today, we have similar languages, by morphology and vocabulary, derived from the Serbo-Croatian language: Serbian, Croatian, Montenegrin and Bosnian. As more than 20 million people from the Ex-Yu region have reading skills in all of the languages of the region, multi-lingual search capabilities over texts and metadata written in those languages should be implemented in order to enable scientists to query the system in their native language and find publications in other regional languages that would otherwise be overlooked due to the language differences. There are two approaches for developing ex-Yugoslav language tools which would support those multi-lingual search capabilities. The first one includes development of common preprocessing techniques including stemming and parallel vocabularies. This approach is possible because of very similar morphology and vocabulary of ex-Yugoslav languages. The second approach includes creation

Related publications for the selected Ph.D. dissertation

Selected dissertation

Thesis or dissertation

Author: Milić Uroš

Publication language: Serbian

Title: *Production and quality assessment of fruit wines from native plum (Prunus domestica L.) varieties*

Theses/dissertation type: PhD dissertation

Institution: Faculty of Technology at Novi Sad, University of Novi Sad

Year: 2015

Download: MILJIĆ, U., PUŠKAŠ, V., VUČUROVIĆ, V. and RAZMOVSKI, R. (2014) Acceptability of wine obtained with increased content of grape seed and stem as functional food. *Journal of the Institute of Brewing*, 120 (2), pp. 149-154MILJIĆ, U. and PUŠKAŠ, V. (2014) Influence of fermentation conditions on production of plum (*Prunus domestica* L.) wine: A response surface methodology approach. *Hemjska industrija*, 68 (2), pp. 199-206

Publications

Selection

Čokić Jelena, Puškaš Vladimir, Stanić Uroš, Torović Ljilja, Rakić Dušan, Varietal phenolic composition of Probes, Rumenika and Frankovka red wines from Fruška Gora (Serbia) and changes in main compounds during maceration, <i>European Food Research and Technology</i> (ISSN: 1438-2377), Vol 242, No 9, 2015, pp. 1319-1328	
Milić Uroš, Puškaš Vladimir, Velikanović Aleksandra, Mašković Pavle, Orešković Dragoljub, Vujić Jelena, Chemical composition and in vitro antimicrobial and cytotoxic activities of plum (<i>Prunus domestica</i> L.) wine, <i>Journal of the Institute of Brewing</i> (ISSN: 0046-0750), Vol 122, 2016, pp. 342-349	
Milić Uroš, Puškaš Vladimir, Vučurović Vesa, Investigation of technological approaches for reduction of methanol formation in plum wine, <i>Journal of the Institute of Brewing</i> (ISSN: 0046-0750), 2016	
Mašković Pavle, Mašković Zorana, Puškaš Vladimir, Milić Uroš, Čokić Jelena, Torović Ljilja, Chemical composition and in vitro antimicrobial and cytotoxic activities of plum (<i>Prunus domestica</i> L.) wine, <i>Journal of the Institute of Brewing</i> (ISSN: 0046-0750), Vol 120, No 1, 2015, pp. 103-108	
Mašković Zorana, Mašković Pavle, Puškaš Vladimir, Milić Uroš, Čokić Jelena, Torović Ljilja, Chemical composition and in vitro antimicrobial and cytotoxic activities of plum (<i>Prunus domestica</i> L.) wine, <i>Journal of the Institute of Brewing</i> (ISSN: 0046-0750), Vol 120, No 2, 2015, pp. 159-163	
Milić Uroš, Puškaš Vladimir, Vučurović Vesa, Radojka, Characterization of water-soluble compounds from different stages of wine production, <i>Journal of the Institute of Brewing</i> (ISSN: 0046-0750), Vol 120, No 2, 2015, pp. 139-143	
Milić Uroš, Puškaš Vladimir, Vučurović Vesa, Radojka, Acceptability of wine obtained with increased content of grape seed and stem as functional food, <i>Journal of the Institute of Brewing</i> (ISSN: 0046-0750), Vol 120, No 2, 2014, pp. 149-154	
Milić Uroš, Puškaš Vladimir, Influence of fermentation conditions on production of plum (<i>Prunus domestica</i> L.) wine: A response surface methodology approach, <i>Hemjska industrija</i> (ISSN: 0367-598X), Vol 68, No 2, 2014, pp. 199-206	
Vučurović Vesa, Razmovski Radojka, Milić Uroš, Puškaš Vladimir, Removal of cationic and anionic azo dyes from aqueous solutions by adsorption on maize stem tissue, <i>Journal of the Taiwan Institute of Chemical Engineers</i> (ISSN: 1076-1070), Vol 45, No 4, 2014, pp. 1700-1703	

