Seminar on Providing Access to Grey Literature 2013: The 6th year of the seminar focused on storage and providing access to the grey literature, 23th October 2013 [online]. Praha: National Library of Technology, 2013. Available at WWW: http://nusl.techlib.cz/Sborniky. ISSN 1803-6015.

CENTRAL REGISTRY OF THESES AND DISSERTATIONS AND THE ANTI-PLAGIARISM SYSTEM AS A COMPREHENSIVE SOLUTION AT NATIONAL LEVEL

JMÉNO A PŘÍJMENÍ (BEZ TITULŮ)

juraj.noge@cvtisr.sk, marta.duskova@cvtisr.sk

Slovak Centre of Scientific and Technical Information, Slovakia

Abstract
The objective of this paper is to present the Central Registry of Theses and Dissertations in the SR and a document originality checking system as a comprehensive solution at national level for higher education institutions, the objectives as well as the starting point of the project, the resolution of organisational and legislative problems, the technical solutions, implementation, and the experience from almost four years of operation. We present some statistical data, including the system benefits and citations, as well as an outline of the future development and use of the system in the future. I will also concisely mention other centralised systems operated by the SCSTI SR.

Keyword
The Central Registry of Theses and Dissertations, CRTD, control of the originality of documents, collection of theses and dissertations, anti-plagiarism system, Antiplag, theses and dissertations records system, academic information system, AIS, comparison corpus, Slovak Centre of Scientific and Technical Information, document storage space

INTRODUCTION

The Slovak Centre of Scientific and Technical Information is a public institution, and so its constitution covers a wide spectrum of activities it should perform, tasks it should fulfil, and services it should provide primarily for the scientific/research community. The name itself indicates that from the national perspective it should cover them in a central manner, which today is unthinkable without the support of information systems – and it is precisely the design, implementation and operation of central information systems that has become one of the key activities of our institution in recent years. Several systems have already been in operation for some years, while others are still in the development phase. This means, for example, SK CRIS – the Current Research Information System, the Central Registry of Publication Activity (CREPČ) and the Central Registry of Evidence of Art Works and Performance (CREUČ), which I will concisely mention below, as well as the CRTD and the APS, which I will focus on in more detail. We are also preparing other systems – for example the Integrated System of Services (ISS) or the digital repository.

CREPČ/CREUČ - Central Registry of Publication Activity / Central Registry of Evidence of Art Works and Performance**.**

The CREPČ/CREUČ systems originated as a development project of the Ministry of Education of the Slovak Republic in 2007 and 2008. Their objective is the comprehensive bibliographical registration of publication and artistic activities of higher education institutions. The contributors are all public and state-run higher education institutions and selected private higher education institutions; at the present time totalling 31 higher education institutions. The task of the SCSTI SR is the administration and verification of the data contained in both systems.

Thanks to legislative support it is possible to build systems at nationwide level. The CREPČ/CREUČ are enshrined in the following legislative documents:

- Act on Higher Education Institutions No 455/2012,

- MESRaS SR Decree No 456/2012 on a Central Registry of Publication Activity and a Central Registry of Evidence of Art Works and Performance,

- Ministry of Education SR Directive No 13/2008-R on Bibliographic Registration and Categorisation of Publications, Artistic Activities and Citations.

The systems contain data for 2007-2013. CREPČ contains over 315 00 records, CREUČ around 13 500 records. The data are publicly available through searches directly through the portal: www.crepc.sk.

The primary objective was that the systems serve the Ministry of Education SR for the calculation of the annual state subsidy to public higher education institutions, however today it also functions as:

- a comprehensive overview of the publication behaviour and artistic output of employees at higher education institutions in the SR,

- an online information resource for mapping publication and artistic outputs in the SR, biographical research,

- statistical overviews, overviews of the profile and performance of scientific/research and pedagogical workplaces at higher education institutions or individuals,

- an information resource for grey literature produced at higher education institutions.

