

Customization of digital library of PhD dissertations for citizens

Ljubomir Paskali

University of Novi Sad, Serbia

ljubomir.paskali@gmail.com

Dragan Ivanovic

University of Novi Sad, Serbia

dragan.ivanovic@uns.ac.rs

Abstract

PHD UNS is digital library of PhD dissertations defended at University of Novi Sad. A web page for basic and advanced search has been developed in order to improve discoverability of dissertations stored in the digital library. This paper presents customization of PHD UNS web search pages for citizens out of academy. The customization includes extension of available representation styles and implementation of automatic recommendations of PhD dissertations. Representation styles are extended with textual representation specially designed for non-academic citizens and visual representation based on word clouds. Automatic recommendations are based on collaborative approach built on PhD download history, i.e., performed on the basis of what other 'similar' users have found useful. The PHD UNS digital library logs information for each dissertation downloading. Besides basic information about downloaded dissertation, those logs also contain information about client machine which requested downloading. Those logs have been used in order to prove our customization really improve non-academic users' experience.

Keywords: PHD UNS, representation style, word cloud, ELK

Introduction

The CRIS UNS system (<http://cris.uns.ac.rs>) is a research information system that has been developed at the University of Novi Sad and has been continuously maintained and updated since 2008. It was originally developed within the project DOSIRD UNS

(<http://dosird.uns.ac.rs/>). The digital library of Ph.D. dissertations of the University of Novi Sad (PHD UNS) stores the entire process of a dissertation's submission and defence. The digital library is integrated with the CRIS UNS system in order to create a unified central catalog of all scientific–research outputs published by the University (Ivanovic et al., 2012b).

Practically all digital libraries present a search interface whereby users can search for relevant content by expressing their needs in the form of queries. However, users may not even know exactly what they are looking for, and most often find it difficult to articulate their information needs through search queries. Notwithstanding these difficulties, users typically find it easy to identify those documents that satisfy their information needs. Automatic recommendation systems exploit this fact by constructing user models based on the users' prior behaviour, and then use these models to provide recommendations. Personalization of the search as well as automatic recommendations of results to users has been studied by many researchers (Azzopardi et al., 2016). Part of the search personalization refers to the personalization of results representation. The process of presenting results to users in textual and non–textual formats has been studied by many researchers in the past in order to improve user experience in search tools.

Although scholars are main users of web page for searching a digital library, citizens can also use this web page to have insights into the scientific and research outputs that are partly financed from the state budget to which the citizens themselves contribute through taxes. Moreover, citizens could be interested in results of some PhD dissertations applicable in business or everyday life. This paper presents customization of PHD UNS web search pages for citizens by extending list of available search results' representation styles and implementation of automatic recommendations of PhD dissertations.

Literature review

DOSIRD UNS

University of Novi Sad consists of 14 faculties and 2 research institutes. The faculties have two–folded mission: educational and research. For the educational domain, information systems have been developed for recording information about students and subjects, there are vice–deans for education, procedures and documents have been developed. Thus, for the educational domain, certain software and organizational infrastructures and legal regulative have been developed. On the other side, the situation is much worse for the research domain. Because of that the DOSIRD UNS project (<http://dosird.uns.ac.rs>) was launched in 2009. The aim of this project is to develop software and organizational infrastructure as well as legal regulative for the research domain of the University of Novi Sad. The first result of this project is the previously mentioned CRIS UNS system (<http://cris.uns.ac.rs>) which is the information system of the research domain of the University of Novi Sad. The development of this system has been started in 2009 and is still active. During the CRIS UNS system specification and implementation, the two main requirements were:

- the system has to be in accordance with international standards that

have been adopted in the research domain and

- the system has to meet all local requirements prescribed by the University, the Province of Vojvodina and the Republic of Serbia.

