DIGITALIZAČNÍ PROJEKT DML – CZ

Oldřich Ulrych a Jiří Veselý

Univerzita Karlova, Matematicko-fyzikální fakulta,
Sokolovská 83, 186 75 Praha 8 – Karlín

Pro mnoho z nás je ideálem mít všechny potřebné knížky a časopisy za zády na polici knihovny. Je to ideál prakticky nedostižný, i když to možná již brzo platit nebude. V posledních letech jsme svědky toho, jak den ze dne přibývají na síti elektronické verze časopisů – je to velmi příjemné, ale často to má jednu podstatnou vadu. Za přístup k informacím je třeba platit nezanedbatelné částky. Řada velkých hráčů na tomto trhu s informacemi usiluje o získání práv k plným textům časopisů, které pak digitalizují a zpřístupňují na komerční bázi. I když si na některá místa zatím počítač těžko vezmete a užití klasické knižní formy je vám příjemnější, práce s digitalizovanými časopisy se stává postupně stále jednodušší. Nabízí také řadu dalších výhod, které zdaleka nejsou zanedbatelné – snadné a rychlé vyhledávání, menší prostorové i finanční nároky a dostupnost všude tam, kde se můžete připojit k Internetu.

Zároveň se objevuje jiný trend: získat nezávislost na těchto placených zdrojích informací a zpřístupnit informace časopisecké i knižní povahy zadarmo či za podstatně lepších finančních podmínek. Všimněme si tedy blíže tohoto trendu v oblasti matematiky. Popišme několik příkladů: francouzské matematické časopisy byly digitalizovány v rámci projektu NUMDAM (Grenoble) (URL viz [NU]). Novější ročníky řady matematických časopisů jsou volně přístupné prostřednictvím EMISu (viz [EM]). Mnoho časopisů (též dokonce i některých českých) a mnoho svazků klasických matematických prací zpřístupnilo během posledních let digitalizační centrum v Göttingen (viz [GO]). Projekty stojící v pozadí i zaměření a rozsah těchto snah jsou odlišné, ale jsou součástí úsilí, skrývajícího se pod zkratkou WDML (World Digital Mathematics Library). Pokud jsme nahlédli na [WD], zjistili jsme v době vzniku této informace, že WDML momentálně zpřístupňuje 2249 knih v elektronické formě ($\geq$ 515650 stránek) a 223 periodik (časopisy, semináře, $\geq$ 4051328 stránek). Seznam nejdůležitějších poskytovatelů pro matematiku relevantních zdrojů nalezne čtenář např. na [WD1].

U neziskového portálu JSTOR je to již s přístupem složitější, řada čtenářů ho však patrně měla možnost využít. Ten mj. plní i archivační roli a zpřístupňuje starší ročníky mnoha časopisů až k tzv. *moving walls*, což jsou meze, ke kterým se

Podporováno projektem 1ET200190513 *DML-CZ: Česká digitální matematická knihovna* financovaného v rámci programu "Informační společnost" Akademie ČR (Národní program výzkumu a vývoje TP2, 2005-2009).

zveřejnění plných textů časopisů provádí: 90 % časopisů na JSTOR má tuto bariéru mezi 0 – 10 roky od zveřejnění (viz ev. [JS]).

Také u nás se snažíme k tomuto celosvětovému trendu přispět: V projektu DML-CZ spojili své síly pracovníci Masarykovy univerzity v Brně, Akademie věd ČR a Univerzity Karlovy v Praze k postupnému zpřístupnění našich matematických časopisů a dalších pramenů v digitální podobě. Zdánlivě přímočará taktika „naskenovat a vystavit“ zdaleka však nepostihuje vše to, co je třeba udělat, aby vložená práce a prostředky přinášely uživatelům maximální užitek.

Nestačí se pouze seznámit se zahraničními postupy, je třeba je zhodnotit a upravit tak, aby byla zachována kompatibilita s ostatními zdroji a přitom byla respektována specifika oboru – stačí orientace na černobílý tisk, avšak je třeba počítat se značnou jazykovou variabilitou, s nutností propojení na databáze ZMath a MathSciNet referativních časopisů Zentralblatt für Mathematik (a Jahrbuch über die Fortschritte der Mathematik) a Mathematical Reviews, se zachováním knihovnických standardů apod. Je nutno se věnovat právním aspektům zveřejňování, postarat se o možnosti rychlého a spolehlivého vyhledávání, o archivaci materiálů, jejich účelnou strukturu pro tisk i vystavování – k tomu bylo třeba vytvořit editační nástroje i pomůcky pro tvorbu metadat (tj. informací charakterizujících článek ve vztahu k časopisu, uživateli i struktuře, v jaké je v knihovně DML-CZ uložen) včetně pravidel, co tato metadata a jak budou zahrnovat (a co nikoli), jaká transkripční pravidla se budou užívat, velkou část titulů článků přeložit do angličtiny apod.