THE CRTD AND APS - THE CENTRAL REGISTRY OF THESES AND DISSERTATIONS AND THE SYSTEM FOR DETECTING PLAGIARISM (ANTI-PLAGIARISM SYSTEM)

The CRTD and APS are two closely cooperating systems. The CRTD serves as a central storage space for the long-term storage of bachelor, diploma, doctoral and qualification postdoctoral publications, and the APS provides checks of their originality. They are characterised in particular as follows:

- all documents from higher education institutions in Slovakia operating according to the body of laws of the Slovak Republic must be sent to the Central Registry (where they are stored for 70 years) and subject to an originality check before their defence,

- the result of the originality check is the Originality Report, intended to support higher education institutions in their decision-making in matters of plagiarism,

- the CRTD contains almost 300 000 documents; every new document is compared with documents from the CRTD and with more than five million documents primarily from internet sources,

- documents sent to the register since 1/9/2011 are also published on the crzp.sk portal

- the CRTD has become a digital repository of grey literature.

HOW IT BEGAN

Since 2000 it has been clear to all those involved in Slovakia that it would be necessary to resolve the issue of growing plagiarism in a decisive manner. It is a fact that plagiarism was most evident and remains most evident in education, which is caused in particular through the rapid development and availability of information/communications technology for practically everybody, and also the dramatic growth in content available through the internet. Another not negligible factor has been the rapid growth in the number of higher education institutions and students: between 1989 and 2012 the number of higher education institutions grew from 13 to 40 and the total number of higher education institution students almost quadrupled.

The first institution to check the originality of documents was School of Management in Trenčín in 2001, followed in 2008 by the University of Economics in Bratislava, and in 2009 by Comenius University in Bratislava (Kravjar, 2013).

In spite of this it has not been possible to adequately resolve the situation on a nationwide basis. The Slovak Rectors' Conference (SRK) attempted to change this situation in 2006 through the adoption of a plenary's opinion on plagiarism entitled "Measures to eliminate plagiarism in the processing and presentation of bachelor's, master's and dissertation theses" and also through the adoption of the "Code of ethics for higher education institution employees". However, not even these documents impacted the existing situation in practice and in principle the declared objective was not fulfilled.

In 2008 the Ministry of Education of the SR decided that it would implement a comprehensive solution at national level covering the collection, processing and control of originality of specified documents. Higher education institutions in Slovakia became obligated to use this comprehensive solution and an Originality Report will be a condition for the release of a document for defence. Necessary legislative support in the form of an amendment to the Act on Higher Education Institutions in 2009 significantly contributed towards the successful implementation of the stipulated objective.

Seminar on Providing Access to Grey Literature 2013: The 6th year of the seminar focused on storage and providing access to the grey literature, 23th October 2013 [online]. Praha: National Library of Technology, 2013. Available at WWW: http://nusl.techlib.cz/Sborniky. ISSN 1803-6015.

BASIS FOR THE APS PROJECT

In around mid-2009, at the beginning of the resumption of the CRTD/APS, the approach was based around the existence of the following:

- a clear opinion of representatives of higher education institutions regarding the prevention of the spread of plagiarism and the need to address the situation,

- a decision by the Ministry of Education of the SR on the implementation of a comprehensive solution,

- a change in legislation in the form of an amendment to the Act on Higher Education Institutions,

- the existence of the study "Prevention and detection of plagiarism" (Skalka et al, 2009a),

- the CRTD solution already under preparation at the Constantine the Philosopher University in Nitra,

- a requirement from the Ministry of Education of the SR to commence the collection of documents as early as in the 2009/2010 academic year,

- the unresolved method of financing of the APS,

- the incomplete analytical materials of the solution,

- unknown responsible APS solver,

- unknown APS contractor,

During the 2nd half of 2009 things moved forward. The Constantine the Philosopher University in Nitra focused on completing the document on the central storage space for theses and dissertations. The SCSTI SR was authorised to cover the organisational aspects of the project and to implement the anti-plagiarism superstructure. The Ministry of Education released funds for the pre-financing of the project. A software supplier was selected through a public tender. A quality team composed of local as well as external workers was created at the SCSTI SR, and tasked with rapidly preparing the technical infrastructure, methodologies, and the organisational coordination of all the entities involved, implementing the CRTD information system with the anti-plagiarism superstructure, and ensuring its operation.

