Juraj Noge

CENTRAL REGISTER OF THESES AND DISSERTATIONS IN SLOVAKIA AND DOCUMENT ORIGINALITY VERIFICATION AS A CENTRALLY PROVIDED SERVICE

Zajímavosti z oboru

Abstrakt:
Obsahom príspevku je stručný popis okolností vzniku Centrálneho registra záverečných prác v SR a systému na kontrolu originality dokumentov ako komplexného riešenia na národnej úrovni, cieľov i východísk projektu, krátky popis riešenia i implementácie systému, skúseností z prevádzkovania. Na záver je uvedených niekoľko štatistických údajov, zhrnutie prínosov a ohlasov poskytovanej služby ako i naznačenie ďalšieho možného rozvoja a využitia systému v budúcnosti.

Klíčová slova: *Centrálny register záverečných prác (CRZP), kontrola originality dokumentov, plošný zber záverečných a kvalifikačných prác (ZP), antiplagiátorský systém (APS), systém evidencie záverečných prác (EZP), akademický informačný systém (AIS), porovnávací korpus, Centrum vedecko-technických informácií SR (CVTI SR), Internet, úložisko dokumentov, Dátové centrum pre výskum a vývoj (DC VaV)*

Abstract:
The paper deals with the brief description of situation associated with creation of the Central Register of Theses and Dissertations in the Slovak Republic and with a system for document originality verification as a complete solution on the national level, with aims and background of the project, with short description of solution and implementation of the system as well as experience from its operation. There are some statistic data, analysis of benefits and feedback to the delivered service, outline for further development and system applications in the future are given in the final part of the paper.

Keywords: *Central Register of Theses and Dissertations (CRTD), document originality verification, non-point collection of theses and dissertations (TD), antiplagiarism system (APS), system of theses and dissertations evidence (ETD), academic information system (AIS), comparative corpus, Slovak Centre of Scientific and Technical Information (SC STI), Internet, document repository, Data Centre for Research and Development (DCRD)*

1 Introduction

To analyse now in deep the phenomenon of plagiarism, i.e. unauthorized adoption and/or imitation of intellectual activity of anybody else, seems to make no sense in this place. It is a fact, however, that with rapid development and accessibility of information and communication technologies to almost anybody, as well as with dramatically increased contents which is open via the Internet, the plagiarism in our country is easier than ever. The most evident examples can be found in the area of education where a part of students is misusing the possibility to attain the whole elaborated texts easily and without their own contribution to works.To disclose and combatthis undesirable phenomenon is quite demanding and hard social task.

The problem of plagiarism of theses and dissertations is medially probably best known, but at the same time really the most harmful one. The authors – plagiarists – are enriched with undeserved professional, social and often also financial benefits. In the case of disclosure, on the other hand, the plagiarists cause loss of credibility of the institution enabling them to defend such a work. The applied practices to solve this problem at individual colleges and universities were really insufficient and low effective. Consequently in 2008,the complete solution on the national level was started to be found from the initiative of the Ministry of Education of the Slovak Republic. To design and implement the central repository for all theses and dissertations completed as outputs of study at Slovak colleges and universities with guarantee of their long-term and safe storage was the primary goal of this initiative. The primary goal was followed by the idea of using this central repository as a comparative corpus to verify originality of all documents and making it available to all Slovak colleges and universities.

2 Stages of Solution

The solution was obtained in several stages. The initial, analytical , conceptual and organisational stage implemented under the auspices of Constantine the Philosopher University in Nitra, was the crucial stage. It was focused on repository for all theses and dissertations. The SVOP Ltd. company has been selected as an external supplier of the system.During this phase, the initial state in the field of collecting theses and dissertations has been defined in co-operation with the supplier, then the solution concept has been designed in which legal aspects, copyright, licencing agreements, financial and organisational arrangments for the project as a whole were concerned , logistics identified and influence on academic information systems determined. The results obtained in mapping the initial situation have shown that theses and dissertations are collected in electronic form and stored in local academic information systems by majority of colleges and universities in Slovakia. As the topics

