CREPČ – Centrálny register publikačnej činnosti

Mgr. Ján Grman
SVOP spol. s r.o., Bratislava
grman@svop.sk

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

Abstrakt

Evidencia publikačnej činnosti je jednou z ostro sledovaných parametrov činnosti vysokých škôl. Vykazovanie tejto činnosti má odozvu nie len v oblastiach akreditácie, posudzovania grantov a projektov, ale čoraz viac sa zvyšuje vplyv publikačnej činnosti aj na rozpočet vysokých škôl. Myšlienka realizácie centrálneho registra prirodzene vyplynula z potreby efektívnej automatizácie dodávania podkladov pre hodnotenie verejných vysokých škol (VVŠ) a následnej kontroly a prezentácie týchto dát grantovým agentúram, autorom, odbornej a aj laickej verejnosti. V prvej fáze bol definovaný formát pre výmenu údajov medzi informačnými systémami VVŠ pre evidenciu EPC a centrálnym registrom. Rovnako bol vytvorený portál CREPČ a importované dáta všetkých 20 VVŠ za rok vykazovania 2007 pre potreby tvorby rozpočtu.

Praktickým dôsledkom existencie dát v jednej databáze bola schopnosť nasadiť kontrolné mechanizmy a prezentovať problémy dodaných dát na priereze univerzít s cieľom odhaliť duplicity, či iné nezrovnalosti. V týchto dňoch prebehol opakovaný import dát za rok 2007 po korekciách v lokálnych IS VVŠ na základe reportovaných problémov. V nasledujúcom období projekt CREPČ čakajú ďalšie inovácie a najmä import dát s rokom vykazovania 2008 a rozšírenie záberu kontrolných mechanizmov nie len na priereze VVŠ, ale aj na priereze rokov. Projekt je realizovaný pod gesciou MŠ SR a implementovaný riešiteľskou skupinou Univerzity Konštantína Filozofa v Nitre, Žilinskej univerzity v Žiline a firmy SVOP spol. s r.o.

Definícia cieľov

Systém vysokého školstva v Slovenskej republike tvoria verejné, štátne a súkromné vysoké školy. Medzi základné úlohy, ktoré plnia, patrí aj výskum a vývoj v rôznych oblastiach. Priamym dôsledkom a výstupom tejto činnosti je publikovanie výsledkov autormi. V období, v ktorom sa pojem „znalostná spoločnosť“ stal doslova zaklínadlom, sa výmena vedeckých informácií a najmä kvalita vedeckej práce javí ako kľúčová. Jedným z nástrojov, ktorý bol zvolený pre podporu vedeckej činnosti, bolo zvýšenie vplyvu evidovanej publikačnej činnosti v jednotlivých jej kvalitatívnych kategóriách na výšku pridelenej finančnej dotácie pre vysokú školu. Táto zmena samozrejme vytvára tlak na presnosť a kvalitu evidencie publikačnej činnosti (EPČ).

Hlavným cieľom projektu centrálneho registra na evidenciu publikačnej činnosti (CREPČ) bolo vytvorenie riešenia schopného v čo najvyššej miere automatizovať proces získavania a vyhodnocovania údajov súvisiacich s publikačnou činnosťou univerzít [1]. Tieto údaje na úrovni rezortu školstva totiž (okrem iného) slúžia ako podklad pre rozdeľovanie finančných prostriedkov jednotlivým (univerzitným) pracoviskám. Sekundárnym cieľom a dôsledkom realizácie primárnej úlohy je vznik unikátneho centrálneho zdroja informácií, ktorý umožňuje získať prehľad o činnosti všetkých zapojených pracovísk odbornej i laickej verejnosti na jednom mieste - portále.

