Národní registr VŠKP a systém na odhalování plagiátů

Michal BRANDEJS, Miroslav KŘIPAČ, Jitka BRANDEJSOVÁ, Jan KASPRZAK
Masarykova univerzita, Brno
brandejs@fi.muni.cz, kripac@fi.muni.cz, brandejsova@fi.muni.cz, kas@fi.muni.cz

INFORUM 2008: 14. konference o profesionálních informačních zdrojích
Praha, 28. - 30.5. 2008

Abstrakt.
Jedním z mimořádně tíživých problémů vysokého školství, jemuž značný význam přikládá i široká veřejnost, je plagiátorství vysokoškolských prací. V loňském roce si sedmnáct vysokých škol vytyčilo cíl přispět ke zvýšení kvality těchto prací a vysokoškolského vzdělávání vytvořením celonárodního registru vysokoškolských kvalifikačních prací (VŠKP) spojeného s úložištěm kvalifikačních prací včetně odhalování plagiátů (podobných textů). Projekt navazuje na technologické zkušenosti Masarykovy univerzity, která disponuje vlastním úložištěm elektronických verzí závěrečných prací od roku 2004 a od roku 2006 veřejně přístupným archivem závěrečných prací v rámci svého informačního systému pro podporu studia (Informační systém Masarykovy univerzity, IS MU) včetně jedinečné služby pro odhalování plagiátů. Dále také navazuje na zkušenosti koncepčního a metodologického charakteru, které jsou nezbytné pro sběr nebo zpřístupňování prací a souvisejí s univerzitními procesy. Masarykova univerzita je koordinátorem projektu. Od ledna 2008 je na adrese http://theses.cz/ již k dispozici testovací verze systému. Ukázky a informace k aktuálnímu stavu projektu jsou hlavní částí příspěvku. Projekt využívá v oblasti struktur a funkcí národního registru VŠKP zkušeností Odborné komise pro otázky elektronického zpřístupňování vysokoškolských kvalifikačních prací v čele s Vysokou školou ekonomickou v Praze. Projekt se neomezuje jen na práce závěrečné, ale navazuje i na práce seminární nebo publikační díla. K projektu přistoupily také soukromé a zahraniční školy.

Historie.
Myšlenka projektu Národního registru VŠKP a systému na odhalování plagiátů vznikla jako přímý důsledek tržní poptávky, zejména ze strany vysokých škol, po službě odhalování plagiátů mezi vysokoškolskými pracemi. Již v roce 2006 dala Masarykova univerzita svým studentům, vyučujícím a zaměstnancům k dispozici velmi kvalitní nástroj pro odhalování podobných dokumentů v Informačním systému Masarykovy univerzity (IS MU). Učinila tak, jak je jejím zvykem, na základě poptávky vlastní akademické obce. MU je univerzitou otevřenou a maximálně podporuje zpřístupňování elektronických studijních materiálů ve výuce stejně jako elektronického archivu závěrečných prací aj. Na druhé straně si je vědoma, že každá otevřenost nese riziko potenciálního zneužití. Nový technologicky úspěšný nástroj poskytuje účinné prostředky pro technické odhalení podobných dokumentů, což je první důležitý krok v procesu správného posouzení, zda se jedná o plagiát.
Informace o vlastní ojedinělé implementaci nástroje se rychle rozšířila a zájem z řad veřejnosti o tuto službu byl značný. Posledním podnětem, který převážil misku vah ke kladnému rozhodnutí vytvořit celonárodní projekt, byla právě loňská konference Inforum 2007. Na konferenci byl nominován a oceněn Archiv závěrečných prací Masarykovy univerzity (http://www.inforum.cz/archiv/inforum2007/infoceny/; http://is.muni.cz/lide/absolventi.pl) a o projektu, který by službu odhalování plagiátů nabízel co nejširšímu okruhu zájemců převážně z řad vysokých škol, bylo rozhodnuto. Po rozhodnutí Masarykovy univerzity pustit se v roce 2007 do tohoto projektu nebyla cesta k realizaci první verze systému úplně přímočará. V první fázi bylo nutné o celém projektu a možnosti se do něj zapojit informovat všechny školy. Bylo třeba vysvětlovat, objasňovat a informovat nejen vysoké školy, ale i veřejnost, která se o téma plagiátorství zajímá. Mnohé školy si nedovedly představit, co je možné od projektu očekávat, jaké typy služeb mohou být jeho součástí, s jakými výsledky mohou být konfrontovány nebo jaká by měla být jejich spoluúčast, pokud se projektu zúčastní. Protože jde o projekt, stejně jako o společenský problém, zcela nový a ojedinělý, mnohé otázky jsou průběžně pokládány a další teprve položeny budou. Otázka plagiátorství představuje určitou daň rozvoji Internetu a dluh těmto otázkám.
V září roku 2007 následně došlo k urychlenému spojení s projektem Národního registru připravovaného zástupci Odborné komise pro otázky elektronického zpřístupňování vysokoškolských kvalifikačních prací v čele s Vysokou školou ekonomickou v Praze. Celkový počet zapojených vysokých škol v tomto projektu