Kromě uvedených vcelku zjevných problémů je třeba řešit i další, které jsou skryty v pozadí či které se vynořují někdy vcelku neočekávaně. Obecnou snahou je relevantní informace maximalizovat a co nejvíce zpřesnit. To znamená např. uvádění plných jmen vždy, kdy je to možné, je však nutné se kdesi v pozadí vypořádat s tím, že některá jména byla psána různým způsobem, že je často složité rozlišit jméno a příjmení (u čínských jmen) či že vyhledávač musí spolehlivě pracovat i se jmény různě modifikovanými (a to nejen jednoduše zbavenými všech akcentů). Celý proces skenování musí být co nejméně závislý na lidském faktoru; např. je užitečné automaticky rozeznat číslo stránky a spojit je s obrázkem. Starší články neobsahují údaj o MSC klasifikaci. Každý článek je nutno v metadatech opatřit též anglickým překladem názvu – zde se neobejdeme bez pomoci širšího okruhu specialistů v různých oblastech matematiky, kteří nám obětavě pomáhají. A takových problémů je hodně, nepoměrně více, než je uvedených příkladů.

Zdálo se například snadné získat metadata prostřednictvím výše zmíněných databází, avšak ty obsahují poměrně velké procento chyb. Navíc struktura a způsob vytváření záznamů v těchto databázích se s časem měnily – zde naopak budeme moci přispět k jejich opravám a zúplnění. V pozadí za obrázky stránek slouží k vyhledávání digitalizované texty pořízené OCR (Optical Character Recognition) programy, je však nutné, aby tyto programy pracovaly co nejspolehlivěji: měly by rozeznat jazyk textu a později i vzorce. Tento „rozeznaný text“ je třeba propojit s obrázky stránek. Je potřebné zajistit co nejjednodušší přístup k pracím, uvedeným v citacích.

Do r. 2009 by mělo být zpracováno a zpřístupněno v rámci projektu 150 – 200 tisíc stran. Testovacím materiálem, na němž se účastníci projektu s touto problematikou seznamovali a postupně vytvářeli zázemí pro budoucí masovější využití užitých technologií, byl časopis Czechoslovak Mathematical Journal. Dnes je jich již o trochu

více a zpracovávání dalších stále probíhá. Informace o projektu a stručnou informaci o zpracovávání lze též získat na URL [DM]. Účastníci projektu uvítají náměty a připomínky sdělené emailem na adresu [DME]. V první polovině příštího roku si náš budoucí příspěvek k WDML budete moci detailně prohlédnout, příslušné URL na stránkách Pokroků oznámíme.

Vše, co bylo řečeno, se týká retrodigitalizace, tedy textů publikovaných v předpočítačové době horké sazby. Odtud plyne další problém: vše je nutno sladit i s prací s texty, které již jako digitalizované do dneška vznikly a budou dále vznikat. U nich je třeba zajistit jejich postupné průběžné vystavování i po dokončení projektu. O změnách, týkajících se redakčního zpracovávání českých matematických časopisů a z toho plynoucích úprav při přípravě článků k uveřejnění, proběhl v Praze již seminář s prof. Rossem Moorem, který je autorem komplexního řešení analogického problému v případě australských matematických časopisů. Jiné řešení poskytuje v rámci projektu NUMDAM projekt CEDRAM, který využívají zejména menší nakladatelé. Hlavní jeho myšlenkou je, že v redakcích se při výběru článků, recenzním řízení a přípravy k tisku automaticky připraví i vše nezbytné pro budoucí vystavení v elektronické podobě. Také o těchto změnách budeme čtenáře včas informovat.

Odkazy na síťové zdroje:

[NU] `http://www.numdam.org/`.

[EM] `http://www.emis.de/`.

[GO] `http://gdz.sub.uni-goettingen.de/`.

[WD] `http://www.ceic.math.ca/WDML/dml/` .

[WD1] `http://www.wdml.org/`.

[JS] `http://www.jstor.org/`.

[DM] `http://dml.muni.cz/`.

[DME] `email pro pripomínky`.