Analýza a príprava projektu

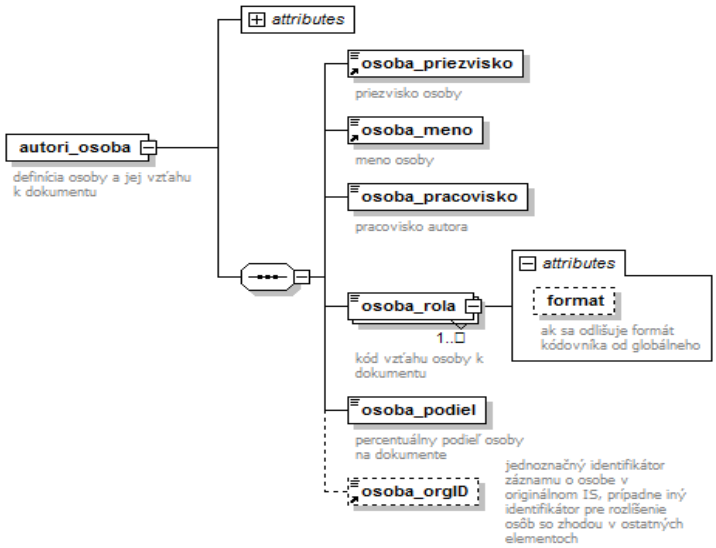
V prípravnej časti projektu bolo potrebné zistiť presný aktuálny stav. Keďže v prvej fáze mali byť do CREPČ zapojené len verejné vysoké školy (VVŠ), práve tam bol zrealizovaný dotazníkový prieskum stavu. Zisťovanie bolo zamerané na:

- celkový objem EPČ a ročný prírastok
- počty pracovníkov EPČ, spôsob zberu dát (centrálne v knižnici, formuláre pre autorov, ...)
- používaný informačný systém pre evidenciu a jeho schopnosti / možnosti v oblasti exportu dát

Ukázalo sa, že väčšina VVŠ používa na evidenciu EPČ profesionálny informačný systém (IS), typicky ako jeden z modulov riešenia vlastnej centrálnej knižnice. Niektoré VŠ používali IS vlastnej výroby a malá skupina realizovala zber a kompiláciu od autorov ručne, teda v neštruktúrovanej podobe formou tvorby výstupov v textovom editore.

Voľba výmenného formátu bola kľúčovou úlohou prípravnej fázy. Existujúce riešenia IS na VVŠ ponúkali formáty UNIMARC, MARC21, XML, či vlastné interné formáty. Ako najvhodnejší sa ukázal formát XML, a to najmä z týchto dôvodov:

- profesionálne KIS v prieskume deklarovali schopnosť produkovať XML
- pri riešeniach, ktoré si VŠ realizovali samé, išlo zvyčajne o riešenia nepracujúce priamo z formátom typu MARC a voľba niektorého formátu MARC by viedla k nutnosti zložitejších konverzií medzi MARC formátmi a dotkla by sa viac producentov
- formát XML je ľahko implementovateľný aj do systémov nie priamo zameraných na spracovanie EPČ, navyše plne rešpektuje legislatívu a odporúčania pre výstavbu štátnych informačných systémov, ich súčastí a protokolov pre výmenu dát



Obrázok 1. Časť schémy definujúcej dátové entity a ich vzťahy

Formát XML-CREPČ bol definovaný štandardne pomocou XSD schémy [2] a to s cieľom:

- striktne definovať minimálny záznam pre potreby CREPČ a to najmä: identifikáciu zdroja katalogizácie, druh dokumentu, kód kategórie dokumentu a časové zaradenie, príslušnosť záznamu k pracovisku / pracoviskám, príslušnosť záznamu k autorom (informácie o roli autora a jeho podiele na zázname)
- využiť súbor údajov pre potreby bibliografického spracovania a prezentácie záznamu,
- definovať povinné i voliteľné elementy a striktne definovať ich vzájomné vzťahy,
- minimalizovať prácnosť prevodu z používaných formátov UNIMARC a MARC21,
- zabezpečiť prevod všetkých dát nevyhnutných pre funkcie registra (povinné polia) a ďalších dát, ktoré prinášajú dôležitú informačnú hodnotu pre potenciálnych používateľov CREPČ (vedecké agentúry, akreditačne komisie, verejnosť a pod.),
- minimalizovať chyby vyplývajúce z rozdielov implementácie noriem rôznymi informačnými systémami, či dokonca rôznymi pracoviskami používajúcimi systém rovnakého výrobcu,
- zabezpečiť maximálnu unifikáciu záznamov pochádzajúcich z rôznych zdrojov a interných formátov,
- zabezpečiť formát ľahko implementovateľný aj v informačných systémoch nepoužívajúcich normu ISO 2709 a niektorý z MARC formátov,
- použiť formát, ktorý bude ľahko rozšíriteľný a modifikovateľný aj pre budúce (a dosiaľ neznáme) funkcie CREPČ alebo jeho systémové rozšírenia

Pri návrhu formátu sa kládol dôraz na funkčné potreby, ale aj na to, aby boli redukované náklady producentov dát na konverziu. V kódovníkových hodnotách napríklad pre jazyk je možné dodávať

hodnoty na báze oboch najrozšírenejších formátov MARC21 aj UNIMARC. Použitie kódovníka sa vyznačuje pre každý dátový element samostatne, čo sa na prvý pohľad javí ako zbytočné, no ukázalo sa to ako užitočné. Jeden z producentov dát používal ako nosný formát MARC21, no kódovníky v niektorých poliach používal z historických dôvodov ešte z UNIMARCu. Flexibilita XML v tejto oblasti mu značne zjednodušila situáciu.

