Od knižnično-informačného systému, cez evidenciu publikačnej činnosti po systém pre odhaľovanie plagiátov v akademickom prostredí.

Mgr. Ján Grman, PhD.

Anotácia
Na pôde akademických pracovísk sa objavuje celá plejáda informačných systémov. Tieto systémy podporujú činnosti univerzít. Spoločnosť SVOP spol. s r.o. je dlhoročným partnerom univerzít a predstavuje riešenia v oblasti informačnej podpory štúdia a výskumu. Komplexný knižnično-informačný systém DAWINCI je efektívnym nástrojom pre prezentáciu fondov univerzitnej knižnice pre študentov a zamestnacov. Nástroje evidencie publikačnej činnosti spolu so systémom Centrálneho registra publikačnej a umeleckej činosti sú dôležitou súčasťou systémov pre podporu a prezentáciu vedy a výskumu. Internet a dostupnosť zdrojov vytvárajú priestor pre nekontrolované šírenie ale aj zneužívanie myšlienok. Túto problematiku riešia systémy pre tvorbu lokálneho úložiska záverečných prác (EZP), Centrálny register záverečných prác (CRZP) a systém pre odhaľovanie plagiátov (ANTIPLAG). Príspevok je prierezom uvedených a realizovaných IKT riešení.

Annotation
Nowadays, a variety of information systems are in use in the academic workplaces. These provide a support for all of universities activities. Company SVOP Ltd. is an universities longstanding partner, which offers solutions in the field of information support of education and research. To presentation an university library collection to its students and employees DAWINCI presents a powerfull and complex library information system. Aplications of the evidence of academic publication activities together with the system of The Central registrel of academic publication activities and the Evidence of art works and performance are important parts of those systems, which support and present science and research. Internet environment and huge resources availability compile a space for not only uncontrolled spread but also for exploitation of ideas and thoughts. These problems solve systems which create local storages of thesis and dissertations (Evidence of Thesis and Dissertations - ETD), The Central Register of Thesis and Dissertations (CRTD/ETD) and system for plagiarism detection (ANTIPLAG). The contribution is an intersection of above mentioned and realized solutions of library information technologies.

Úvod
V súvislosti s automatizáciou akademických pracovísk sa objavuje na pôde univerzít celá škála informačných systémov. Všetky špecifické úlohy nie je možné realizovať jediným informačným systémom (IS), je však možné postaviť súbor úzko spolupracujúcich IS. Prostriedkom, ktorým je možné dosiahnuť maximálnu kooperáciu, je použitie otvorených a flexibilných IS od spoľahlivých dodávateľov. Dôležitým faktorom je najmä schopnosť systémov pružne reagovať na zmeny legislatívy a očakávaní rôznych skupín užívateľov. Zmeny financovania, zmena pomeru denných a externých študentov, zvyšovanie dôrazu na výsledky vo vedeckej práci kladie čoraz vyššie nároky aj na IT infraštruktúru VŠ.

Univerzitná knižnica
Knižnično-informačný systém (ďalej len KIS) sa nepovažuje práve za kritický IS pre fungovanie VŠ ako celku. Jeho prvotnou úlohou bolo a zostáva evidencia fondov (dokumentov) knižnice, riešenie výpožičného procesu, kvantitatívnych, výkonnostných a čiastočne ekonomických ukazovateľov knižnice (štatistiky). Principiálne ide o evidenciu rozsiahlych metadátových štruktúr s vysokou variabilitou. Množstvo druhov evidovaných

dokumentov je v prostredí akademickej knižnice naozaj vysoké. Funkcie KIS sú teda pomerne špecifické. Základné funkcie sú však ideovo jasné: evidencia a dostupnosť informácií pre študentov a zamestnancov.

