Vize budoucího zpracování časopisů digitalizovaných v DML-CZ

Motivační studie

Petr Sojka

Fakulta informatiky, Masarykova univerzita v Brně
sojka@fi.muni.cz

1 Motivace

Již třetím rokem úspěšně běží projekt české digitální matematické knihovny DML-CZ [6,1]. Má být naším příspěvkem k uskutečnění vize celosvětové World Digital Mathematics Library (WDML) [3]. Cílem je zajistit dostupnost matematické literatury publikované v českých zemích v elektronické podobě. Projekt, řešený v letech 2005–2009, je podporován grantem Akademie věd ČR v rámci programu Informační společnost. Celkový rozsah projektu je odhadován na 200 000 stran dostupných v elektronické podobě. Webová informační stránka projektu je `http://dml.muni.cz`.

Ještě letos by se na `http://www.dml.cz` měly objevit první digitalizované časopisy, počínaje pilotním Czech Mathematical Journal (CMJ) z let 1950 až 1991. Tento časopis bude v dalším textu používán jako ilustrační příklad diskutovaných postupů.

2 Vývoj přípravy matematických časopisů

V historii časopisů lze identifikovat nejméně tři etapy.

1. Sazba je pořizovaná horkou sazbou. Elektronická podoba je možná pouze digitalizací. U CMJ léta 1950 až 1991.
2. Sazba je pořizovaná sice elektronicky (born-digital), typicky v TEXu, ale hlavním a většinou jediným výstupem je předloha pro tiskárnu. Pro CMJ je to období od 1992 do dneška (i když od cca 1996 vzniká část metadat pro web a archivuje se PostScript jednotlivých článků).
3. Sazba a celý redaktorský proces je prováděn elektronicky tak, že zároveň s přípravou předlohy vznikají všechna potřebná data i metadata pro prezentaci článků časopisu v digitální knihovně resp. repozitáři nakladatele. Data vznikají jako „vedlejší produkt“ elektronického zpracování časopisu, a nepředstavují při dodržení standardního značkování primárních zdrojů článků zdražení produkce a zatížení redakce časopisu. U CMJ je přechod na tento typ zpracování ve fázi specifikace značkování a rozhraní pro import dat do repozitáře DML-CZ.

Autoři dnes již dobře ví, že jejich citační index (h-index, g-index) roste mnohem rychleji, když jejich publikace jsou relativně rychle k dispozici v elektronické podobě [4]. Některé časopisy proto lákají autory a zvyšují svou kvalitu a impakt faktor tím, že na omezenou dobu po vydání čísla časopisu jsou články volně k dispozici v elektronické podobě na Internetu.

Kvalitní webová prezentace časopisu a poskytování digitálního obsahu ve vhodné formě (metadata přes OAI-PMH, data v podobě vhodné k indexaci v Google Scholar či Google Books) stojí nemalé úsilí a množství peněz. Starosti o aktualizace, zálohování, internet konektivitu, datové úložiště, podporu nových digitálních formátů, konverze, indexaci atp. vyžadují v dnešních rychle se měnících podmínkách specializované odborníky, o kterých je v redakcích menších nakladatelství nouze. I velké nakladatelské domy jako Springer zpracování časopisů zajišťují outsourcingem, například v Indii (River Valley Technologies) nebo Litvě (VTEX).

3 Možné perspektivy řešení

DML-CZ se v mnohém inspiruje projektem NUMDAM, v rámci kterého bylo digitalizováno již přes půl miliónu stran. Pro usnadnění přidávání nových ročníků do digitální knihovny NUMDAMu byl navržen projekt CEDRAM, který využívají zejména menší nakladatelé. Hlavní idea je ta, že redakce se koncentrují na výběr článků a kvalitní recenzování a přípravu podkladů čísla v předem dohodnutém formátu. Pak přebere zdrojová data CEDRAM (Cellule MathDoc), který zároveň s přípravou předloh pro tisk vše importuje do digitálního repozitáře včetně všech metadat, nalezení Zbl a MR identifikátorů všech citací článků atd. Technické aspekty postupu zpracování jsou popsány v článku [2] a prezentaci The cedram journal production system.

Podobné řešení zvolila Australská matematická společnost, když zpracováním svých časopisů pověřila prof. Rosse Moore, který vyvinul vysoce automatizovaný postup zpracování článků [5], abstrakt.

Oba výše zmíněné postupy umožňují za cenu vhodného značkování vstupních dat bezchybná (meta)data článků včetně bibliografických citací a plných textů článků pro další využití v digitálním repozitáři.

4 Shrnutí

Shrnuli jsme podstatné aspekty, které argumentují proto, aby byl zefektivněn proces přípravy časopisů dostupných v DML-CZ tak, aby po skončení projektu bylo možné bezproblémové vkládání nových čísel do vytvořeného digitálního repozitáře. Článeček byl sepsán s vírou, že aspoň částečná shoda na způsobech zpracování v redakcích časopisů zahrnutých v DML-CZ umožní ekonomicky schůdné plnění a využití repozitáře DML-CZ i po ukončení projektu v roce 2009.

Reference

1. Miroslav Bartošek, Martin Lhoták, Jiří Rákosník, Petr Sojka, and Martin Šárfy. DML-CZ: The Objectives and the First Steps, September 2006. accepted for publication as book chapter CMDE 2006 (A.K. Peters).
2. Thierry Bouche. A pdfLaTeX-based automated journal production system. *TUGboat*, 27:45–50, 2006.
3. Allyn Jackson. The Digital Mathematics Library. *Notices of the AMS*, 50(4):918–923, 2003.
4. Adam Langley and Dan S. Bloomberg. Google Books: Making the public domain universally accessible. In *Proceedings of SPIE — Volume 6500, Document Recognition and Retrieval XIV*, pages 1–10, San Jose, CA, January 2007. The International Society of Optical Engineering. `http://www.imperialviolet.org/binary/google-books-pdf.pdf`.
5. Ross Moore. Extending the Reference Web using modern TeXniques. In *Proceedings of ICIAM 2007, the Sixth International Congress on Industrial and Applied Mathematics*, Zurich, Jul 2007.
6. Petr Sojka. From Scanned Image to Knowledge Sharing. In Klaus Tochtermann and Hermann Maurer, editors, *Proceedings of I-KNOW '05: Fifth International Conference on Knowledge Management*, pages 664–672, Graz, Austria, June 2005. Know-Center in coop. with Graz Uni, Joanneum Research and Springer Pub. Co.