Česká digitální matematická knihovna

Miroslav Bartošek, ÚVT MU

Koncem roku 2009 byla uživatelům z celého světa zpřístupněna plná verze České digitální matematické knihovny `http://dml.cz` – nejrozsáhlejšího a nejpropracovanějšího systému digitální knihovny, na jejímž vzniku se Ústav výpočetní techniky MU dosud podílel. Knihovna zpřístupňuje na 26 000 vědeckých matematických článků (přes 275 000 stran textu) z časopisů, sborníků a monografií vydaných na území České republiky od poloviny 19. století do současnosti a zahrnuje tak podstatnou část české matematické odborné literatury. Kromě plných textů článků ve formátu PDF (většinou volně dostupných komukoliv) nabízí knihovna podrobná článková metadata včetně bibliografických referencí (seznamu použité literatury), věcnou klasifikaci všech článků podle systému MSC (Mathematics Subject Classification), propojení článků na jejich recenze ve světových referenčních databázích MathSciNet a Zentralblatt MATH, nabídku obsahově příbuzných článků podle míry podobnosti vypočtené na základě strojové analýzy textu a řadu dalších vymožeností. To vše je zpřístupněno v rámci digitální knihovny poskytující bohaté možnosti procházení obsahu podle různých rejstříků a vyhledávání v metadatech a plných textech.

Česká digitální matematická knihovna (zkratka DML-CZ) je výsledkem projektu řešeného v letech 2005-2009 v rámci programu *Informační společnost* podporovaného Akademií věd ČR (projekt č. 1ET200190513) [1]. Na projektu se podíleli informatici, matematici, knihovníci a studenti z pěti akademických pracovišť: Matematického ústavu AV ČR v Praze (vedení projektu, výběr materiálu, autorsko-právní problematika, matematický popis), Knihovny AV ČR (digitalizace tištěného materiálu), Matematicko-fyzikální fakulty UK v Praze (sklízení a zpracování metadat z matematických referenčních databází), Fakulty informatiky MU (pokročilé technologie OCR, strojová textová analýza, zpracování born-digital materiálů) a Ústavu výpočetní techniky MU (vývoj softwarových nástrojů, technická koordinace, implementace a provoz vlastní digitální knihovny).

Projekt DML-CZ byl unikátní v mnoha směrech. Nešlo v něm jen o běžnou digitalizaci; součástí projektu byl i výzkum a vývoj mnoha pokročilých podpůrných technologií: OCR matematických textů s rozpoznáváním matematických výrazů, textová analýza pro vyhledávání podobných článků s využitím metod strojového učení, metadatový editor na komplexní podporu všech činností při tvorbě článkově orientované digitální knihovny, systém pro automatizovanou tvorbu nových digitálních časopiseckých čísel na bázi TeXovských technologií a jejich import do DML-CZ, prezentační systém digitální knihovny nad univerzálním repozitářovým systémem DSpace a další. Některé z těchto technologií nalezly již své uplatnění i v dalších digitalizačních projektech v rámci ČR a MU (například metadatový editor je využíván jako základní nástroj v projektu digitalizace sborníků prací Filozofické fakulty MU). Připravuje se i jejich nasazení v právě zahajovaném projektu evropské digitální matematické knihovny EuDML.

K obsahu DML-CZ

Matematické texty představují obecně velmi vhodný materiál pro digitalizaci a široké zpřístupnění, bez ohledu na věk. Mají trvalou hodnotu, starší výsledky nejsou nahrazovány novými, ale tvoří jejich základ. Hodnota matematické literatury je podmíněna její celistvostí, umocněnou vzájemným provázáním prostřednictvím referencí a dalších odkazů. Je prokázáno, že asi polovina citací v současných matematických pracích směřuje k literatuře staré alespoň deset let a čtvrtina citací k literatuře starší než dvacet let. Matematickou literaturu hojně využívají i nematematici. To vše jsou charakteristiky, které matematiku odlišují od jiných oborů a které ukazují, proč by měla být matematická literatura v co největším rozsahu pečlivě archivována, indexována, uchovávána a zpřístupňována v dlouhodobém horizontu [2]. Z pohledu tvůrců digitální knihovny k tomu přistupují ještě další příjemné vlastnosti, mezi něž patří zejména vysoký stupeň standardizace a organizovanosti matematické literatury a systému

jejího publikování v národním i celosvětovém měřítku.