Špecificky v akademickom prostredí sa akcentuje význam evidencie publikačnej činnosti. Sledovanie jej parametrov je nevyhnutné tak pre potreby akreditačného procesu, sledovania odbornej spôsobilosti a kvality pracovníkov a doktorandov, ale súvisí aj s prideľovaním financií, dokonca často na VŠ implikovalo zmeny vnútorných pravidiel oceňovania aktivity pracovníkov. Táto agenda vyžaduje dôslednú a presnú evidenciu a univerzitná knižnica v tomto procese hrá významnú a aktívnu úlohu. Vývoj v oblasti EPC je príkladom toho, ako zmeny zákonov a smerníc a automatizácia procesov na centrálnej úrovni spôsobí tlak na automatizáciu procesov na úrovni lokálnej. Vývoj a spustenie centrálneho registra publikačnej činnosti (www.crepc.sk, rovnako produkt spoločnosti SVOP spol. s r.o.) malo vplyv na unifikáciu existujúcich riešení evidencie a vznik nových a presnejších a navyše automatizáciu tohoto procesu aj tam, kde nebola dovtedy realizovaná. V prostredí KIS ponúkame riešenie pre evidenciu publikačnej činnosti už od roku 2003. Od tej doby vzniklo viacero modulov evidencie, kontroly a exportu údajov (od evidencie prostriedkami KIS v rámci univerzitnej knižnice, cez prezentáciu na OPAC a Z39.50, až po formulárové systémy tvorby návrhov záznamov priamo autormi a podobne).

Evidencia publikačnej činnosti

Motiváciou pre tvorbu projektu centrálneho registra EPČ (ďalej CREPČ) bolo sústrediť, unifikovať, archivovať, analyzovať, kontrolovať a porovnávaťdáta na priereze rokov globálne pre všetky VŠ. Zároveň dôsledkom existencie registra sú funkcie umožňujúce tieto dáta prezentovať pre kontrolu VŠ, agentúram, akreditačným komisiám a odbornej aj laickej verejnosti. Riešenie registra CREPČ reflektuje požiadavky na Evidenciu publikačnej činnosti definovaných smernicou č. 8/2007-R z 31. mája 2007 a jej doplnkov.

Výmenným formátom registra je XML dátová dávka formátovaná na základe definovanej schémy. Kódovanie znakov je v UTF-8. Výsledný formát je ľahko implementovateľný aj v informačných systémoch nepoužívajúcich klasické knižničné normy (ISO2709, niektorý z MARC formátov) a je tiež v súlade s Výnosom Ministerstva financií Slovenskej republiky z 9. júna 2010 o štandardoch pre informačné systémy verejnej správy.

Výstupov registra je súborná databáza EPČ dostupná pre verejnosť, ale aj oficiálne štatistiky a normalizované výstupy pre účely kontroly a prezentácie (pre VŠ a MŠ SR). Projekt CREPČ je dostupný na adrese www.crepc.sk a aktuálne končí štvrtý ročník zberu dát (roky 2007 až 2010).

Kooperácia IS

Koexistencia a kooperácia viacerých IS v uzatvorenom prostredí je nevyhnutná. Takmer každý akademický informačný systém pracuje s informáciami o zamestnancoch alebo študentoch. Zdieľanie a propagácia dát je v tomto kontexte základnou pracovnou metódou.
Už v prípade KIS je možné pozorovať prvé externé informačné toky. Tvorba konta čitateľa knižnice, resp. tvorba autoritného záznamu pracovníka pre potreby evidencie publikačnej činnosti je predurčená na automatizovanú propagáciu už raz centrálne evidovaných údajov o zamestnancovi. V prípade DAWINCI ide o používanie informácií na čipových kartách pracovníkov a študentov, resp. o čítanie LDAP záznamov autorizačného centra, prípadne komunikáciu na AIS. Uvedené funkcie sú prejavom flexibility IS a jeho schopnosti prepojiť

sa a používať údaje iných systémov. Dôležitá je však aj otvorenosť IS smerom na iné systémy za podpory noriem a priemyselných štandardov.

Ak hovoríme o produktoch spoločnosti SVOP spol. s r.o., môžeme hovoriť o prepojení na systémy prostredníctvom HTTP/SOAP/Z39.50/RSS/OAI a získavať štruktúrované informácie vo formátoch XML, ISO, HTML, XHTML, RTF, CSV, XLS a ďalších. Vďaka otvorenej SQL architektúre je možné aj priame dátové prepojenie na iné IS.

V oblasti IS je dôležitou zložkou aj podpora zo strany dodávateľa a jeho skúsenosti a "know-how" v odborných otázkach. Samozrejmosťou je dodávka nových verzií, poradenstvo v oblasti periférií, riešenie problémov úplne, či čiastočne súvisiacich s dodanými systémami a ich procesmi. Všetky naše riešenia sú typu klient-server čoho dôsledkom je ich ľahká šíriteľnosť (na rôznych počítačoch v práci ale aj z domu). Tejto filozofii zodpovedá aj model licencovania systémov založený na sledovaní počtu aktuálnych spojení (concurrent licence model). V oblasti anonymného prístupu (www/http rozhrania) a serverových služieb sa uplatňuje licencovanie na server. Neobmedzuje sa tým počet užívateľov pre funkcie vyhľadávania, možnosti používať RSS kanály, elektronické žiadanky na tituly, rezervácie a informácie o stave účtu a podobne.
Ak hovoríme o informačných systémoch, hovoríme najmä o ukladaní dát. V tejto otázke sa riešenia spoločnosti opierajú o platformu spoľahlivých a stále neprekonaných relačných databáz typu SQL.