Save

Select all

Deselect all

Fig. 1. Establishing link

Search results

Number of results is 1

Sort by: relevance

Style: structured

1 MILJIĆ, U. (2015) *Production and quality assessment of fruit wines from native plum (Prunus domestica L.) varieties*. (PhD dissertation), Faculty of Technology at Novi Sad 

Additional data MARC 21 Dublin Core ETD MS Digital document Related papers

MILJIĆ, U., PUŠKAŠ, V., VUČUROVIĆ, V. and RAZMOVSKI, R. (2014) Acceptability of wine obtained with increased content of grape seed and stem as functional food, *Journal of the Institute of Brewing*, 120 (2), pp. 149-154MILJIĆ, U. and PUŠKAŠ, V. (2014) Influence of fermentation conditions on production of plum (*Prunus domestica* L.) wine: A response surface methodology approach. *Hemjska industrija*, 68 (2), pp. 199-206

Fig. 2. Search result

Record preview

Thesis or dissertation

Author: Milić Uroš

Publication language: Serbian

Title: *Production and quality assessment of fruit wines from native plum (Prunus domestica L.) varieties*

Theses/dissertation type: PhD dissertation

Institution: Faculty of Technology at Novi Sad, University of Novi Sad

Year: 2015

Download: 

Related records

Milić Uroš, Puškaš Vladimir, Influence of fermentation conditions on production of plum (*Prunus domestica* L.) wine: A response surface methodology approach, *Hemjska industrija* (ISSN: 0367-598X), Vol 68, No 2, 2014, pp. 199-206.Milić Uroš, Puškaš Vladimir, Vučurović Vesa, Razmovski Radojka, Acceptability of wine obtained with increased content of grape seed and stem as functional food, *Journal of the Institute of Brewing* (ISSN: 0046-0750), Vol 120, No 2, 2014, pp. 149-154.

Fig. 3. Record preview

of a multi-lingual semantic space using Latent Semantic Analysis (LSA). An LSA approach would entail a creation of a virtual semantic space that relies on the assumption that, provided the languages are using the same script (latin or cyrillic), the underlying similarities between the languages can be exploited to perform a semantically driven search taking into account synonymy and co-occurrence information in all the languages that are to be supported (Serbian, Croatian, Montenegrin and Bosnian). The objective of both those approaches is to ensure that results published in the

Serbian language PhD dissertation are also visible to the scientists from ex-Yugoslav region, and the two approaches will be presented in the following two subsections.

5.1. *Ex-Yugoslav common preprocessing tool*

The developed ex-Yugoslav language tool performs preprocessing of PDF texts and metadata before its indexing. This language tool is developed as a Lucene Analyser⁹. Pre-processing includes the following tasks:

- Case folding - all capital letters are transformed to small letters.
- Alphabet conversion - conversion of the Cyrillic characters into their corresponding characters from the Latin alphabet. Those two alphabets are used in Ex-Yugoslav countries. Also, those two alphabets have the same number of letters (30) and each letter of Cyrillic alphabet can be unambiguously mapped to a Latin letter. This conversion enables ex-Yugoslav-alphabets insensitive search of PhD dissertations. It means that PhD dissertation which is written using Cyrillic alphabet can be found using a query written using Latin alphabet and vice versa.
- Stop word removal - common stop words for Serbo-Croatian languages are defined, and all those words are removed from the output of preprocessing.
- Stemming - Creation of a unique stemming is possible, because all ex-Yugoslav languages have a similar morphology. Rules for the stemming of Croatian text¹⁰ have been used as basis. Sets of rules have been slightly changed in order to take into account different dialects used in the ex-Yugoslav region, as well as to take into account some specificities of the Serbian language. Stemming rules are expressed in Snowball language and transformed to a Lucene SnowballFilter using the Snowball compiler for generating Java platform output.
- Vocabularies translation - The small parallel Serbo-Croatian vocabulary has been created. This vocabulary contains about 11,000 pairs of Serbian and Croatian language terms. A Lucene Filter, which applies this parallel vocabulary to replace specific Croatian terms with Serbian terms, has been developed. However, this vocabulary is not a complete Serbo-Croatian parallel vocabulary, and there is a problem with sustainability of this vocabulary taking into account that languages evolve.
- Characters conversion - conversion of non-ASCII characters to their corresponding characters from the ASCII character set. There are 30 letters of ex-Yugoslav alphabet and some of them include accents (diacritics: é, ě, š, ž). Accents (diacritics) are removed as a result of the preprocessing tool.

