Projekt digitalizace vysokoškolských prací MU

Miroslav Bartošek, ÚVT MU

1 Úvod

V loňském roce proběhla na MU významná digitalizační akce. Jednalo se o projekt podporovaný grantem Fondu rozvoje VŠ, na kterém se podílelo šest fakultních ústředních knihoven pod vedením Knihovnicko-informačního centra MU při ÚVT. Cílem projektu bylo výrazně rozšířit objem vysokoškolských kvalifikačních prací – VŠKP (disertačních, diplomových, bakalářských) přístupných on-line, a tím na jedné straně zlepšit jejich dostupnost uživatelům, na druhé straně snížit provozní náročnost knihoven (uchovávání, půjčování, manipulace s VŠKP).

Masarykova univerzita si udržuje vůdčí roli mezi všemi univerzitami v ČR při přechodu k elektronickým vysokoškolským kvalifikačním pracím a při jejich širokém zpřístupňování uživatelům. Jako jedna z mála našich vysokých škol má propracovanou legislativu a infrastrukturu, která umožňuje – a také vyžaduje – vytváření a povinné odevzdávání všech VŠKP v elektronické podobě, a také jejich efektivní zpřístupnění uživatelům. Všechny práce odevzdané po 1. lednu 2006 jsou zveřejněny volně na internetu prostřednictvím Archivu závěrečných prací v IS MU, `http://is.muni.cz/thesis` resp. `https://is.muni.cz/auth/lide/absolventi.pl` (v autentizované části IS MU), viz [1]. Práce odevzdané v elektronické podobě před tímto datem mohou být zveřejněny pouze uvnitř MU, tj. v autentizované části IS MU, pokud autor neposkytl explicitní souhlas se zveřejněním na internetu.

Přestože některé fakulty zavedly odevzdávání VŠKP v elektronické podobě již před rokem 2006, nebylo pokrytí starších let příliš rozsáhlé. Navíc nebyly sjednoceny formáty a elektronické verze byly často odevzdány pouze jako CD/DVD příloha tištěných prací. Projekt FRVŠ na digitalizaci VŠKP MU usiloval o rozšíření on-line přístupu k VŠKP zpětně až do roku 2001 (v prosinci 2000 vstoupila v platnost novela autorského zákona 121/2000 Sb., která teprve upravila zacházení s tzv. školními díly). S tvůrci IS MU bylo dohodnuto, že digitalizované práce nebudou vystavovány samostatně, ale budou přihrány k již existujícím pracím v Archivu elektronických prací IS MU.

2 Postup řešení projektu

Do projektu se zapojilo celkem šest fakult – byly to všechny fakulty MU kromě Ekonomicko-správní fakulty (ta měla fond svých VŠKP kompletně zdigitalizovaný již dříve – viz [2]), Fakulty informatiky (neprojevila o digitalizaci starších prací zájem) a Fakulty sportovních studií (malý objem prací). Každá fakulta si určila sama podle svých priorit a potřeb typy a časové rozpětí VŠKP k digitalizaci. V ústředních knihovnách zapojených fakult byla zřízena digitalizační minicentra vybavená počítačem a stolním skenerem s automatickým podavačem. Všude tam, kde to bylo možné, byly digitalizované práce rozřezány a volné listy se skenovaly dávkově přes automatický podavač. Proces skenování se tím velmi urychlil (stovky stran za hodinu oproti desítkám stran při ručním obracení stránek). Práce byly skenovány v rozlišení 300 dpi, v bitonálním režimu (pouze barevné přílohy se skenovaly barevně) a skenery byly nastaveny tak, aby vytvářely přímo jeden pdf-soubor pro celou digitalizovanou práci.

Pro řízení a podporu digitalizačního procesu byla vytvořena speciální webová aplikace, která zajišťovala celý digitalizační postup:

- Přenos naskenované práce (případně i posudků a příloh na CD/DVD) po síti do centrálního meziúložiště na ÚVT.
- Dohledání bibliografického záznamu práce v knihovním systému Aleph-MU (typicky na základě čárového kódu tištěné práce).
- Propojení bibliografického záznamu práce s odpovídajícím záznamem studia ve studijní evidenci IS MU. Protože jde o dvě rozdílné evidence, které nemají žádný společný jednoznačný identifikátor, musela aplikace využívat různých heuristik pro správné automatizované přiřazení co největšího množství prací.

fakulta	BP	DP	disertace	ostatní	prací celkem	stran celkem
FF	468	1 237	180	13	1 898	209 008
FSS	519	403	0	0	922	80 737
LF	252	71	193	19	535	42 694
PřF	1	348	212	0	561	55 268
PraF	2	359	20	23	404	42 524
PedF	0	311	52	2	365	40 656
Celkem	1 242	2 729	657	57	4 685	470 887