The paper (Ivanović et al., 2011a) proposes a CERIF (Common European Research Information Format) compatible data model based on the MARC 21 format. In that data model, the part of the CERIF data model representing research outputs is mapped to the MARC 21 format. The CRIS UNS system was built on that model. More information on system architecture and implementation can be found in papers (Ivanovic, 2010; Ivanovic et al., 2010; Milosavljevic et al., 2010; Ivanovic & Milosavljevic, 2010). Automatic extraction of metadata from research outputs such as title, keywords, abstract is presented in the paper (Kovacevic et al., 2011). The CRIS UNS reporting component is described in the paper (Dimić-Surla & Ivanovic, 2012). The possibility of exchanging metadata with other systems is discussed in the paper (Ivanovic, 2011). The CRIS UNS ontology is the topic of papers (Ivanovic et al. 2012d, Dimić-Surla et al. 2012). The CRIS UNS search service is based on a CQL profile for CRIS systems presented in the paper (Penca et al., 2012). The papers (Ivanović et al., 2011b, Ivanović et al., 2012c) present an extension of the CERIF data model for semi-automatic evaluating of published research outputs. The extension is based on the semantic layer of the CERIF model, by which it is possible to classify research domain entities and their relationships by different classification schemes. This model has been verified against rules prescribed by the Rulebook on the Evaluation of Scientific Research Outputs of University of Novi Sad (Surla et al., 2012). A service for evaluating of results outputs published in journals based on bibliometrics indicators is described in the paper (Nikolić et al., 2012).

The digital library of Ph.D. dissertations (PHD UNS) that is the topic of this paper has been integrated within the CRIS UNS system. The development of PHD UNS began in 2010 with the following features:

- PHD UNS is integrated within CRIS UNS.
- The digital library is CERIF compatible, that is, it can exchange metadata with CERIF compatible research information systems.
- A PHD UNS dissertation is described by a metadata set that covers all metadata prescribed by Dublin Core and EDT-MS metadata format, that is, the system can exchange data in Dublin Core or EDT-MS format via OAI-PMH protocol.
- The PHD UNS library has such a data model and architecture that it can be easily integrated with a library system based on the MARC 21 format.
- The user interface allows entry of dissertation data without knowing the standardized metadata formats on which the PHD UNS library is built.

More information on the development of the PHD UNS system can be found in the published papers (Ivanovic et al., 2012a; Ivanovic et al., 2012b; Ivanovic, 2013; Ivanovic et al., 2013; Ivanovic and Surla, 2012).

Search results representation

Part of the search personalization refers to the personalization of search results representation.

A word or tag cloud is a visual representation of word content commonly used to represent content in different environments (Scanfeld et al., 2010). For instance, tag clouds have been used in PubCloud for the summarization of results (Kuo et al., 2007) based on words extracted from the abstracts returned by the query.

Automatic recommendation

There are two groups of automatic recommendation techniques: content-based techniques – recommendation is performed based on similarity between the documents' contents; and collaborative techniques – recommendation is performed based on what other similar users have found useful.

Content-based and Collaborative recommendation systems have pros and cons comparing to each other. Collaborative recommendation systems are generally considered more simple and straight-forward to implement (Das et al., 2007). Moreover, this approach is domain and language independent (Bordogna et al., 2006; Das et al., 2007; Schafer et al., 2007). However, collaborative systems may lead to more unexpected recommendations (Schafer et al., 2007). On the other hand, content-based systems do not require a large number of users and could be applied in a limited user environment (Bordogna et al., 2006; Schafer et al., 2007).