tak ve finále dosáhl počtu sedmnácti, jejich seznam je uveden na stránkách http://theses.cz/. 1. října 2007 podaly tyto školy na Ministerstvo školství, mládeže a tělovýchovy centralizovaný rozvojový projektu MŠMT „Národní registr VŠKP a systém na odhalování plagiátů“ a projekt podporu ze strany MŠMT získal.

Cíle a vlastnosti projektu.

Hlavním cílem projektu je vytvořit celonárodního úložiště[1] závěrečných prací v elektronické formě a umožnit vyhledávat v něm plagiáty (podobné soubory). Součástí tohoto cíle je vytvoření celonárodního registru[1] vysokoškolských kvalifikačních prací (VŠKP).

Vedlejším cílem pak je vznik archivu[2] závěrečných prací, ve kterém si bude škola moci volit úroveň zveřejňování prací od úplného zákazu zveřejňování po plné zveřejnění.

Proto systém předpokládá diferencovaný přístup jednotlivých škol ve škále od poskytování pouze údajů do registru až po poskytování plných textů prací za účelem vyhledávání podobných textů.

Od těchto cílů se odvíjí vlastnosti systému. Práce mohou být v elektronické podobě zveřejňovány a systém nabízí jednotný přístup ke všem pracím z jednotlivých škol. Systém bude nabízet vyhledávání informací v pracích, tedy možnost vyhledávání podle obsahu a zpřístupnění zajímavých výsledků v různých oblastech studia, ale i vyhledávání informací o pracích, například, kde lze získat práci, není-li k dispozici v elektronické formě. Nejvýznamnější funkcí však zůstává vyhledávání podobností v pracích za účelem odhalení plagiátů.

Základní doménou pro systém byla zvolena theses.cz a server nese název Vysokoškolské kvalifikační práce.

Aktuální stav projektu.

Podávaný centralizovaný projekt se dělí na hlavní technické řešení, za něž odpovídá Masarykova univerzita, a dílčí organizačně-technická řešení, které realizují spolupracující vysoké školy.

Masarykova univerzita se velice pečlivě připravovala na vstup do projektu, v němž se zavázala ostatním školám a bylo zřejmé, že nejpozději v květnu bude muset být celý systém v základní variantě k dispozici, aby školy mohly začít vkládat červnové závěrečné práce.

1. ledna 2008 byl spuštěn systém, resp. jeho část „Národní registr VŠKP“ a školám byly dány k dispozici základní vlastnosti pro zřízení přístupu. Současně jim byla předána dokumentace týkající se koncepce aktuální části projektu, popisu prvků, formátu importu souboru, příklad souboru, seznam škol, seznam jazyků. Systém v této úvodní fázi také obsahoval fulltextové hledání, tematické hledání, správu dat, možnost importu i mazání dat, správu osob a další užitečné aplikace (pro řízení projektu, záložky apod.).

13. ledna 2008 následovala aplikace pro nastavení konfigurace systému, kde škola svými výroky nastavuje, jak se má systém chovat ke vkládaným datům, například zda se metadata předávaná Národnímu registru zveřejňují a komu, zda se zveřejňují plné texty prací a komu, zda se vyhledávání v plných textech prací povoluje/nepovoluje světu.