of collecting graduate (bachelor, diploma) and degree(rigorous, dissertation and habilitation) works concerns many legislation aspects (University Law, LibraryLaw, Copyright Law, etc.) it was necessary in this stage to prepare legislation measures in the form of issued Methodological Directions in co-operation with Ministry of Education of the Slovak Republic. As a result, after building the system, the graduate schools are obliged to send theses and dissertations to the central register in order (to detect for plagiarism is the prerequisite to admit the work for defence). Among issues to be solved the following issues can be mentioned: the burning issue of defining requirements concerning the CRTD for universities; estimation of expenditures for adaptation of local academic information systems (AIS); the issue of concluding agreements with authors;modification of the agreement between the Ministry and register provider;modification of the methodological directions for universities and preparation of generally obligatory legislation and the concept of the sole solution.

The second stage, the activities of which markedly interfered with those of the first stage, included implementation of the CRTD. It was necessary to develop technical details of the solution, program equipment, technical infrastructure, to instal the proper SW equipment at the provider and at college/ university workplaces, to perform tests, etc. During the second stage the Ministry of Education of the Slovak Republic (MESR) has decided to change the provider. The Slovak Centre of Scientific and technical Information (SC STI) – the institution directly managed by the MESR, with its headquarters in Bratislava, became the new provider. Building of the required infrastructure on the side of colleges and universities and testing of functions of the transfer interface between the central register and AIS created the integral part of this stage running in 2009. To prepare instructions, explanations and similar activities associated with the adopted legal standard and the forced arrangement of the systems at colleges and universities in the context of export of works into the central register - was the fundamental part of activities in this stage.

As early as in 2009, from the initiative born again at the MESR, it has been decided that the built CRTD should serve also as a comparative corpus to apply and verify documents for originality – the s.c. anti-plagiarism system (APS). The MESR has granted funds for this purpose, which can be referred to as the third stage of the solution. At the same time the SC STI was asked to provide antiplagiarism superstructure over the CRTD. After completion of market research and determining the requirements on such a sophisticated system, the public procurement (PP) has been announced. The competition documents were taken by nine potential suppliers. However, due to hard conditions (mainly due to short time asit was intended to compare final works for originality already in the 2009/2010 Academic year, due to quite limited funds and difficulties of the Slovak language), no direct offer from any

of the interested part was submitted. In the subsequent direct negotiations conducted with the three selected suppliers (Masaryk University in Brno being one of them), the SC STI has finally conluded Agreement with SVOP Ltd. company – i.e. the CRTD register supplier. As the system to be superstructured by APSwas well-known in details to this company, the initial position of SVOP Ltd. company was faciliated. The SC STI, as theprovider, took a risk because the supplier has offerred original, but till now not verified solution, never applied to practice. After signing the Agreement, the time of delivery was observed and the system for document originality verification was put into operation by the end of April 2010, i.e. at the beginning of massive collection of theses and dissertationsin 2009/2010 Academic year.

3 The Process of Collecting and Verifying the Works

As a result of a survey performed within analytical procedures in 2009, approximately 75 % of colleges and universities were already collecting electronic versions of works and majority of others was preparing to do so. The conditions and rules had to be assessed in such a manner that the works completed at colleges and universities are stored in the required form in the CRTD to comply with legislation in force. The works were, however, stored from local systems of evidence of theses and dissertations (ETD), other individual SW application or by means of special Web interface. This local repository should be equipped with interface compatible with the CRTD interface through which data between the CRTD and the ETD are transferred.

The proper deposition of theses and dissertations in the CRTD is preceeded by collection process performed at colleges and universities. During this process, the completed thesis or dissertation is transformed to the required format, equipped with defined metadata and the parameters of Licence agreement concerning open access to full texts are defined. Theses and dissertations prepared in this way are stored in the ETD of the college or university and arranged to batches wait for the order to be copied into the CRTD. After obtaining order/request, metadata are tranferred through interface, where information about location of documents of i.e. thesis or dissertation are extracted. Based on this information files with thesis or dissertation are downloaded from the ETD to the CRTD data repository.