Log analysis

After implementing the system, one way to improve the system and user experience is to analyze system usage, find user profiles and personalize the system according to user profiles. Moreover, system usage analysis can identify trends and leaders in a particular research area. Bollen and colleagues (2003) analyzed the use of digital libraries to identify trends in a research area and determine the measure for journal popularity. Bollen and Luce (2002) analyzed messages about the operation of digital libraries to determine user preferences and user profiles of a digital library. Zhang and colleagues (2001) analyzed the log messages of South Korea's digital dissertation library to determine usage patterns. They have come to the conclusion their digital library has increased international influence over the years, that is, increased the number of users from abroad, as well as the number of returning users. Jones and colleagues (1998) analyzed the usage patterns of a New Zealand digital library (www.nzdl.org/cgi-bin/library.cgi). The results of their analysis showed that users rarely change the default settings for defining queries and presentation of search results. Chau and colleagues (2005) compared usage patterns of digital libraries search and web search engines. They found that the average number of terms in the queries was similar, while the search topics and terms were significantly different. Sisodia and Verma (2012) analyzed the navigational behaviour of users using systems' log messages. The results of that type of

analysis can be used to analyze web traffic, modify the website, improve the system and personalize the user interface.

Methodology

We had to complete the following step to reach our goal and customize PHD UNS for citizens:

1. Preparation of PhD dissertations' index structure which should enable power search mechanisms.
2. Design and implementation of a web page for searching digital library.
3. Techniques and well-known formats for textual and visual representation of digital textual files have been analyzed, adopted and implemented within PHD UNS.
4. Implementation of customization of search results representation through user interface.
5. Log messages have been analyzed to determine the user's paternity behaviour based on query types and client device
6. Making a module for collecting log messages after downloading a dissertation.
7. Making a module for automatic recommendation for the certain user based on previously collected logs for that user and other PHD UNS users.
8. Evaluation of various approaches for automatic evaluation based on implicit feedback stored in download logs.
9. Integration of the best approach within the PHD UNS library
10. Collecting explicit and implicit users' feedback.
11. Evaluation of the integrated module for automatic dissertations' recommendation.
12. Analysis of usage of PHD UNS digital library by non-academic citizens.

Results and discussion

In this section we described how we implemented steps from the methodology section and discussed results.

1. Preparation of PhD dissertations' index structure which should enable power search mechanisms

The Lucene library has been used for implementation of information retrieval features. Language tools for Serbian language enable search insensitive on morphological and inflectional Serbian words' changes. Index structure includes basic metadata of the Thesis entity shown in Figure 1.