STAGES OF THE SOLUTION

The project was implemented over several stages. The initial analytical/conceptual/organisational stage was of key importance in particular, and was implemented under the auspices of the Constantine the Philosopher University in Nitra. It addressed the central storage space for all theses and dissertations. The company SVOP, spol. s r.o. was selected as the external supplier of the system. During this stage and in cooperation

with the supplier, the initial state in terms of the collection of theses and dissertations was identified, and a conceptual solution proposed, including legislation, the issue of copyright and licencing contracts for documents, the financial and organisational securing of the project throughout its area of effect, the identification of logistics and impacts on academic information systems, and so on. The results of the mapping of the initial situation have, for example, also shown that the majority of higher education institutions already collect theses and dissertations in electronic form and place them in their academic information systems (AIS) and/or libraries. As the issue of the collection of final (bachelor, diploma, doctoral) and qualification (postdoctoral and habilitation) documents is the subject of several laws (Act on Higher Education Institutions, Act on Libraries, Copyright Act, ...), it was necessary in this stage to prepare, in cooperation with the Ministry of Education of the SR, Methodological Regulations in such a way that after the creation of the system higher education institutions would be obligated to send theses and dissertations to the CRTD (a plagiarism check is a condition for passing a document on for defence) (Noge, 2011). In addition, the issue was also addressed of defining the requirements on higher education institutions vis-à-vis the CRTD and an estimate of the costs for the modification of their local information systems, the issue of concluded licencing contracts with authors, an amendment to the contract between the Ministry of Education of the SR and the operator of the CRTD/APS, and the preparation of a generally binding legal regulation as well as a concept for the technical solution itself.

The second stage, which in terms of scheduling partially overlapped the activities of the first stage, was the implementation of the CRTD. This stage included the preparation of a technical design for the solution, software development, the construction of technical infrastructure, the installation of the necessary SW at the operator, testing work etc. This work was mainly carried out by an external contractor. During this stage the Ministry of Education of the SR decided to change the operator from the Constantine the Philosopher University in Nitra to an organisation directly managed by the Ministry – the SCSTI SR, based in Bratislava. Part of this stage, which took place in 2009, was also the construction of the required infrastructure at the higher education institutions and the testing of the functionality of the communication interface between the CRTD and AIS. A significant part of this activity in this stage was also the preparation of instructions, explanations, etc. related to the adopted legal norm, and the necessary modification of the systems at the schools in connection with the export of documents to the CRTD.

Also in 2009, once again upon the instigation of the Ministry of Education of the SR, it was decided that the constructed CRTD would also serve as the comparison corpus for the document originality checking application – the so-called anti-plagiarism system (APS). The Ministry of Education of the SR released funds for this purpose, which we can consider as the third stage of the solution, and entrusted the SCSTI SR with the provision of an anti-plagiarism superstructure for the CRTD. After a complex public tender process, the SCSTI SR eventually signed a contract with a contractor for the CRTD registry – the company SVOP, spol. s r. o. This was intended to simplify the initial situation primarily because it perfectly understood the system for which the APS was to be the superstructure. The SCSTI SR took a risk in particular because the contractor's bid was with an original, but not verified, solution, which it had never previously implemented in practice. However, the delivery date

was complied with and the document originality checking system was put into full operation at the end of April 2010, meaning at the start of the bulk collection of theses and dissertations for the 2009/2010 academic year.

The next stage will be the migration of the CRTD/APS system to the newly constructed Data Centre for Research and Development (DC RD) constructed with funds from the European Regional Development Fund through the Operational Programme Research and Development. The modern, secure and reliable ICT equipment will provide excellent conditions for operating a system that is demanding in terms of sufficient data capacity and, above all, during the data collection period, demanding in terms of computing power. During this period, after an amendment to the Act on Higher Education Institutions, documents stored in the CRTD were implemented and made available to the public (from 1/9/2011).