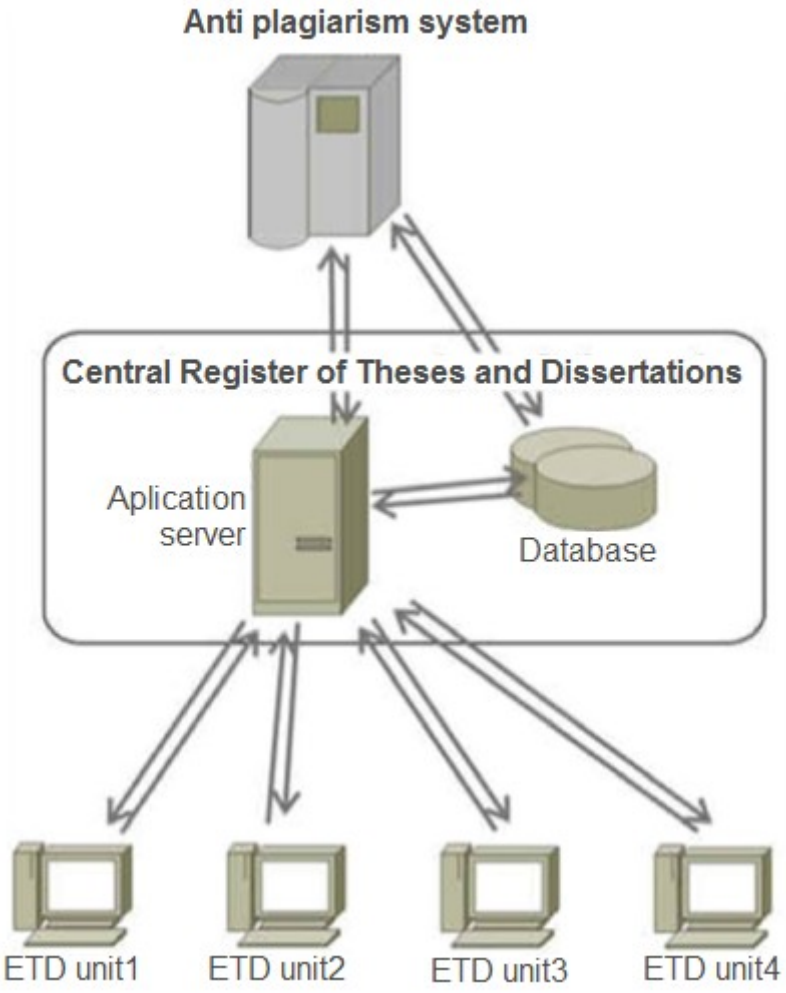

Fig. 1 Position of the CRTD in the process of final work collection[1]

On the basis of college or university request,the accepted batch of final works is subject to plagiarism detection in comparison with the comparative corpus made from the works sent and stored in the previous seasons, and/or in comparison with other resources (e.g. from the Internet). The results obtained in comparing to detect plagiarism are then joined with the respected final works and sent automatically by the system to the respective college or university. The sole texts of the works are incorporated in the comparative corpus of the anti-plagiarism system.

1 SKALKA, Ján; VOZÁR, Libor; DRLÍK, Martin; GRMAN, Ján: Centrálny register záverečných prác a metodika ich zberu. In IT*lib. 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. [On-line] [Dátum: 16. 4 2011.]

After delivery of the comparison results to the college or university,these serve only as basis for evaluation and classification of works by the appropriateExamining Commission. The final judgement whether it is an attempt for plagiarism or not (e.g. the work contains a higher number of correct quotations) and final decision lays always on the Examining Commission.

4 A Brief Description of Technical Solution

From the technical point of view, both the ETD and CRTD systems are designed as systems of cooperating servers.

- Application serverplays the role of communication server visible in the Internet environment, it provides portal services and ensures downloading of full texts
- Storage server represents a versatile repository for both original files and plain texts, output protocols, logs, etc.
- Database serverserves for operation of MS SQL databases
- Antiplag server contains index and searching core of algorithm to detect plagiarism (the numberof servers of this type will be increasing with increasing number of works)