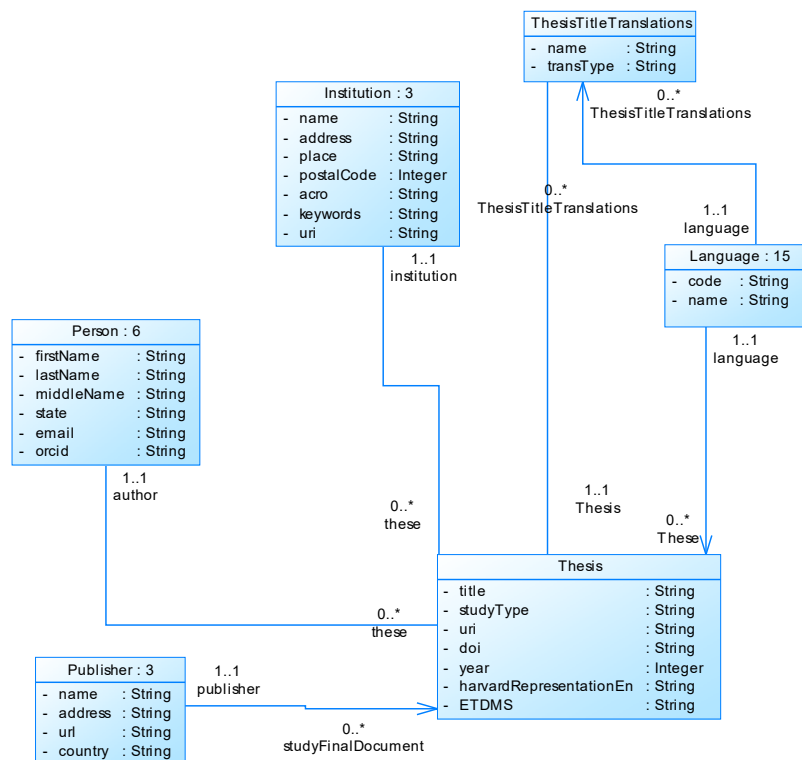




Figure 1. Data model

1. Design and implementation of a web page for searching digital library

Web page for searching is implemented using JSF and RichFaces frameworks for UI development. The web page supports basic and advanced search. Sorting and filtering results of the search by various parameters (publication dates, institutions, etc.) have been implemented. Moreover, controlled (authorized) downloading of PhD dissertations is also an option in user interface. More details about this user interface can be found in the paper (Ivanovic et al., 2013).

2. Techniques and well-known formats for textual and visual representation of digital textual files have been analyzed, adopted and implemented within PHD UNS

PhD dissertations' textual representation styles for three group of users have been defined: researchers (the Harvard citation style), librarians (MARC 21, Dublin Core, ETD-MS), and citizens (PhD metadata in HTML structured format – Figure 2). Moreover, word cloud stile for visual representation has been adopted, implemented and integrated within the PHD UNS library (Figure 3).

 GVOZDENAC, S. (2016) *Biological potential of cultivated plants in detection of water and sediment contamination*. (PhD dissertation), Faculty of Agriculture at Novi Sad 

Additional data	MARC 21	Dublin Core	ETD MS	Digital document
<p>GVOZDENAC, S. (2016) <i>Biological potential of cultivated plants in detection of water and sediment contamination</i>. (PhD dissertation), Faculty of Agriculture at Novi Sad</p> <p>Author data: First name: Соња Last name: Гвозденац Father name: Мирослав Fulfilled preconditions: Магистарске студије, 28/12/2000 Faculty: Пољопривредни факултет, Нови Сад</p> <p>Dissertation data: Title: доктор пољопривредних наука Defended on: 26/09/2016 Promotion date: 26/12/2018 Advisor: др Душанка Инђић, ред. проф. Пољопривредни факултет, Нови Сад Board members: др Сања Лазич, ред. проф. Пољопривредни факултет, Нови Сад, председник др Срђан Рончевић, ванр. проф. Природно-математички факултет, Нови Сад др Душанка Инђић, ред. проф. Пољопривредни факултет, Нови Сад, ментор и није члан</p>				

Figure 2. PhD metadata in the HTML structured format



Figure 3. Word cloud – visual representation of search results

3. Implementation of customization of search results representation through user interface

When search results are displayed, users can change results representation format, i.e., user can switch between representation styles shown in Figure 2 and Figure 3. PHD UNS stores a log message about changing representation style (Listing 1). Message logging have been implemented using the log4j library.

```
[INFO]23.04.2017. 11:47:31 (SearchDissertationsManagedBean:setRepresentationStyle)
Date and time: Sun Apr 23 11:47:31 CEST 2017| milliseconds: 1492940851164| + session
id: 8EBE86EC80D0539B2B75114ED8F17440| userId: 149294082881676| ip address:
78.30.151.18| location: city: Belgrade, postal code: null, regionName: null (region: 00),
```

```
countryName: Serbia (country code: RS), latitude: 44.818604, longitude: 20.468094| user
agent (device): Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML,
like Gecko) Chrome/51.0.2704.79 Safari/537.36 Edge/14.14393| new representation style:
htmlRepresentation
```

Listing 1. A log message about changing representation style

4. Log messages have been analyzed to determine the user's paternity behaviour based on query types and client device

Collected logs have been analyzed using ELK (ElasticSearch, LogStash, Kibana) stack technologies. Analysis shows that much more PHD UNS users prefer textual representation style than visual. Some users changed representation styles several times and it is assumed that different queries and type of results require different representation style. It was also found that for queries producing long lists of results, it is more transparent to see results in visual mode. Based on this, we can conclude that the ability to personalize the search results representation style is useful functionality and to improve the PHD UNS users' experience.