THE DOCUMENT COLLECTION AND VERIFICATION PROCESS

The actual storage of the theses and dissertations in the CRTD is preceded by the collection process at higher education institutions. During this process the finished theses and dissertations are converted into the required format, furnished with the required metadata, and the parameters of the licencing agreement for full-text access are defined. Works prepared in this way are stored in the Electronic Theses and Dissertation (ETD) system of the higher education institution and there, placed into batches, await the instruction to be copied to the CRTD. After the request/instruction, the metadata of the documents describing the theses and dissertations is transferred through the data interface to the CRTD, where information is extracted about the location of the documents (theses and dissertations). Based on this, the theses and dissertations files are automatically uploaded from the ETD to the data storage space of the CRTD (Skalka et al, 2009b).

Subsequently, on the basis of a request from the higher education institution, a plagiarism check is launched on the received batch of theses and dissertations against the comparison corpus created from documents previously sent and stored, respectively from other sources (e.g., from the internet). The results of the plagiarism comparison are then linked to the corresponding theses and dissertations and the system automatically sends them back to the higher education institution in question. The actual texts of the documents are incorporated into the comparison corpus of the anti-plagiarism system (Image 1)

The delivery of the comparison results to the higher education institution only forms the basis for the assessment and evaluation of the document by the relevant commission. This evaluates whether or not plagiarism is involved (for example the document actually only contains correctly attributed quotations) and the commission in question always has the last word during the defence.