The developed ex-Yugoslav language tool has been put into operation in 2013. It is integrated as part of PHD UNS digital library indexing process and query execution. The tool is used for text and metadata preprocessing before indexing, as well as for query preprocessing before its execution. The PHD UNS digital library logs information about downloaded PhD dissertations including the location of a client device. There are more than 10,000 downloads per month. About 5% of the total number of downloads are from Croatia, Montenegro and Bosnia and Herzegovina (three ex-Yugoslav countries). We suppose that the ex-Yugoslav language tool helps users from those countries to find PhD dissertations using queries expressed in their national languages. Also, we received a few email requests for non-open PhD dissertations from scientists from Croatia.

5.2. *Ex-Yugoslav common semantic space*

There are at least two drawbacks of the previous approach:

- Time-prone process of creation of parallel vocabularies - it is almost impossible to create a complete parallel vocabulary for Serbo-Croatian languages.
- Even though if we have a complete parallel vocabulary, it should be maintained taking into account that languages evolve.

Because of that, we have been researching a new approach for this issue based on LSA. An exploratory case study was performed in 2017 that investigated the creation of multi-lingual semantic spaces for Information Retrieval.

Specifically, this experiment targeted the Serbian and Croatian languages. The underlying belief, as explained previously, is that these languages are very similar and, thus, it may be possible to create a joint, or mixed, semantic space utilising Latent Semantic Analysis, for semantic similarity comparison purposes¹¹.

5.2.1. Latent Semantic Analysis

Latent Semantic Analysis, or LSA^{12,13}, is an Information Retrieval technique whose origins are found in factor analysis. The central idea behind LSA is to generate the classic $m \times n$ term by document matrix (typically referred to as A) from a corpus of documents. The matrix is then factored as follows via the process of *singular value decomposition* (SVD) such that $A = TSD^T$. T is a $m \times r$ matrix, the D^T matrix is $r \times n$ where each of these is made up of orthonormal eigenvectors. The S matrix (containing the *singular values*) is a $r \times r$ diagonal matrix where r designates the *rank* of matrix A ¹³. This creates what we refer to as a *semantic space*. The rows (vectors) in the matrix T correspond to the *terms* in this newly created semantic space and the columns (vectors) of D^T represent the documents. The vectors in D^T representing documents can now be compared with one another, using a metric such as *cosine similarity*, in order to ascertain the semantic similarities between them. The same feature is available for comparisons between *term* vectors as represented by T . The only caveat involved with this is that the associated term or document vector must first be scaled by the S matrix (for full details of the mathematics one can refer to¹²). Additional documents (or queries) can be *folded into* the semantic space via projection by multiplying the new document vector by TS^{-1} . Any new document/query projected into the semantic space in this manner can now be compared with other documents in the semantic space (thus, a search operation can be carried out).

The true benefit of LSA is only apparent when *dimensionality reduction* is employed. This is carried out by choosing some value, say k , which is less than r and reduce the dimensionality of the singular value matrix such that it now only has k dimensions (with the rest essentially being zeroed out). This has the net effect of having the document and term vectors represented in k dimensions rather than r . The net gain from this is to attempt to exploit any underlying latent semantic structures in the set of documents that is potentially hidden by the nature of word choice (having some degree of randomness to it). This dimensional reduction attempts to filter out this *noise* in order to exploit the underlying latent semantics (word relationships) that may exist in the set of documents¹³.