Realizačná fáza

Na úrovni hárdvéru bol pre účely CREPČ zabezpečený hlavný a záložný server s príslušenstvom. Základná systémová úroveň sa zabezpečuje na platforme Microsoft Windows Server 2003, databázový stroj je MS SQL 2005 a samotné funkcie CREPČ vznikli modifikáciami a realizáciou nových funkcií na platforme vývojovej platformy už existujúceho knižnično-informačného systému DAWINCI.

Pilotná komunikačná schéma pre prvú dávku záznamov za obdobie 1.11.2006 až 30.10.2007 (obdobie MŠ SR volilo úmyselne z dôvodu synchronizácie na tvorbu rozpočtu) bola zvolená takto:

- VVŠ vo vlastnej réžii spracováva EPČ pracovníkov a v určenom termíne odovzdáva dátový súbor XML s dávkou pre potreby evidencie v CREPČ
- Pre upload XML dávok bolo realizované rozhranie pre každú VŠ, prostredníctvom privátnej sekcie na portáli projektu na adrese www.crepc.sk. Prostredníctvom rozhrania prebehla základná validácia dávky oproti XDS schéme a výstupom je súborný štatistický výstup odovzdaných záznamov v členení sledovaných kategórií v súlade s požiavkami smernice pre EPČ

Samotný import údajov prebehol začiatkom roka 2008 a v rovnakom čase bol vytvorený aj portál, ktorý plní tieto funkcie:

- anonymné rozhranie pre vyhľadávanie údajov a získavanie štatistík EPČ na priereze rokov, univerzít i univerzitných pracovísk, zamestnancov, kategórií publikácií a pod.,
- autorizované rozhranie na odovzdávanie dávok univerzitami
- poskytovanie hodnotenia doručených dávok prostredníctvom súborov generovaných systémom a na základe manuálnych kontrol (kontrola kódov EPČ – správnosti zaradenia)
- štatistiky pre účely MŠ SR podľa potreby a zadania

Po importe dát prebehli analytické práce s cieľom implementovať funkcie schopné detekovať zdvojené alebo viacnásobné záznamy a propagovať ich na kontrolu a vyjadrenie producentom dát. Kompletným importom dát všetkých dotknutých organizácií vznikla unikátna databáza, na základe ktorej bolo možné analyzovať typické chyby, najmä multiplicity i chyby percentuálneho rozdelenia publikácií medzi viac pracovísk (problém autorov z rôznych VVŠ). Kontroly boli zamerané na procesy:

- automatizovanej kontroly úplnosti položiek záznamov (realizovanej už vlastnosťami výmenného formátu),
- automatizovanej kontroly duplicít (duplicita publikácie v rámci odovzdanej dávky, duplicita publikácie autora v rámci viacerých VVŠ s kontrolou rozdelenia
- poloautomatizovanej kontroly údajov vo vnútri CREPČ prostredníctvom špecializovaného rozhrania pre hodnotiteľov so zámerom umožniť:
    - kontrolu zaradenia publikácie do kategórie na základe rozsahu dokumentu,
    - kontrolu zaradenia publikácie do kategórie na základe obsahu dokumentu,
    - potlačiť falošné duplicity, vyznačiť naopak skryté duplicity
    - evidovať poznámky ku kvalite popisu

Výsledkom kontroly boli reporty z podozreniami na chyby (nie všetky duplicity nimi naozaj boli), ktoré boli distribuované a dôsledkom ktorých boli opravy v systémoch VVŠ. Následne bola realizovaná opakovaná dodávka dát za rok 2007 a dodané vyjadrenie VVŠ k výsledkom kontroly.