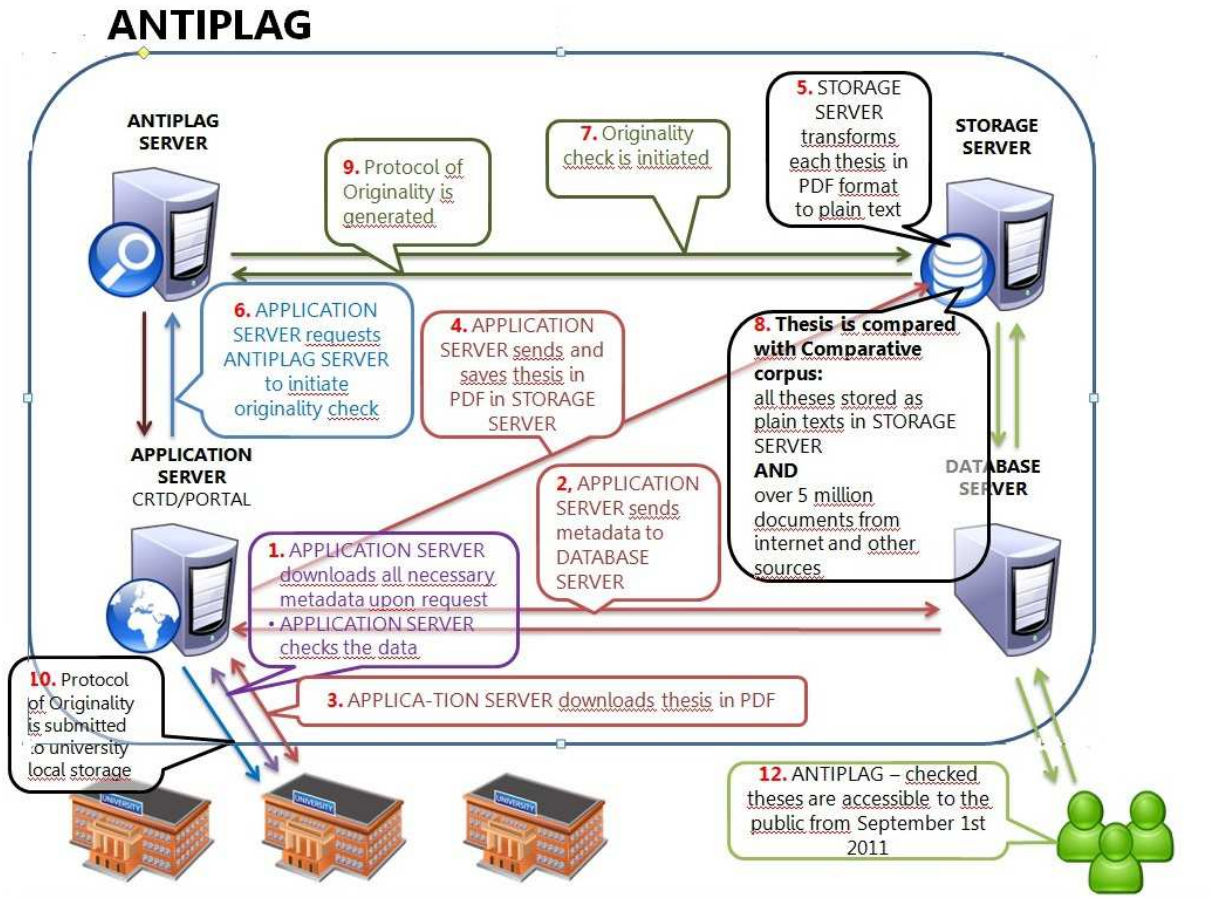

Image 1 The document collection and verification process

CONCISE DESCRIPTION OF THE TECHNICAL SOLUTION

Technically speaking, the CRTD and APS systems are constructed as a system of cooperating servers:

- The application server fulfils the function of communication server visible on the internet, provides portal services and full-text downloading,

- The storage server is a universal storage space for the original files and also plain text files, output reports, logs and so on,

- The database server serves for the operation of MS SQL databases,

- The Antiplag server contains the indexation and scanning core of the algorithm for detecting plagiarism (the number of these servers will be increased in step with the number of documents and the comparison corpus),

- The actual anti-plagiarism system is a set of several applications and original algorithms created by SVOP, spol. s r. o. and forms part of its know-how. They have been designed as an agent system and are launched dynamically as required.

EXPERIENCE WITH OPERATION

From our perspective the operation of the system has 2 dimensions – organisational and technical. From the operational perspective it was necessary, particularly at the start, to ensure cooperation with the administrators of the local storage spaces at the higher education institutions. We provide users with year-round support in all aspects, and at the present time can say that there are no significant problems. Cooperation with the system administrator, the Ministry of Education of the SR, is also good.

From the technical perspective we provide continuous 24x7 operation. The delivered system is reliable and works without outages, in particular after migration to the DC RD environment –information is secure and the capacity is sufficient. The seasonal nature of the addition of documents and the parallel plagiarism comparisons – see Image 2 – is also provided for. We are capable of meeting the legally stipulated deadline for the delivery of the Originality Report – within 48 hours – with a reserve, even when up to 5,000 documents are received a day at peak times.