Aktuální verze DML-CZ nabízí tři typy dokumentů – odborné matematické časopisy, sborníky konferencí a monografie. V časopisecké části je zastoupeno 10 nejvýznamnějších českých a jeden slovenský matematický časopis. Patří mezi ně i časopis *Archivum Mathematicum* vydávaný Přírodovědeckou fakultou Masarykovy univerzity, stejně jako například *Časopis pro pěstování matematiky a fyziky* – první matematický časopis vydávaný (od roku 1872) v zemích tehdejšího Rakousko-Uherska. Každý časopis je v digitální knihovně dostupný od svého prvního čísla až po současnost. Ve sborníkové části je zařazeno pět kompletních konferenčních řad, včetně významné mezinárodní konference EQUADIFF o diferenciálních rovnicích a jejich aplikacích, pořádané od roku 1962 střídavě brněnskou univerzitou, Matematickým ústavem AV ČR v Praze a Komenského univerzitou v Bratislavě. Monografická část digitální knihovny pokrývá především práce Bernarda Bolzana – historicky patrně nejvýznamnějšího matematika působícího na našem území, ale i několik vybraných knih předních novodobých českých matematiků, například Čechovy *Bodové množiny*. Zařazena je také kolekce monografií *Dějiny matematiky* (zatím několik prvních svazků, postupně budou doplňovány další) poskytující třeba středoškolským učitelům matematiky materiál ke zpestření výuky historickými komentáři a souvislostmi.

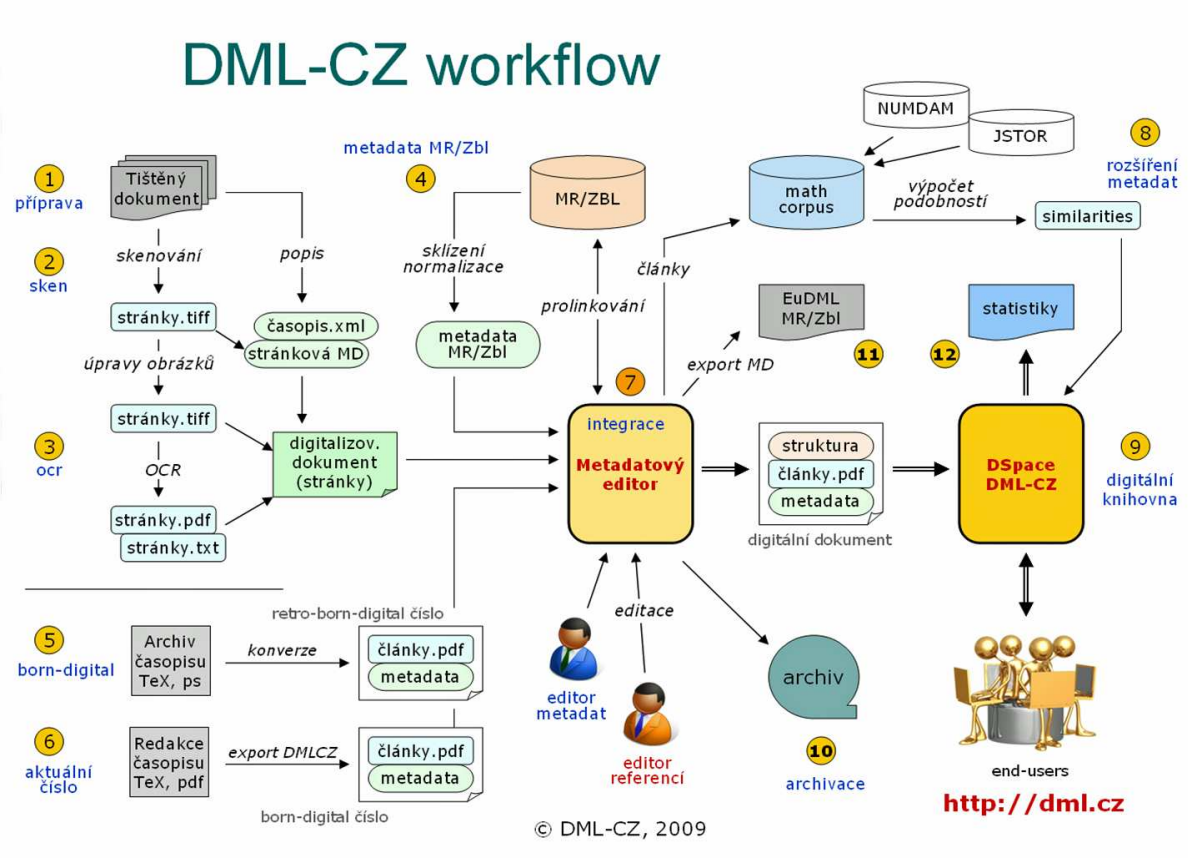
Obsah digitální knihovny není uzavřen. Třebaže vlastní projekt skončil, digitální knihovna DML-CZ bude dál trvale rozvíjena a doplňována (vlastníkem a koordinátorem digitální knihovny je Matematický ústav AV ČR, provoz a údržbu zajišťuje Ústav výpočetní techniky MU). Průběžně budou přidávána nově publikovaná čísla časopisů a sborníky konferencí, souběžně budou ale zařazovány i nové dokumenty – v závislosti na získaných finančních prostředcích a ošetření autorských práv.