Už základné vykonané analýzy ukázali niektoré nedostatky dát a okrem priameho dôsledku v podobe ich opráv (resp. odmazania duplicitných záznamov) odštartovali aj konštruktívnu diskusiu o samotnom procese evidencie. Ukázali sa mierne odchýlky v spracovaní a oživila sa diskusia o niektorých problematických otázkach. Typickým príkladom je aktuálna prax evidencie ohlasov v MARC formátoch, či nejasnosti okolo obdobia vykazovania záznamov pre potreby rozpočtu a podobne. Už samotné spájanie dát z viacerých zdrojov napríklad v oblasti priradenia záznamu k pracovisku implikuje diskusiu o centralizovanej správe kódovníka pracovísk a propagácií jeho zmien.

Dôsledky

Hlavným výstupom projektu je Centrálny register evidencie publikačnej činnosti (CREPČ) umiestnený na http://www.crepc.sk. Základnými vlastnosťami a výsledkami riešenia je:

- bezpečné uloženie, archivácia a jednotná prezentácia informácií,
- vyhľadávanie záznamov podľa definovaných kritérií (autor, kľúčové slová, pracovisko autora, rok vydania atď.),
- možnosť kontroly dát z viacerých pohľadov a súvislostí,
    - pohľad na priereze rokov sa prejaví až importom dát za rok 2008
    - rok 2008 sa bude dodávať na viac častí
- priame výstupy a štatistiky pre potreby manažmentu finančných prostriedkov,
- značná redukcia času od získania údajov, cez ich kontrolu, až po získanie požadovaných analytických podkladov

Nezanedbateľným dôsledkom je vznik relevantného zdroja informácií aj pre iné zložky participujúce na procesoch súvisiacich s problematikou výskumu, vývoja, ich evidencie a hodnotenia. Dohoda a zjednotenie metodiky hodnotenia EPČ umožnia používať dáta CREPČ aj grantovým agentúram (VEGA, KEGA, ...) a znížiť tak administratívne nároky na VŠ redukciou duplicitnosti vykazovania. Okrem základnej evidencie a prezentácie verejnosti a prípadne pre MŠ SR sa spravidla v inej metodike požadovali výstupy pre akreditačné agentúry, grantové projekty a podobne.

Záver

Dôležitým aspektom je zaiste aj fakt, že vďaka realizácií projektu sa oživili odborné diskusie o evidencii EPČ ako takej. Jej tlakom na vyrovnanie rozdielov v evidencii a jej skvalitnenie. Prenesene je tlakom aj na samotných autorov vzhľadom na vyššiu dostupnosť údajov, a tým väčšej verejnej kontroly ich práce.

Pri riešení stanovených úloh sme samozrejme narazili aj na problémy, resp. námety ako doplniť a vylepšit procesy centrálneho registra. Bola vytvorená pracovná skupina, ktorá sa bude v nasledujúcom období zaoberať ďalším vývojom projektu. Z diskutovaných tém je možné spomenúť:

- problém štandardizácie číselníkov pracovísk v rámci interných KIS univerzít a CREPC a propagáciu a údržbu zmien (ktoré sú pomerne dynamické),
- problematika evidencie ohlasov, ich aktualizácie s ohľadom na prax ich zápisu v MARC formátoch,
- problematika pravidiel pre kontrolu rozsahu publikácií a percentuálne podiely spoluautorov v rámci viacerých VVŠ,
- problematika automatizácie overovania karentovaných časopisov voči aktuálnej databáze

Aktuálna je aj úloha sprístupniť a zapojiť do CREPČ aj ďalšie subjekty realizujúce vzdelávaciu a vedeckú činnosť ako napríklad štátne a súkromné vysoké školy, SAV a podobne). Pozitívne je, že samotné VŠ mimo CREPČ majú záujem participovať a samé oslovujú riešiteľov s otázkou na možnosti a podmienky zapojenia sa.

V súvislosti s automatizáciou procesov a na pozadí tvorby centrálnych registrov v rezorte školstva, (register študentov, register pracovníkov a garantov, ...) projekt CREPČ plne zapadá do koncepcie budovania štátneho informačného systému. Riešiteľom sa podarilo pilotnú fázu projektu realizovať naozaj v krátkom čase a úspešne uviesť do praxe.

Referencie

[1] Skalka, Ján – Grman, Ján – Vozár, Libor: Centrálny register EPCA. Príspevok na konferencii UNINFOS 2007. 13. - 15.11.2007, EUBA, Bratislava.

[2] XML schema diagram (XSD) - http://www.w3.org/XML/Schema. Dostupné 30.4.2008.

[3] Portál CREPC - http://www.crepc.sk.