5. Making a module for collecting log messages after downloading a dissertation

The module for collecting download logs has been implemented using the log4j library. Besides basic information about downloaded dissertation, those logs also contain information about client machine which requested downloading (Listing 2). The collecting of logs was started in 2015 and the system collects log about more than 10,000 downloads per month.

```
[ INFO] 28.03.2017. 19:48:34 (FileManagerServlet:handleDownload)
Date and time: Tue Mar 28 19:48:34 CEST 2017| milliseconds: 1490723314534| + session id:
not defined| userId: 149072331453494| ip address: 207.46.13.110| location: city: Redmond,
postal code: 98052, regionName: Washington (region: WA), countryName: United States
(country code: US), latitude: 47.6801, longitude: -122.120605| user agent (device):
Mozilla/5.0 (iPhone; CPU iPhone OS 7_0 like Mac OS X) AppleWebKit/537.51.1 (KHTML, like
Gecko) Version/7.0 Mobile/11A465 Safari/9537.53 (compatible; bingbot/2.0;
+http://www.bing.com/bingbot.htm)| download record id: (BISIS)85285| file id: 1138| file
name: Karmela Filipovic doktorat.pdf| source: CRIS UNS| license: Attribution-NonCommercial
| record: Filipović Karmela, Glukozamin sulfat u lečenju osteoartroze kolena, Medicinski
fakultet u Novom Sadu, Univerzitet u Novom Sadu, 2014
```

Listing 2. A log message about downloading dissertation

6. Making a module for automatic recommendation for the certain user based on previously collected logs for that user and other PHD UNS users

We investigated with content based and collaborative based approaches for automatic recommendation. Approaches are language-independent and were built on download logs

collected in 2016 year. More information about this can be found in the paper (Azzopardi et al., 2016)

7. Evaluation of various approaches for automatic evaluation based on implicit feedback stored in download logs.

We used a set of download logs covering the period 20/2/2016 13:14:26 till 1/4/2016 08:50:53 to train the system, and a second set of download logs covering the period 2/4/2016 18:47:18 and 8/4/2016 10:49:59 to evaluate the system. We evaluated the 27 recommendation configurations. From the results, one can note that on the whole, collaborative based recommendation generally performs better than content-based recommendation. The best performing system is Collab-SimUsers that works on binary ratings using LSA and Pearson similarity, with a recall of approximately 42%.

8. Integration of the best approach within the PHD UNS library

The recommendation component with best result (see previous point) has been integrated within PHD UNS. It is put into operational mode on: <http://www.cris.uns.ac.rs/searchDissertations.jsf> as from October 2016. Considering that recommendation is performed on the basis of the previous download history of each user, the user models are updated every night taking into the account new downloads from the previous day.

9. Collecting explicit and implicit users' feedback

Icons with thumbs up and thumbs down have been used for collecting explicit users' feedback for recommended dissertations, while downloading of dissertations has been used as true positive implicit users' feedback.

10. Evaluation of the integrated module for automatic dissertations' recommendation

Evaluation of the integrated module for automatic dissertations' recommendation has been performed based on explicit feedback and the ELK stack technologies. The PHD UNS system has logged 172 positive user feedbacks and 24 negative user feedbacks for the first six months - starting from the moment that the feature that presents recommended dissertations was put in operational till April 2017.

11. Analysis of usage of PHD UNS digital library by non-academic citizens

An analysis of usage of PHD UNS digital library by non-academic citizens has been performed based on download logs and ELK stack technologies. Download logs used in the analysis cover the complete 2017 year and include more than 130.000 downloads. The analysis has shown that the most frequently downloaded dissertations deal with issues that are of interest for citizens outside the academic community: Physical activity of preschool children; agriculture; literature; history.