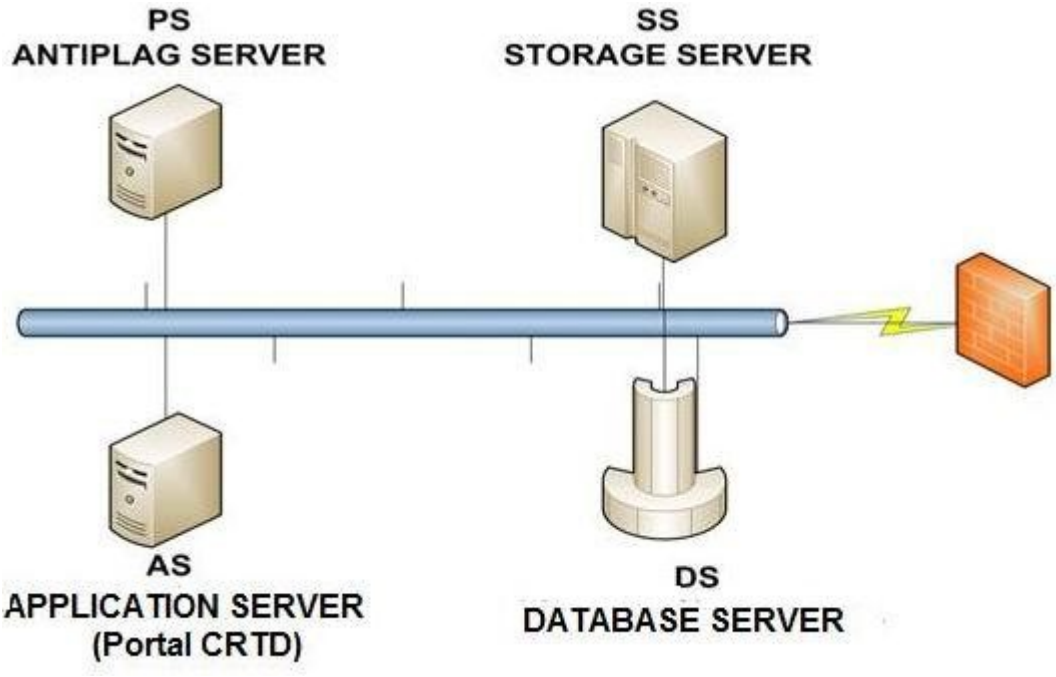

Fig. 2 A scheme of interconnection of individual parts of the system[2]

2 GRMAN, Ján: *Systémový popis APS (antiplagiátorský systém).* Technická dokumentácia k APS, 2010

The anti-plagiarism system itself is a set of several applications and original algorithms developed at the SVOP Ltd. company. They are designed as agent system and operated dynamically on the basis of impetuses.

Communication agents

- Input – downloads files from the localrepository (metadata and full texts) and converts them into plain texts.
- Output – transforms the results obtained in comparison for plagiarism into the PDF format, generates the results in the form of metadata, uploads files for agents of the central portal.

Index agents

- decompose (fragment) texts, detect documentlanguage, identify texts which are formulated artificially to defraud the system

Detection agents

- detectmatches in texts by means of special algorithms.

An example of a protocol of originality verification[3]:

Protokol o kontrole originality

Kontrolovaná práca

Citácia	Percento*	
Bibliographical Identification of tested document plagID: 765	13,61%	<< GP

*Číslo vyjadruje percentuálny podiel textu, ktorý má prekryv s indexom prác korpusu CRZP.

Práce s nadprahovou hodnotou podobnosti

Dok.	Citácia	Percento*	
1	Bibl. Identification of found document num.1	2,64%	<< P1
2	Bibl. Identification of found document num.2	13,09%	<< P2

*Číslo vyjadruje percentuálny prekryv testovaného dokumentu len s dokumentom uvedenom v príslušnom riadku.

3 RAVAS, Rudolf; GRMAN, Ján : *Technical aspects of plagiarizm detection system, presentation.* CVTI SR, 2011

5 Figures and Statistics

Already 33 colleges and universities send at present final and qualification works to compare them for plagiarism in the CRTD/APS system.

Approximately 71 000 final and qualification works were collected in the comparative corpus during its operation in the first Academic year, covering approximately 160 GB of storage space. And such amount of works is expected to be accepted every year. For comparison for plagiarism reasons about 3.4millions of documents covering approx. 1.4 TBwere downloaded from the Internet.The expected increase in the forthcoming years is about 70 000 new works yearly and hard to be determined increase of the other sources.