V úvodní části si každý řešitel zvolil správce systému za danou školu, jehož úkoly jsou: spolupráce se zástupci projektu MU, zavádění osob do systému pro vlastní školu, přidělování a nastavování práv, příprava importu dat do systému a další úkoly. Školy byly následně informovány o aktuálním stavu a vyzvány k připomínkování i testování stávajících aplikací.

K 28. lednu 2008 byla rozšířena a upravena koncepce pro uživatele systému, zavedena stránka Častých otázek a odpovědí „FAQ“, zprovozněna aplikace „Vývěska a aktuality“ pro informování uživatelů a „Diskusní fóra“ pro možnost diskusí. Nově byly přidány stránky „Lidé“ o uživatelích systému a „Úschovna“ pro ukládání velkých souborů nebo „Můj web“ sloužící jako osobní úložiště dokumentů. Další změny se týkaly dílčích úprav stávajících aplikací.

V únoru probíhala zejména komunikace mezi zástupci tvořícími metadatový standard, resp. formát importu do systému. Připomínkové řízení bylo ukončeno k 1. březnu 2008. V tomto období byl v systému implementován „Dokumentový server“ a doplněna dokumentace o číselníky a formáty importu do Národního registru VŠKP.

7. března 2008 byly školám předloženy smlouvy k připomínkování, které byly dále postupně uzavírány.

Projekt tedy probíhá podle harmonogramu. Na straně některých škol dochází ke zpoždění, což však zatím není problematické pro ty z nich, které čekají především na nejvážnější fázi projektu – „Systém pro odhalování plagiátů“, jež by měl být k dispozici v květnu 2008. Teprve pak nastane hlavní testovací období systému a sběr metadat/závěrečných prací.

Nezanedbatelnou organizační část projektu tvoří potřebná a užitečná komunikace se zástupci již zúčastněných škol, ale také se zástupci, kteří teprve projevili zájem se do projektu zapojit. K sedmnácti

zúčastněným školám v projektu se přidávají další z řad státních, soukromých a zahraničních vysokých škol, očekává se postupný vstup i dalších subjektů do projektu.
Spoluřešitelé projektu jsou z řad prorektorů, zástupců oddělení strategie a rozvoje škol, knihovníků nebo ředitelů knihoven, vedoucích ÚVT nebo informatiků – jde o poměrně různorodou škálu delegovaných osob, kde každý ze zástupců může mít odlišnou motivaci, proč se zapojil (nebo i „byl zapojen“) do projektu.
V nejbližší době bude nutné se zaměřit nejen na testovací provoz pro sběr dat a postupné spuštění do ostrého provozu, tedy budování softwarové části, ale také na zajištění hardwarové části projektu.

Projektové řízení.
Způsob řízení projektu využívá klasického řízení projektu kombinovaného s metodou extrémního programování. Díky tomuto způsobu řízení se podařilo vývojovému týmu IS MU na Masarykově univerzitě posunout systém pro podporu studia k neobvykle rozsáhlému využívání studenty a vyučujícími a úspěch MU v této oblasti dokazuje, že tento způsob se osvědčil.
Předinvestiční a investiční etapa byla daná, projekt se v roce 2008 omezuje na provozní a vyhodnocovací řízení projektu. V roce 2007 jsme však věnovali hodně času přípravě projektu: vycházeli jsme ze SWOT analýzy, dílčí analýzy trhu a stanovili jsme předpokládané náklady, plány a postupy pro rok 2008 aj.
Klasické řízení si projekt vyžaduje například v oblasti projektového týmu. Již z povahy rozvojového projektu jsou členy projektu zástupci vývoje MU a zástupci zúčastněných škol. Za MU jde o hlavního koordinátora projektu (doc. Ing. Michal Brandejs, CSc.), programátory systémové, aplikační, organizačního pracovníka, testera, uživatelskou podporu a designéra, ekonoma, právníka a další. Na druhé straně stojí spoluřešitelé za školy, informatici a knihovníci (zatím v rozsahu jedné až čtyř osob za školu). Ne však vždy jsou tito pracovníci v projektu na plný úvazek (záleží na etapách a harmonogramu projektu).
Harmonogram je daný rozvojovým projektem, ale je nutné ho přizpůsobovat reálné situaci, kterou někdy je a někdy není možné ovlivnit. Zde je snaha přistupovat ke školám individuálně a posílat jim vždy informace o uvolněných verzích aplikací, koncepci a jiné informace, které jsou vázané k aktuální etapě projektu. Současně s tím jsou požádány o konkrétní kroky (odpověď, připomínky, testování atd.). Z toho vyplývá, že jsou vynechávány etapy „analýz“, resp. dlouhodobého nebo krátkodobého sběru informací (o jejichž významu máme v tomto typu projektu pochybnosti vycházející ze zkušeností), ale je vždy dána k dispozici hotová část systému k připomínkám. Na první pohled se zdá, že změna implementace na základě připomínek může přinést vícenáklady, ale není tomu tak. Až při konkrétní podobě aplikace si uživatel totiž může uvědomit, jak si chování a fungování ve skutečnosti představoval. A pokud není notorickým „šťouralem“, dokáže jasně formulovat své zásadní potřeby na změnu nebo souhlas s navrženým řešením. Součástí projektu je také smluvní příprava, dohody a podepsání smluv. Změny jsou obvykle rychlé a konstruktivně vedené.