Conclusions

This paper presented a customization of PHD UNS web search pages for citizens out of academic

community. Analyses based on PHD UNS logs and ELK stack technologies have been conducted and have produced the following conclusions:

1. For queries producing long lists of results, it is more transparent to see results in visual (non-textual) mode
2. There is much more positive than negative explicit feedback for dissertations' automatic recommendation, thus this is useful functionality which improves the PHD UNS users' experience
3. The most frequently downloaded dissertations deal with issues that are of interest for all citizens (not only for the academic community): Physical activity of preschool children; agriculture; literature; history. Thus, PHD UNS has attracted users out of academy.

References

- Azzopardi, J., Ivanovic, D., & Kapitsaki, G. (2016). 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* (pp. 100–111). Springer, Cham. DOI: 10.1007/978-3-319-53640-8_9
- Bollen, J., & Luce, R. (2002). Evaluation of digital library impact and user communities by analysis of usage patterns. *D-Lib Magazine*, 8(6), 1–13.
- Bollen, J., Luce, R., Vemulapalli, S. S., & Xu, W. (2003). Usage analysis for the identification of research trends in digital libraries. *D-Lib Magazine*, 9(5), 1082–9873.
- Bordogna, G., Pagani, M., Pasi, G., & Villa, R. (2006, June). A flexible news filtering model exploiting a hierarchical fuzzy categorization. In *International Conference on Flexible Query Answering Systems* (pp. 170–184). Springer, Berlin, Heidelberg.
- Chau, M., Fang, X., & Liu Sheng, O. R. (2005). Analysis of the query logs of a web site search engine. *Journal of the American Society for Information Science and Technology*, 56(13), 1363–1376.
- Das, A. S., Datar, M., Garg, A., & Rajaram, S. (2007, May). Google news personalization: scalable online collaborative filtering. In *Proceedings of the 16th international conference on World Wide Web* (pp. 271–280). ACM.
- Dimić-Surla, B., & Ivanović, D. (2012). Software component for reporting in the CRIS systems. In *Proceedings of the CRIS 2012 Conference, Prague, June 6–9* (pp. 61–66).
- Dimić Surla, B., Segedinac, M., & Ivanović, D. (2012). A BIBO ontology extension for evaluation of scientific research results. In *Proceedings of the Fifth Balkan Conference in Informatics* (pp. 275–278). ACM, DOI: 10.1145/2371316.2371376.
- Ivanović, D. (2010). A scientific-research activities information system. PhD dissertation, Faculty of Technical Sciences, University of Novi Sad