5.2.2. Multi-Lingual Search With LSA

LSA has been employed in multi-lingual search in the past (typically referred to as ML-LSA). The classic approach is to utilise a parallel corpus of documents for creating a semantic space. This parallel corpus of documents will be a set of documents but each document will have a version in each of the languages being targeted. These documents that are represented in multiple languages are then concatenated together such that each linguistic version of the same 'document' is combined into a single document (with respect to our A matrix). Thus each document represented in A contains the text for the same document in all the languages being supported. The only drawback to this is that such a parallel corpus must exist and be comprehensive enough to generate a meaningful semantic space via SVD (this has led to the text from the Bible as being a popular target for research with parallel bodies of text, as it has been translated into so many languages, as demonstrated by the work of Young¹⁵ in a ML-LSA context). For multiple languages this can be an inherent drawback. An experiment was performed using this approach using the Hansard Collection, which is a collection of Canadian parliament interactions (in both the official languages of English and French). They evaluated this using a 'mate-retrieval' test which involves selecting the same document in both languages (that does not exist in the semantic space) and folding them in to see if they naturally retrieve their 'mate'. The average success rate for this test was 98.4%¹⁴.

A similar experiment was carried out utilising a parallel corpus of documents in both Serbian and Croatian. The underlying hypothesis that was tested is to see if a 'mixed' semantic space, that is a semantic space of documents that contain text from the Serbian and Croatian languages that are *not* concatenated, can yield reasonable results from a similar experiment. That is, the semantic space was composed of documents that were either in Croatian or Serbian. It was found that the 'mate' documents could be found as the main search result more than 90% of the time and the mate would actually be returned in a list of top 5 search results almost 95% of the time. This demonstrates a 'mixed' semantic space, as described above, has some promise for enabling a multi-lingual search in scenarios such as the one described here. That is to say due to the inherent similarities between Serbian, Bosnian, Montenegrin and Croatian it may be possible to generate a semantic space with a mixture of these languages that can be utilised for search scenarios such as those envisioned in this paper. The full experiment described above can be found in¹¹.

5.2.3. Future Research Directions Utilising LSA

Further work is planned for this avenue of enquiry. It would be prudent to verify similar results can be obtained with a semantic space generated from documents written in Serbian, Croatian, Montenegrin and Bosnian. The size of documents and usage of a training set that is not parallel in nature will also need to be experimented with in order to determine if similar positive results can be achieved. It would be ideal to demonstrate that a parallel corpora of documents is not required at all as this may not be a realistic constraint for practical application. The end goal of this line of enquiry is to see if it is possible to create a search repository that can be freely searched in any of the four languages targeted above.

Another application of the semantic space, that can potentially address the first shortcoming listed at the beginning of this section, could be the automatic generation of a synonymy vocabulary list between the different languages contained in a semantic space. If a semantic similarity comparison was carried out between each term of each of the languages contained within it should be possible to generate such a list and supplement the parallel dictionary utilised in the previous approach.

6. Conclusion

This paper presents features implemented at the University of Novi Sad which improve discoverability of Serbian-language PhD dissertations for the non-Serbian scientific community. Those features have been implemented within PHD UNS digital library which is integrated with CRIS UNS system. PHD UNS is a digital library of PhD dissertations defended at University of Novi Sad, while CRIS UNS is research management system of the University.

An approach which makes accessible research presented in Serbian PhD dissertations for the English scientific community through connected journal papers has been described. Those connected journal papers are usually written in English and present a part of the research results described in PhD dissertations. Those connected journal papers haven't been stored in digital format in CRIS UNS system, but its metadata are stored. Because of that, the connected journal papers are displayed to the PHD UNS / CRIS UNS users, but users have to find and download those journal papers using other scientific output repositories: Scopus, Google Scholar, etc. Taking into account that is outside of our system, we can only detect PHD UNS system information that connects journal papers that have been displayed to the user. We can't analyse how useful this feature is nor how many times users actually downloaded the connected journal papers.