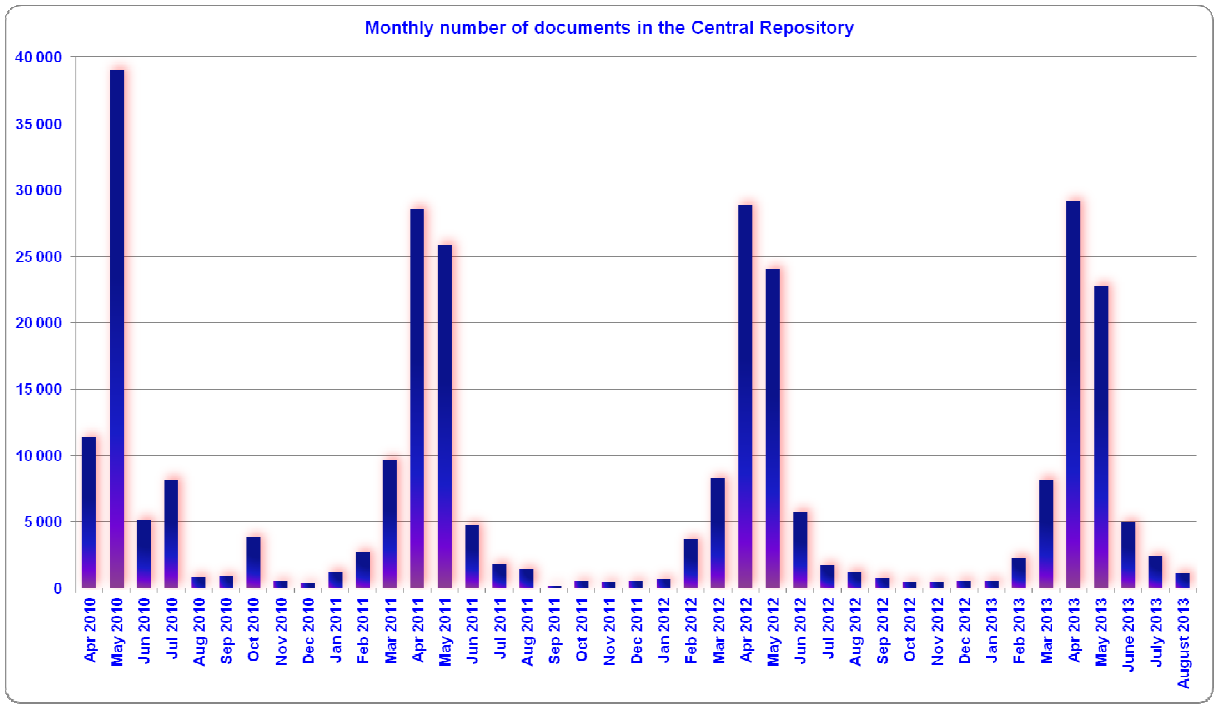

Image 2 Monthly increases in the number of documents in the CRTD

At the present time 33 higher education institutions send theses and dissertations to the CRTD/APS system.

Around 75,000 theses and dissertations are added to the CRTD annually, and it currently stores around 300,000. They take up around 600 GB of storage space. Over 5 million documents, representing around 2 TB of data, have been downloaded from the internet for the plagiarism comparisons.

A comparison of one document against today's corpus, including the processing of metadata and the generation of a PDF record, takes on average 10 seconds. The system even coped with the maximum daily inflow of documents, 4,900 files over 10 hours, without a problem.

BENEFITS AND CITATIONS (AWARDS)

The publication of the fact that the CRTD/APS was being put into operation already had a positive preventative/behavioural effect. Both students and pedagogues have realised that their document will be checked for plagiarism and have begun approaching the creation of written work and work with literature more seriously. The general obligation to send work into the system has established a level playing field for all higher education institutions.

Citations from the media regarding the launching of the CRTD/APS all agree that it has contributed towards:

- making the behaviour of both students and pedagogues more responsible,

- increasing students' independence during the document creation process,

- increasing the quality of the work,

- more thorough citations.

The CRTD/APS was awarded in a competition for the best product in the digitisation of public administration at the ITAPA 2011 conference, where the SCSTI SR was ranked second in the new services category.

In a Europe-wide competition announced by the European Commission in 2013 "European Prize for Innovation in Public Administration" the SCSTI SR was successful in the category "Initiatives in Education and Research" with the CRTD/APS project, winning first prize in "Initiatives in Education and Research". We accepted the prize of € 100,000 directly from EU Commissioner for Research, Innovation and Science, Máire Geoghegan-Quinat, at the gala prize awards in Cork (Ireland) as part of the Week of Innovative Regions Europe Conference (WIRE IV).

FURTHER DEVELOPMENT PLANS

The perspectives for the further development of this system are relatively extensive on several levels. They are a reaction to external influences and efforts to adopt new technological trends and possibilities:

- In the short term we are planning the implementation of the winning algorithm for originality checks from the international competition in Amsterdam - PAN 2011 Lab (Evaluation, 2011). This algorithm not only won overall, but also in each of the four monitored indicators. One of these indicators was the detection of translation plagiarism.

- We are planning to expand the number of licences for the Antiplag server, which will further speed up the response time for users.

- There will also be a web version of the Originality Report available that will show, unlike the paper report, all the test matches that ANTIPLAG detected, not only the matches from the first fifteen documents with the highest percentage of over-the-limit matches (this is a limitation of the paper version of the report). One unique feature of the web-based report will be the possibility (for the reviewer) to establish the exact percentage of matches in the text of the controlled document with the unattributed "plagiarised" text found in the document.