Jak se dělá taková digitální knihovna

Police plné knih ještě netvoří knihovnu a disk plný dat ještě nepředstavuje digitální knihovnu. Od klasické knihovny se očekává, že umožní spolehlivé uchování dokumentů, snadnou navigaci, rychlé vyhledání požadovaného dokumentu. U digitální knihovny k tomu přistupují další funkcionality umožněné digitálním formátem dokumentů: rychlé třídění podle zvolených kritérií, full-textové prohledávání, více či méně snadné zpracování textů, vzájemná provázanost dokumentů a jejich propojení s bibliografickými databázemi – a samozřejmě vzdálený nepřetržitý přístup [2].

K vytvoření takové digitální knihovny je zapotřebí poměrně složitý postup zpracovávající různé druhy dokumentů a různé formy vstupních dat. Materiály zařazované do DML-CZ pocházejí ze tří publikačních období a jsou zpracovávány rozdílnými způsoby:

1. *Tištěné dokumenty*: jde o časopisy, monografie a sborníky vydané zhruba před rokem 1990, které existují pouze v tištěné podobě. Tyto dokumenty jsou skenovány, obrazy stránek jsou digitálně zpracovávány (prahováním, odstraňováním šumu, narovnáním, rozpoznáváním textu), následně jsou stránky seskupovány do článků a jsou generována dvouvrstvá článková PDF obsahující obrazovou a textovou vrstvu, pro jednotlivé články jsou vytvářena popisná metadata a jsou zpracovávány reference.
2. *Starší digitální dokumenty*: materiály od počátku devadesátých let do současnosti. Tyto dokumenty již existují jako články v „nějaké" digitální podobě, takže není třeba je skenovat a zpracovávat jednotlivé stránky. Avšak zdrojové digitální podklady (texty článků vysázené ve formátu TeX) i jejich výsledná prezentační forma (soubory ve formátu pdf nebo postscript) nejsou jednotné a často se v průběhu doby několikrát měnila jejich struktura i v rámci jednoho seriálu (časopisu nebo sborníkové řady). Takže je nezbytné konvertovat je do požadovaného jednotného tvaru. Samotná metadata článků lze obvykle extrahovat z digitálních podkladů (ne vždy jsou ovšem k dispozici), je však třeba zohledňovat přitom specifika použité sazby.
3. *Nové digitální dokumenty*: v průběhu řešení projektu byl pro jednotlivé redakce vydávající



Obrázek 1: Schéma postupu při vytváření DML-CZ

časopisy zařazené do DML-CZ vytvořen systém, který umožňuje, aby při přípravě nového čísla pro tisk byla automaticky vygenerována i digitální forma připravená pro import a začlenění do digitální knihovny DML-CZ. Zařazování nově vydávaných časopiseckých čísel do DML-CZ pak může probíhat automatizovaně, bez nutnosti náročné ruční práce a složitých konverzí.