Zber záverečných prác
Zber záverečných prác v papierovej forme vyplýva zo zákonov o biblografickej registrácii a archivácii dokumentov. V oblasti elektronického zberu je motiváciou nie len výhoda efektívneho archivovania a prezentácie, ale aj otázky plagiátorstva a jeho odhaľovania.

Plagiátorstvo je činnosť, ktorá sa akosi rýchlo zakorenila v myslení študentov a to najmä vďaka internetu a jednoduchosti kopírovania textov z webových stránok. Tento problém je v živote vysokých škôl naozaj živý, pretože záverečné práce prinášajú svojim autorom profesionálny a často i finančný a spoločenský prospech. Odhalené plagiátorstvo navyše škodí menu inštitúcie, ktorá umožní takúto prácu obhájiť.

Záverečná práca je vyvrcholením istého stupňa a formy štúdia a jej obhajoba je nevyhnutným predpokladom ukončenia štúdia a získania príslušného titulu. Väčšina záverečných prác už vzniká v elektronickej forme použitím textového editora. To je veľmi dobrý predpoklad pre zber elektronických verzií ZP.

So zberom ZP mala spoločnosť SVOP skúsenosti už na platforme KIS. Tieto práce v papierovej forme boli totiž typicky samostatným evidenčným fondom. Akademická knižnica je totiž podľa Zákona o knižniciach povinná záverečné práce zbierať a uchovávať, no nemá právo na ich ďalšie rozširovanie ani poskytovanie ďalším subjektom. Tvorba špecializovaného modulu pre tvorbu lokálnych úložísk elektronických verzií ZP bola teda len logickým dôsledkom a využitím získaného „know-how“.

Ako nevyhnutné vyvrcholenie úsilia a potlačenie rozmáhajúceho sa plagiátorstva sa na konci roka 2008 zrodila iniciatíva Ministerstva školstva SR s cieľom realizovať komplexné riešenie zberu a testovania ZP na národnej úrovni. Bola vypracovaná štúdia a metodika pre národný zber záverečných prác za účelom odhaľovania plagiátorstva. Zber elektronických záverečných prác v rámci VŠ sa riadi podľa Metodického usmernenia MŠ SR č. 14/2009-R. Uzatvorenie

licenčnej zmluvy medzi autorom a vysokou školou umožnuje jej ďalšie spracovanie (napr. antiplagiátorským systémom).

Spoločnosť SVOP na základe výberového konania realizovala systém centrálneho registra zberu záverečných a kvalifikačných prác (CRZP - v prvej polovici roku 2009). Následne vyrobila tiež referenčnú aplikáciu pre lokálne úložisko ZP určenú primárne na testovanie CRZP a neskôr prípadným záujemcom, teda vysokým školám ktoré v tej dobe takéto úložisko ešte nemali vybudované.

Navrhnutý distribuovaný systém zberu ZP je možné po takmer roku prevádzky označiť za úspešný a životaschopný. Univerzity používajú niekoľko odlišných systémov lokálnych úložísk a takmer všetky komunikujú s CRZP úplne automatizovane bez zásahu obsluhy. Systém lokálneho úložiska EZP od našej spoločnosti využili najmä zákazníci KIS DAWINCI, ale kedže ide o nezávislý a otvorený systém, prevádzkujú ho aj VŠ s odlišnými AIS a KIS systémami. Môžeme smelo tvrdiť, že prichádzame do kontaktu a máme komunikačne zvládnuté takmer všetky kombinácie nasadenia CRZP, EZP, KIS, AIS a LDAP systémov na slovenských univerzitách.