- An expansion of the comparison corpus to include documents in Czech from internet sources.

- Preparation for comparisons of the documents in foreign languages (primarily in English) directly against the internet source.

- There is also a plan to expand the CRTD/APS system to provide services to other institutions than higher education institutions. There is also consideration of, for example, checks of documents that are the outputs of research and development projects, or documents that form part of projects financially supported from public sources. This, however, must firstly be enshrined in legislation.

CONCLUSION

During the amendments to the Act on Higher Education Institutions, no thought was given to feedback. In addition, the higher education institutions do not have any obligation to publish data such as how many cases of suspicion of plagiarism have been investigated; how many of them were positive findings; and what sanctions were applied against the plagiarists. This means that we do not have available precise data to evaluate how the numbers of attempts at plagiarism changed after the blanket application of the system.

A positive side-effect of the implementation of the CRTD (on which no emphasis was placed when the project was put into operation) is, from the perspective of the processing of grey literature, the existence of a semi-complete (only documents from 2010) and convenient specialised storage space for grey literature (theses and dissertations). The CRTD is available at: www.crzp.sk

Seminar on Providing Access to Grey Literature 2013: The 6th year of the seminar focused on storage and providing access to the grey literature, 23th October 2013 [online]. Praha: National Library of Technology, 2013. Available at WWW: http://nusl.techlib.cz/Sborniky. ISSN 1803-6015.

References

- Skalka, Ján a kol.: Prevencia a odhaľovanie plagiátorstva : zber prác za účelom obmedzenia porušovania autorských práv v kvalifikačných prácach na vysokých školách. Nitra : UKF, 2009. 126 s. ISBN 978-80-8094-612-8. Dostupné na internete: http://www.crzp.sk/dokumenty/prevencia_odhalovanie_plagiatorstva.pdf [Online] [Dátum: 8.1. 2011.]
- Skalka, Ján; Vozár, Libor; Drlík, Martin; Grman, Ján: Centrálny register záverečných prác a metodika ich zberu. In ITlib. Informačné technológie a knižnice [online], 2009, č. 04 [cit. 2009-12-14]. Dostupné na internete <http://www.cvtisr.sk/itlib/itlib094/cr_zaver_prace.htm>. ISSN 1336-0779. [Online] [Dátum: 16. 4 2011.]
- NOGE, Juraj. Central register of theses and dissertations in Slovakia and document originality verification as a centraly provided service. In: ProInflow : časopis pro informační vědy [online]. - 2011, roč. 3, č. 2 , s. 110-120. - ISSN 1804-2406. Dostupný na: http://pro.inflow.cz/central-register-theses-and-disserations-slovakia-and-document-originality-verification-centrally-pr.
- KRAVJAR, Július. CRZP/APS : míľniky, aktuálny stav, pripravované zmeny. In: INFOS 2013. Zborník príspevkov z 37. medzinárodného informatického sympózia Inovatívne knižnice a pamäťové inštitúcie. Hľadanie odpovedí na nové výzvy znalostnej spoločnosti. Stará Lesná, 8. – 11. apríla 2013. - Bratislava : Spolok slovenských knihovníkov, 2013. - ISBN 978-80-89586-07-3. - S. 104-116. Dostupné na internete: http://www.infolib.sk/files/Novy_portal_infolib_subory/janka_nemethyova/infos_2013/infos_2013_zbornik.pdf
- Kravjar, Július; Noge, Juraj : Strategies and Responses to Plagiarism in Slovakia. In: Zborník príspevkov International Conference Plagiarism across Europe and Beyond, projektu IPPHEAE, Brno, 12. – 13. Júna 2013 – BRNO: Mendelova Univerzita Brno, 2013. ISBN 978-80 7375 765-6. - S. 201-215. Dostupné na: http://ippheae.pefka.mendelu.cz/files/proceedings.pdf.

- Evaluation Results. 2011 [ONLINE] Available at: http://www.uni-weimar.de/medien/webis/research/events/pan-11/pan11-web/plagiarism-detection.html#results. [Accessed 12.9.2013].