Problémy.
Jsou ☺.
Stručně:

- zvyk na jiný, méně efektivní způsob řízení projektu;
- nedodržení termínu/ů;
- nepřipravenost na projekt (míněno v době spuštění testovacího období);
- závislost na dodavateli, který zajišťuje evidenci VŠKP;
- neobjasněné a nevysvětlené důvody (sdělené i nesdělené obavy);
- a jiné a jiné a… ;-).

„Je lepší zapálit svíčku, než si stěžovat na tmu.“ Čínské přísloví

Plány.
Jedním z důležitých úkolů pro úspěšné zavedení systému do produkčního provozu je vyřešení řady technických otázek. První z nich se týká samotného technologického vybavení. Vzhledem k tomu, že výběr a dodání jednotlivých počítačů vyžaduje řadu měsíců, běží testovací fáze projektu na dočasném hardware MU, který postačuje pro nižší výkon. Na základě zkušeností z minulých let však byla navržena začátkem roku nová architektura serveru, která plně podporuje distribuované zpracování dat včetně unikátního distribuovaného úložiště plných textů prací. V dubnu a květnu tohoto roku probíhá výběrové řízení na dodavatele potřebných komponent.

Vedle nasazení hardwarového vybavení bude v nejbližších dnech zprovozněna také zcela nová generace algoritmů na vyhledávání podobností. Ta se vyznačuje jak schopností pojmout velké množství dat (v současné době systém zpracovává přibližně 750.000 dokumentů), tak vysokou propustností vyhledávání podobností (všechny dokumenty jsou prohledány v řádu jednotek hodin) nebo rychlostí dohledání podobností k nově přidanému dokumentu. Na základě zkušeností s praktickým provozem byly současně optimalizovány algoritmy pro vyšší účinnost dohledání podobností tak, aby nebylo snadné opsaný dokument modifikovat a zabránit detekci plagiátorství.
Rovněž existující technologie fulltextového vyhledávání v plných textech závěrečných prací není pro nasazení ve velkém měřítku vhodná. Proto v současné době pracujeme opět na zcela nové verzi vyhledávání, která bude lépe využívat distribuovaný charakter ukládání dat a přidá nové vlastnosti jako je vyhledávání bez ohledu na tvar slova apod. Důležité je však zachování stávající schopnosti systému vyhledávat přesně, tedy s ohledem na přístupová práva a v celé databázi, ne pouze ve vybraných (dosud indexovaných) dokumentech.
Důležitou vlastností registru je také možnost zviditelnění prací v dalších registrech a internetových vyhledávačích, ať už formou cílené optimalizace pro automatické roboty vyhledávačů, nebo pouhým zveřejněním dalších odkazů na práci. Tato schopnost může přinést velké výhody nejen samotným školám a studentům, ale i široké veřejnosti, která se k vystavovaným pracím snadněji dostane.
Předpokládá se také napojení na projekt kontroly plagiátů v seminárních vysokoškolských pracích, publikacích a napojení na vybrané zdroje Internetu.

Literatura.

Vysokoškolské kvalifikační práce. Oficiální stránky Theses.cz [online]. 2007. Dostupný z www: http://theses.cz/. [cit. 2008-04-30]