Tabulka 1: Počty digitalizovaných prací

Problém spočíval v tom, že bibliografické záznamy VŠKP neobsahovaly jednoznačný identifikátor autora (UČO), a naopak IS MU neobsahoval u všech studií před rokem 2003 informace o závěrečné práci, nebo se existující údaje (název práce) neshodovaly vždy přesně s těmi v knihovním katalogu. Ve spojení s tím, že za období 2001-2006 absolvovaly na MU tisíce studentů (někdy stejného jména a příjmení i v rámci stejné fakulty, roku a oboru), u žen docházelo ke změnám příjmení, a některé osoby absolvovaly (úspěšně či neúspěšně, někdy i opakovaně) více studií ve stejném nebo různých oborech ve stejném nebo různých letech, představovalo automatizované propojení bibliografických záznamů VŠKP se záznamy studií v IS MU netriviální problém. V případech, kdy systém nebyl schopen spolehlivě propojení určit, nabídl možnost ručního dohledávání na základě různých množin společných nebo podobných znaků. I tak zůstalo několik desítek prací, u nichž přiřazení správného autora a jeho studia vyžadovalo téměř detektivní práci.

- Kontroly úplnosti a správnosti propojení.
- Statistiky a přehledy postupu prací na jednotlivých fakultách a u jednotlivých pracovníků/studentů najatých na digitalizaci.
- Export prací a jejich metadat z meziúložiště a předání pro import do Archivu závěrečných prací IS MU.

3 Počty digitalizovaných prací

Skenování prací na fakultách bylo zahájeno v květnu 2007. Do konce roku 2007 bylo zdigitalizováno 4 685 prací o rozsahu 470 887 stran textu. Přehled po jednotlivých fakultách uvádí tabulka 1 (DP značí diplomová práce, BP je bakalářská práce):

Z celkového počtu bylo čtyřicet procent všech prací zdigitalizováno na Filozofické fakultě MU. Relativně nižší počet prací na PedF a PraF byl způsoben vyšší pracností digitalizace, protože velkou část VŠKP na těchto fakultách nebylo možno rozřezat a musely se digitalizovat s ručním obracením stránek. Na některých fakultách pokračuje digitalizace i v roce 2008.

4 Zpřístupnění digitalizovaných prací

Jak již bylo uvedeno výše, pro zpřístupnění digitalizovaných prací bylo využito stávajícího Archivu závěrečných prací v IS MU. Pro převod prací z meziúložiště ÚVT do Archivu IS MU byl vytvořen speciální importovací profil. Součástí importu bylo zpracování digitalizované práce programem OCR. Ze souboru `scan.pdf` (výsledek digitalizace) byly automaticky vygenerovány soubory `text.pdf` (dvouvrstvé pdf obsahující jak obrázky stran tak rozpoznaný text – pro možnost vyhledávání v textu práce) a `text.txt` (holý text). Tyto tři soubory, spolu s popisnými metadaty a posudky (pokud existují v digitální podobě) jsou uloženy společně v adresáři dané VŠKP v Archivu závěrečných prací IS MU.

Po nahrání digitalizovaných prací do Archivu byly v knihovním systému Aleph-MU vygenerovány odkazy z bibliografických záznamů VŠKP na odpovídající plné texty v Archivu. Uživatelé tak mají možnost vyhledávat a dostat se k plným textům závěrečných prací prostřednictvím dvou různých systémů:

- Archivu závěrečných prací IS MU,

– celouniverzitního knihovního systému Aleph-MU.

VŠKP vytvořené po 1. 1. 2006 jsou dostupné komukoliv na Internetu, práce vytvořené před tímto datem jsou dostupné pouze autentizovaným uživatelům IS MU.

Co se týče samotného Archivu závěrečných prací IS MU: koncem roku 2007 obsahoval přes 40 000 bibliografických záznamů VŠKP a v 19 000 případech byl k dispozici i plný text práce v digitální podobě (ať již vznikl přímo jako born-digital nebo digitalizací).

Literatura

[1] J. Brandejsová. *Zveřejňování závěrečných prací v IS MU.* Zpravodaj ÚVT MU. ISSN 1212-0901, 2006, roč. XVII, č. 1, s. 12–14.

[2] J. Nekuda, J. Poláček. *Elektronické diplomky a bakalářky na ESF MU: dokončená mise.* Zpravodaj ÚVT MU. ISSN 1212-0901, 2006, roč. XVI, č. 3, s. 1–3. □