Moreover, the paper describes two approaches for making findable research presented in Serbian PhD dissertations findable for the ex-Yugoslav scientific community. The official language of Yugoslavia was the Serbo-Croatian language. Today, we have similar languages by morphology and vocabulary derived from Serbo-Croatian language: Serbian, Croatian, Montenegrin and Bosnian. As more than 20 million people from the Ex-Yu region have reading skills in all of the languages of the region, multi-lingual search capabilities over texts and metadata written in those languages should be implemented in order to enable scientists to query the system in their native language and find publications in other regional languages that would otherwise be overlooked due to the language differences. We presented the two approaches for multi-lingual search capabilities implementation. The first approach is based on common preprocessing tool. This approach has been implemented and integrated into PHD UNS digital library. It is based on similar morphology of ex-Yugoslav languages and a small parallel vocabulary. Taking into account that development and maintenance of a complete parallel vocabulary is time-prone, we started the another approach utilising Latent Semantic Analysis (LSA) techniques. This approach is implemented and verified using Serbo-Croatian parallel corpus. Further development includes application of this approach using parallel corpus for four ex-Yugoslav languages: Serbian, Croatian, Montenegrin and Bosnian.

References

1. Implementing FAIR Data Principles: The Role of Libraries. LiberEUROPE, 2017
2. Wilkinson M. D. et al. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data* 2016; **3** (160018). DOI 10.1038/sdata.2016.18
3. Surla D, Ivanović D, Konjović Z. Development of the software system CRIS UNS. In *Intelligent Systems and Informatics (SISY), IEEE 11th International Symposium* 2013; p. 111-6
4. Jeffery, K, Asserson, A. CERIF-CRIS for the European e-Infrastructure. *Data Science Journal* 2010; **9**: CRIS1-CRIS6.

5. Ivanović L, Ivanović D, Surla D. Integration of a Research Management System and an OAI-PMH Compatible ETDs Repository at the University of Novi Sad, Republic of Serbia. *Library Resources and Technical Services* 2012, **56** (2): 104-12. DOI 10.5860/lrts.56n2.104
6. Ivanović L, Ivanović D, Surla D. A data model of theses and dissertations compatible with CERIF, Dublin Core and EDT-MS. *Online Information Review* 2012; **36** (4):568-86. DOI 10.1108/14684521211254068
7. Ivanović D, Surla D, Konjović Z. CERIF compatible data model based on MARC 21 format. *The Electronic Library* 2011; **29** (1):52-70. DOI 10.1108/026404711111111433
8. Azzopardi J, Ivanovic D, Kapitsaki G. Comparison of Collaborative and Content-Based Automatic Recommendation Approaches in a Digital Library of Serbian PhD Dissertations. In *Semantic Keyword-based Search on Structured Data Sources* 2016, p. 100-11
9. Lucene - Analyzer, https://www.tutorialspoint.com/lucene/lucene_analyzer.htm
10. Ljubešić N, Boras D, Kubelka O. Retrieving information in Croatian: Building a simple and efficient rule-based stemmer. *Digital information and heritage* 2007: 313-20.
11. Layfield C, Ivanović D, Azzopardi J. Multi-Lingual LSA with Serbian and Croatian: An Investigative Case Study. In *Proceedings of the 3rd International KEYSTONE Conference IKC* 2017; Gdansk, p. 155-64
12. Landauer T K, McNamara D S, Dennis S, Kintsch W (Eds). *Handbook of Latent Semantic Analysis* 2007, Mahwah, NJ, US: Lawrence Erlbaum Associates Publishers.
13. Deerwester S, Dumais S T, Furnas G W, Landauer T K, Harshman R. Indexing by Latent Semantic Analysis. *Journal of the American Society for Information Science* 1990, **41** (6):391-407.
14. Dumais S T, Letsche T A, Littman M L, Landauer T K. Automatic Cross Language Retrieval Using Latent Semantic Indexing. *AAAI Technical Report SS-97-05* 1997, p. 18-24.
15. Young P. *Cross-Language Information Retrieval Using Latent Semantic Indexing*. Master's thesis. University of Knoxville, Tennessee, United States of America. 1994.