Comparison of the individual documents with the present corpus including metadata processing and PDF protocol generation takes 10 seconds in average. The system has managed successfully the maximum daily batch of documents which attained 4900 files, taking approx. 10 hours.

6 Feedback, Benefits and Evaluation after the First Year of Operation

The burning issue of plagiarism is very attractive for mass media communications. The SC STI as the system provider was expecting increased interest of mass media and feedback from the academic public. The initial stimulated interest on the side of mass media has fallen off quite quickly as no sensation appeared in this connection. Similarly, feedback from colleges and universities was mostly positive one despite the fact that our concept is based on strictly required comparison of the work for matches prior to its defence – which somebody can consider as unacceptable. Another reason of positive feedback may rise from the successful and reliable operation,the systemgenerates the required outputs with short period of response provided that the rules are observed.

The technical solution proved to be very stable and after initial instruction of users its application was generally trouble free. The system meets the requirements laid down on its selection. So far the comparison algorithms seem to be sufficiently quick and the system generates the results with some time reserve considered as very positive feature. It is a matter of fact that time between the work submission and its defence is evidently very short (approx. 2 weeks) and it is necessary to carry out detection for plagiarism, evaluation of results by the Examination Commission and writing of reviews by the reviewers. After the first year of operation it is difficult to analyze whether the system helped to achieve the determined goal – to reduce plagiarism in the final works. It is, however, evident that several tenths of works were sent back to be overwriten due to percentage of matches found. In my opinion, based on responses, the CRTD/APS has the preventive effect. Further application of the system in the future will bring possibilities of more disineterested evaluation of the system benefits.

Like in Slovakia also in Poland the similar issue of plagiarism has started to be solved. Therefore twoprominent Polish representatives took the invitationof the SC STI for a visit – namely Prof. Jan Kazmierczak ,Sejm deputy, Chairman of the Parliament Committee for Innovation and Informatization and Prof. Zbigniew Marciniak, Deputy minister, Ministry of Science and Higher Education. They highly appreciate and declare inspiring nature of the presented solution considered to be attractive for their pre-planned approach.

7 Perspectives for further Development

Perspectives for further development of the system are promising in several lines, namely:

a) Transfer of the system to the new created Data Centre for Research and Development provided by the SC STI, offerring IKT infrastructure of high quality, enables to speed up comparison processes and achieve higher quality of data protection.

b) Higher quality of comparative corpus can be achieved through agreement with graduate schools to send us also earlier works deposited in ther local AIS or libraries. All attempts in this field failedtill now.It is intended to determine other Internet resources which could be integrated into comparative corpus.

c) To ensure free access to final works deposited in the CRTD to public. It is incorporated in the Amendment to University Law.At present implementation of effects following from the Amendment on the CRTD/APS system is prepared.

d) Higher quality of comparison for originality can be achived by e.g. incorporating directory of synonyms or word roots into algorithms.

e) To improve functionality, user friendliness and possibilities of setting of the system for administrators and operators (e.g. to set different values for tolerated percentage of match for different branches of study).

The CRTD/APS will be extended in the future with providing the service also to other institutions than colleges and universities. Application to e.g. validation of documents which are outputs of R and D projects or documents which create a part of projects funded from public funds.

Link to CRTD/APS web site: www.crzp.sk

References:

1. RAVAS, Rudolf; GRMAN, Ján : *Technical aspects of plagiarizm detection system, presentation*. CVTI SR, 2011
2. MEŠKO, Dušan. *Plagiátorstvo*. [Online] [Dátum: 17. 2 2009.] http://www.etd.sk/doc/plagiatorstvo.doc.
3. 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.

http://www.crzp.sk/dokumenty/prevencia_odhalovanie_plagiatorstva.pdf [Online] [Dátum: 8. 1. 2011.]

4. SKALKA, Ján; VOZÁR, Libor; DRLÍK, Martin; GRMAN, Ján: Centrálny register záverečných prác a metodika ich zberu. In IT*lib*. *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. [On-line] [Dátum: 16. 4 2011.]
5. GRMAN, Ján: *Systémový popis APS (antiplagiátorský systém)*. Technická dokumentácia k APS, 2010