Lokálne úložisko pre zber ZP (SVOP EZP) okrem štandardných funkcií vkladania metadát a súborov plného textu realizuje aj mnohé, nie celkom bežné, ale užitočné funkcie. Ide najmä o textovú validáciu dokumentu z hľadiska vhodnosti použitého formátu (PDF), ktorá odhalí možné problémy s extrakciou „plain textu“ z dokumentu, šifrované alebo neštandardne vytvorené dokumenty nevhodné na ďalšie spracovanie a napokon dokáže overiť aj minimálny rozsah odovzdávanej práce. Túto validáciu je možné realizovať aj v CRZP, ale zakomponovanie do EZP je veľmi výhodné pre študentov. Dôležité je tiež generovanie príslušných dokumentov (licenčná zmluva, potvrdenie o odovzdaní a podobne) a kontrolné a komunikačné mechanizmy úložiska.

Aktuálny stav CRZP je približne 67 tisíc záverečných prác všetkých druhov. V týchto dňoch prebieha indexovanie už získaných a predspracovaných internetových dokumentov v objeme asi 600 až 700 tisíc dokumentov (v tejto fáze s preferenciou na PDF a DOC dokumenty a v slovenskom jazyku). Register prác sa zároveň sprísňuje. Pribúdajú nové kontroly textu s cieľom odhaliť neštandardné kódovania najmä PDF dokumentov. Centrálny register ZP okrem komunikačného rozhrania pre lokálne úložiská a administrátorov VŠ ponúka tiež rozhrania pre:

- Testovanie prevodu práce na čistý text
- Návrhy liniek, dokumentov a portálov pre indexovanie
- Lokálne úložisko pre CVTI pre účely testovania a indexovania dokumentov na požiadanie
- Subsystém na získavanie protokolov na základe známeho identifikátora
- CMS systém pre úpravy informačného portálu www.crzp.sk

Pripravuje sa vyhľadávací nástroj na prehľadávanie metadát záverečných prác s možnosťou prezentácie aj plných textov vo vhodnej forme v súlade s legislatívou (zákon o plošnom zverejňovaní je na pôde parlamentu).

Antiplagiátorský systém

Algoritmus pre odhaľovanie plagiátov je originálnym dielom spoločnosti SVOP. Parametricky aj ideovo ide o unikátny systém agentového typu. V tomto období sa okrem indexovania korpusu začalo s indexovaním „viditeľného webu“. Ide o prepracovanú metódu cieleného dolovania zdrojov bez nutnosti hrubého sťahovania celých webových sídiel a teda

množstva nerelevantného materiálu. Zároveň sa pre druhý ročník pripravuje tiež alternatívny mechanizmus vyhľadávania podobností. Spojením oboch metód sa dosiahne ďalšie zlepšenie detekčných vlastností a možností.
Systém ANTIPLAG je systém pre podporu rozhodovania. Jeho výstupom je protokol, elektronický dokument, generovaný algoritmom na odhaľovanie plagiátov. Tento upozorňuje na dokumenty, ktoré mohli uniknúť pozornosti školiteľa alebo oponenta. Každá práca je porovnávaná voči indexu systému ANTIPLAG. Zo všetkých dokumentov sú vybrané tie, v ktorých sa nachádza nadprahové množstvo podobného textu.
Aktuálne sa pracuje tiež na čistení slovníka slov s dôrazom na detekciu slovenského jazyka, samotný algoritmus je však principiálne použiteľny pre ľubovolný jazyk. V najbližšom období by sme sa chceli špecializovať na jazyk slovenský, anglický a následne na slovanské jazyky. V poslednom odbobí boli nasadené detekčné a kontrolné mechanizmy umožňujúce rozpoznať niektoré manipulácie s dokumentom pri ktorých sa síce nemení vzhľad dokumentu, no mení sa obsah čistého textu a sťažuje sa identifikácia slov. I keď teória hovorí že boj proti záškodníkom je bojom márnym, je možné túto činnosť minimálne značne sťažiť.

Literatúra

[1] Skalka, Ján a kol. 2009. 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.
[2] Grman Ján. 2006. Výhody systému DAWINCI ako prostriedku pre evidenciu a prezentáciu publikačnej činnosti. In: Zborník konferencie UNINFOS 2006 v Nitre.
[3] Grman Ján. 2005. AKIS DAWINCI - aktívna súčasť akademického informačného systému. In: Zborník konferencie UNINFOS 2005 v Banskej Bystrici.
[4] Centrálne registre: www.crepc.sk, www.crzp.sk,

Kontakt: Mgr. Ján Grman, PhD., SVOP, spol. s r.o., Líščie údolie 59, 841 04 Bratislava, Slovenská republika, tel./fax: +421 265 422 752, tel.: +421 905 412 681, e-mail: grman@svop.sk, www.svop.sk, www.dawinci.sk