- Ivanović, D., Milosavljević, G., Milosavljević, B., & Surla, D. (2010). A CERIF-compatible research management system based on the MARC 21 format. *Program: Electronic library and information systems*, 44(3), 229–251, DOI: 10.1108/00330331011064249.
- Ivanović, D., & Milosavljević, B. (2010). Software architecture of system of bibliographic data. In *Proceedings of the XXI Conference on Applied Mathematics PRIM (Vol. 2009, pp. 85–94)*.
- Ivanovic, D. (2011). Data exchange between CRIS UNS, institutional repositories and library information systems. In *Proceedings of 5th International Quality Conference, Kragujevac, May 19–21 (pp. 371–378)*.
- Ivanović, D., Surla, D., & Konjović, Z. (2011a). CERIF compatible data model based on MARC 21 format. *The Electronic Library*, 29(1), 52–70, DOI: 10.1108/02640471111111433.
- Ivanović, D., Surla, D., & Racković, M. (2011b). A CERIF data model extension for evaluation and quantitative expression of scientific research results. *Scientometrics*, 86(1), 155–172, DOI: 10.1007/s11192-010-0228-2.
- Ivanović, L., Ivanović, D., & Surla, D. (2012a). A data model of theses and dissertations compatible with CERIF, Dublin Core and EDT-MS. *Online Information Review*, 36(4), 548–567.
- Ivanović, L., Ivanović, D., & Surla, D. (2012b). Notes on operations: Integration of a Research Management System and an OAI-PMH Compatible ETDs Repository at the University of Novi Sad, Republic of Serbia. *Library Resources & Technical Services*, 56(2), 104–112.
- Ivanović, D., Surla, D., & Racković, M. (2012c). Journal evaluation based on bibliometric indicators and the CERIF data model. *Computer Science and Information Systems*, 9(2), 791–811, DOI: 10.2298/CSIS110801009I
- Ivanović, L., Dimić-Surla, B., Segedinac, M., & Ivanović, D. (2012d). CRISUNS ontology for theses and dissertations. In *Proceedings of the ICIST 2012 Conference (CD), Kopaonik, February 29 (pp. 164–169)*.
- Ivanovića, L., & Surla, D. (2012). A software module for import of theses and dissertations to CRISs. In *Proceedings of the CRIS 2012 Conference, Prague, June 6–9 (pp. 313–322)*.
- Ivanović, L. (2013). Search of catalogues of theses and dissertations. *Novi Sad J. Math*, 43(1), 155–165.
- Ivanović, L., Ivanović, D., Surla, D., & Konjović, Z. (2013). User interface of web application for searching PhD dissertations of the University of Novi Sad. In *2013 IEEE 11th International Symposium on Intelligent Systems and Informatics (SISY) (pp. 117–122)*. IEEE.
- Jones, S., Cunningham, S. J., & McNab, R. (1998). Usage analysis of a digital library. In *Proceedings of the third ACM conference on Digital libraries, May 1998 (pp. 293–294)*. ACM.
- Kovačević, A., Ivanović, D., Milosavljević, B., Konjović, Z., & Surla, D. (2011). Automatic

- extraction of metadata from scientific publications for CRIS systems. *Program: electronic library and information systems*, 45(4), 376–396, DOI: 10.1108/00330331111182094.
- Kuo, B. Y., Hentrich, T., Good, B. M., & Wilkinson, M. D. (2007). Tag clouds for summarizing web search results. In *Proceedings of the 16th international conference on World Wide Web* (pp. 1203–1204). ACM.
- Milosavljević, G., Ivanović, D., Surla, D., & Milosavljević, B. (2011). Automated construction of the user interface for a CERIF-compliant research management system. *The Electronic Library*, 29(5), 565–588, DOI: 10.1108/02640471111177035.
- Nikolić, S., Penca, V., Ivanović, D., Surla, D., & Konjović, Z. (2012). CRIS service for journals and journal articles evaluation. In *Proceedings of the 11th International Conference on Current Research Information Systems*, Prague, Czech Republic (pp. 323–332).
- Penca, V., Ivanović, D., Surla, D., & Konjović, Z. (2012). Development of a Unified Search Profile for CRIS Systems. In *Proceedings of the 6th International Conference on Methodologies, Technologies and Tools Enabling e-Government*, Belgrade, July 3–5, 2012 (pp. 1–10).
- Scanfeld, D., Scanfeld, V., & Larson, E. L. (2010). Dissemination of health information through social networks: Twitter and antibiotics. *American journal of infection control*, 38(3), 182–188.
- Schafer, J. B., Frankowski, D., Herlocker, J., & Sen, S. (2007). Collaborative Filtering Recommender Systems. In *The adaptive web* (pp. 291–324). Springer, Berlin, Heidelberg.
- Sisodia, D. S., & Verma, S. (2012). Web usage pattern analysis through web logs: A review. *Computer Science and Software Engineering (JCSSE)*, 2012 International Joint Conference, May 2012 (pp. 49–53). IEEE.
- Surla, D., Ivanović, D., Konjović, Z. & Racković, M. (2012). Rules for evaluation of scientific results published in scientific journals. *Management Information Systems*, 7(3), 3–10.
- Zhang, Y., Lee, K., & You, B. J. (2001). Usage patterns of an electronic theses and dissertations system. *Online Information Review*, 25(6), 370–378.