Veškeré podklady pro DML-CZ (získávané kterýmkoliv z výše uvedených způsobů) jsou soustřeďovány a zpracovávány ve speciálně vytvořené webovské aplikaci – Metadatovém editoru, který je integračním centrem všech aktivit při vytváření článkově orientované digitální knihovny. Na vlastním zpracování dat se v Metadatovém editoru podílejí různí řešitelé (včetně spolupracujících studentů) s různým stupněm oprávnění k povoleným činnostem. Po kompletaci celého digitálního dokumentu, zkontrolování správnosti a úplnosti všech jeho komponent a po doplnění vazeb na jiné dokumenty jak v rámci DML-CZ, tak i mezinárodním kontextu (vazby na světové matematické referenční databáze a jiné digitální matematické knihovny), je dokument importován do digitální knihovny, jejímž prostřednictvím je zpřístupněn koncovým uživatelům. Jako digitální knihovna je použit univerzální repozitářový systém DSpace, nad nímž byla vytvořena aplikační a prezenční vrstva speciálně pro potřeby DML-CZ.

Přehledné schéma postupu při vytváření DML-CZ je uvedeno na obrázku 1.

Od DML-CZ ke světové digitální matematické knihovně

Snem matematiků je vytvoření světové digitální matematické knihovny, která by zpřístupňovala lidstvu veškerou existující hodnotnou matematickou literaturu. Odhadovaný rozsah současné existující matematické literatury představuje méně než 100 miliónů stran textu [2]. Může se to zdát hodně, avšak dramatické pokroky

v masové digitalizaci spolu s úspěšným nasazením velmi rozsáhlých digitálních knihoven v posledních letech (vzpomeňme jen Google Book Search, ArXiv.org či JSTOR) ukazují, že sen matematiků není nereálný a je dosažitelný již soudobými technologiemi.

Původní představy o vytvoření světové digitální matematické knihovny „shora", prostřednictvím jednoho velkého projektu, se ukázaly jako neschůdné. Matematikové jsou však vytrvalí a svého snu se nevzdali. Začali budovat části své velké knihovny postupně zdola. Digitální knihovna DML-CZ je jednou z takových částí. Nevznikla jako izolovaný národní systém, ale již od počátku byla vytvářena tak, aby mohla být snadno zapojena do většího celku. Při svém vzniku se inspirovala zkušenostmi z obdobných zahraničních projektů (zejména francouzským systémem NUMDAM [3]) a implementovala všechny podstatné náležitosti, které její zapojení do světové matematické knihovny podporují: je plně propojena do světových on-line matematických referenčních databází MathSciNet a Zentralblatt MATH; nabízí anglické uživatelské rozhraní a i pro neanglické články jsou poskytována základní metadata v angličtině; jsou dodržovány mezinárodní standardy pro zápis matematických výrazů; jsou podporovány technické standardy pro interoperabilitu v rámci digitálních knihoven.

Dalším krokem směrem ke světové matematické knihovně je projekt na vytvoření Evropské digitální matematické knihovny EuDML přijatý koncem loňského roku (řešení projektu se zahajuje právě v těchto dnech). Tento tříletý projekt získal grantovou podporu evropského programu ICT Policy Support Programme a účastní se ho 14 partnerů z 9 evropských zemí. Do projektu jsou zapojeni také tvůrci České digitální matematické knihovny – a samozřejmě i knihovna DML-CZ samotná.

Literatura

[1] Webové stránky projektu DML-CZ, `http://project.dml.cz`.

[2] Jiří Rákosník. *DML-CZ: Česká digitální matematická knihovna*. Sborník semináře „Matematika na vysokých školách". Herbertov u Vyššího Brodu, 2009. `http://project.dml.cz/docs/herbertov2009_rakosnik.pdf`

[3] Francouzská matematická digitální knihovna NUMDAM – Numérisation de documents anciens mathématiques. `http://www.numdam.org/` □