Theses.cz

Michal Brandejs, Jitka Brandejsová, Jan Kasprzak, Miroslav Křipač, Martin Stančík

Abstrakt
Projekt Národní registr VŠKP a systém na odhalování plagiátů se stal populárním nejen v akademické sféře jako projekt Theses (podle domény http://theses.cz/). Původně národní projekt přesáhl hranice a stal se projektem mezinárodním díky zapojení slovenských škol. Systému využívají také vysoké školy soukromé a státní a počet zapojených škol stále roste. V příspěvku se účastníci semináře dovědí, co systém představuje po technologické stránce (např. služba lokálního úložiště prací - implementována), jaké jsou zkušenosti z provozu systému s vloženými pracemi a co se plánuje. Seznámí se s praktickými postupy při vyhledávání plagiátů, se zkušenostmi, jak postupovat při podezření z plagiátorství, a s příp. statistikami.

Během první poloviny roku 2008 byl systém postupně spuštěn v pilotním a následně i ostrém provozu. Z původně národního projektu sedmnácti veřejných vysokých škol, které se do projektu zapojily na samém počátku, došlo k rozšíření jak o soukromé školy, tak o školy ze Slovenské republiky. V době konání semináře, v říjnu 2008, pak probíhají jednání s dalšími školami z obou zemí, které o připojení mají také zájem. Systém tak zůstává otevřen pro všechny zájemce, což dále zvyšuje jeho efektivitu.

Jedna z viditelných změn, kterou mohli uživatelé systému Theses.cz zaznamenat, byla změna v designu a grafických stylů aplikacích použitých v autentizované (určené převážně pro správce) i veřejné části systému.

Na žádost některých škol, které kromě hlavních funkcí systému jako je vyhledání plagiátů a centrální registr prací využijí systém jako své primární lokální úložiště, byla do provozu v průběhu léta spuštěna nová aplikace, která umožní vkládání prací samotnými studenty. Tato funkce ulehčí lokálním správcům na dané škole práci se správou lokálního systému sběru závěrečných prací a zároveň ponechává škole možnost zpřístupnit svým studentům a zaměstnancům práce v jiném režimu, než veřejnosti.

Velký důraz při vývoji systému v poslední době byl kladen také na nové možnosti vyhledávání nejen v metadatech, ale i plných textech závěrečných prací. Od počátku byl systém spuštěn s funkcí vyhledávání v plných textech (samozřejmě s ohledem na velmi jemnou a tedy přesnou specifikaci přístupových práv k jednotlivým dokumentům). Původně zvolená technologie, která vychází z produktu Oracle Text, se však ukazuje v rozsáhlém prostředí jako nepříliš vhodná. Oproti jiným způsobům vyhledávání nejen na Internetu sice obsahuje poměrně silný vyhledávací jazyk a zejména možnost vyhledávat v částech slov, současné implementace však vykazují problémy zejména v českém prostředí a neumožňují příliš flexibilně vyhledávat v dokumentech uložených mimo relační databázi v distribuovaném úložišti, které systém Theses.cz používá od počátku vývoje. Proto v současné době připravujeme novou verzi jádra vyhledávacího systému, které umožní tyto nevýhody eliminovat.

Poměrně dlouhým procesem vývoje prošly také formáty pro automatizované i ruční vkládání dat. V současné době tak systém plně podporuje jak vlastní formát Theses.cz 1.0 (vyvinutý přímo pro projekt Theses.cz), tak formát Evskp.cz 1.1 vyvinutý Odbornou komisí pro otázky elektronického zpřístupňování vysokoškolských kvalifikačních prací. Systém dále podporuje přenos údajů o pracích uložených v těchto formátech různými způsoby počínaje ručním vložením pomocí formuláře v autentizované části systému (zejména vhodné pro testování), přes automatické vkládání do systému pomocí HTTPS protokolu až po protokol OAI-PMH.

Další velkou inovací, kterou prošel systém během svého krátkého vývoje byla celková změna v hardwarovém vybavení systému. Předpokládáme, že technické vybavení je na takové úrovni, že bude možné zabezpečit provoz systému nejméně na tři roky dopředu bez toho, aby se musela tato poměrně nákladná část inovovat.

S rozvojem systému v ostrém provozu došlo také k prvním nalezeným podobnostem. Přestože cílem projektu není hledání exemplárních případů plagiátorství jako takové a jako správci

systému se o všech případech nedozvíme, pokud jednotlivé školy nepožádají o součinnost (systém je založen na principu přímé komunikace škol, které vložily navzájem podobné soubory), první případy vyhledaných podobností ukazují, že systém může být pro školy skutečnou pomocí.

Správci pověření za každou školu tak mohou využít nejen přímé nalezení podobností (označením jednoho nebo více souborů v systému), tak aplikaci tzv. globálního vyhledávání plagiátů, kdy správce nemusí znát přesné soubory, které prohledává, ale systém sám ukáže ty, v nichž se našly podobnosti vyhovující kritériím, které může správce stanovit a upřesnit.

K datu 28. září 2008 tak bylo do systému vloženo 22092 závěrečných prací z 5 škol v ostrém provozu, další školy k danému datu vložily testovací záznamy. Pouze 4 školy nastavily, že plné texty se nezveřejňují nikomu, naopak drtivá většina (celkem 16) škol zveřejňují metadata celému světu.

Již první kroky v ostrém provozu systému a přístupu jednotlivých škol ukazují, jak unikátní projekt se daří budovat. To potvrzuje také aktuální zájem slovenských škol, které na své národní úrovni zatím tuto možnost nemají.