Miroslav Bartošek

DML-CZ: ČESKÁ DIGITÁLNÍ MATEMATICKÁ KNIHOVNA V ROCE 2011

Recenzenti článku:
Ing. Milan Janíček
Mgr. Jan Hutař

Abstrakt :

Česká digitální matematická knihovna (DML-CZ) je výsledkem spolupráce tří českých akademických pracovišť (Masarykovy univerzity, České akademie věd, Karlovy univerzity) v rámci stejnojmenného projektu financovaného Českou akademií věd v letech 2005-2009. Od roku 2010 je digitální knihovna v plném provozu. Článek popisuje obsah DML-CZ, její základní funkce a postupy používané při její tvorbě. Zkušenosti získané při budování DML-CZ a vyvinuté nástroje jsou využívány v některých dalších projektech na národní i mezinárodní úrovni, jako je například Digitální knihovna Filozofické fakulty MU nebo Evropská digitální matematická knihovna EuDML.

Klíčová slova: *digitální matematická knihovna; digitalizace; postup při tvorbě digitální knihovny; DML-CZ*

Abstract:

The DML-CZ – Czech Digital Mathematics Library has been implemented by Masaryk University Brno, Czech Academy of Sciences, and Charles University in Prague within the research and development project funded by the Academy of Sciences of the Czech Republic in 2005-2009. Since 2010 the digital library switched over to routine operation. This article presents content of DML-CZ, its implemented features and workflow used in its development. Know-how and tools developed in the DML-CZ are applied in other projects at both national and European levels, examples being the Digital Library of Faculty of Arts at Masaryk University and the European Digital Mathematics Library EuDML.

Keywords: *digital mathematics library; digitization; digital library workflow; DML-CZ*

Úvod

Česká digitální matematická knihovna DML-CZ (http://dml.cz) je v současnosti patrně nejrozsáhlejší a nejpropracovanější českou oborovou plnotextovou digitální knihovnou vůbec. Zahrnuje většinu nejdůležitějších vědeckých a odborných matematických textů publikovaných na území České republiky od druhé poloviny 19. století do současnosti. Většina těchto textů je volně dostupná komukoliv v režimu

otevřeného přístupu[1]. Knihovna DML-CZ vznikla jako výsledek pětiletého projektu programu „Informační společnost“ Akademie věd ČR, řešeného v letech 2005-2009 (projekt 1ET200190513, http://projekt.dml.cz) a je průběžně dále rozvíjena. Iniciátorem a oborovým garantem projektu byl Matematický ústav Akademie věd ČR; spolu s ním se na řešení projektu podílela čtyři další akademická pracoviště: Ústav výpočetní techniky Masarykovy univerzity (implementace a provoz vlastní digitální knihovny), Fakulta informatiky MU (výzkum v oblasti OCR matematických textů, digitální redakční workflow), Matematicko-fyzikální fakulta Univerzity Karlovy v Praze (tvorba metadat) a Knihovna Akademie věd ČR (digitalizace textů). Projekt měl významnou výzkumnou složku, v jejímž rámci se řešila specifika digitalizace, zpracování a zpřístupnění odborných matematických textů – v návaznosti na zkušenosti a poznatky obdobných zahraničních projektů. Hlavním výstupem projektu je však samotná digitální knihovna, kterou si nyní blíže představíme.

Obsah DML-CZ

Digitální knihovna DML-CZ obsahovala v říjnu 2011 přes 30.000 vědeckých a odborných článků od více než 10.000 autorů (tyto články představují v souhrnu zhruba 320.000 stran textu). Jádrem DML-CZ jsou matematické časopisy, dále jsou zařazeny i sborníky některých matematických konferencí a vybrané monografie. V časopisecké části je zastoupeno 10 nejvýznamnějších českých matematických časopisů a jeden slovenský časopis. Patří mezi ně časopis *Archivum Mathematicum* vydávaný Přírodovědeckou fakultou Masarykovy univerzity, stejně jako například *Časopis pro pěstování matematiky a fyziky* – první matematický časopis v zemích tehdejšího Rakousko-Uherska vydávaný od roku 1872 Jednotou československých matematiků a fyziků. 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 například práce Bernarda Bolzana – historicky patrně nejvýznamnějšího matematika působícího na našem území, ale i několik vybraných knih předních novodobých českých matematiků. Zařazena je také kolekce cca 40 monografií *Dějiny matematiky* mapující vývoj matematiky od starověku až do současnosti a poskytující například středoškolským učitelům matematiky vhodný materiál ke zpestření výuky historickými komentáři a souvislostmi. Nově bude v nejbližší době obohacen obsah DML-CZ o digitální archivy prací předních českých matematiků. Dokončen je první

1 Volně dostupných je více než 95 % článků v DML-CZ. Některé časopisy mají nastaveno časové embargo v délce 12 nebo 24 měsíců; články z nově publikovaných časopiseckých čísel mají volně k dispozici pouze metadata, přístup k plným textům je uvolněn až po uplynutí časového embarga.

z nich – kompletní digitální archív významného brněnského matematika Otakara Borůvky (zahrnující nejen všechny vědecké, odborné a popularizační práce Borůvkovy, ale také práce o Borůvkovi; celkem jde o 210 děl v rozsahu 4.000 stran).

Základní informační jednotkou poskytovanou digitální knihovnou je článek (resp. knižní kapitola). Jak je uvedeno výše, těch je v současnosti přes 30.000. Kromě plných textů článků nabízí knihovna také 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 matematických 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.

Během řešení projektu DML-CZ byla vyvinuta řada nástrojů a systémů, které podporují vytváření obsahu digitální knihovny a jeho zpřístupnění uživatelům. Základním nástrojem je Metadatový editor, rozsáhlá webová aplikace která integruje veškeré činnosti související s tvorbou a zpracováním obsahu – od zpracování naskenovaných obrázků jednotlivých stran, přes seskupování stránek do článků a vytváření jejich metadatového popisu, zpracování seznamů referencí, začleňování dokumentů vzniklých v digitální podobě až po generování výsledných článkových PDF souborů. Pro prezentaci obsahu koncovým uživatelům slouží upravený repozitářový systém DSpace, do něhož jsou importována data zkompletovaná Metadatovým editorem. Mezi další významné nástroje podporující pokročilé funkce DML-CZ patří například systém pro OCR matematických textů s rozpoznáváním matematických výrazů nebo automatizovaná textová analýza pro vyhledávání podobných článků s využitím metod strojového učení. Vedle nástrojů pro tvůrce digitální knihovny jsou důležitým prvkem celého systému také nástroje pro poskytovatele obsahu, především pak pro redakce matematických časopisů. Tyto nástroje umožňují sjednotit redakční postupy a standardy takovým způsobem, aby při tvorbě nového tištěného časopiseckého čísla vznikla i jeho standardizovaná digitální verze připravená pro okamžité začlenění do DML-CZ (a to nejen co se týká plných textů, ale také všech potřebných metadat, strukturovaných soupisů referencí a vazeb na další komponenty DML-CZ). Díky tomu je poměrně snadné digitální knihovnu průběžně doplňovat a udržovat ji aktuální.

Jak vzniká DML-CZ

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, fulltextové prohledávání, vzájemná provázanost dokumentů a jejich propojení s bibliografickými databázemi – a samozřejmě vzdálená nepřetržitá dostupnost[2].

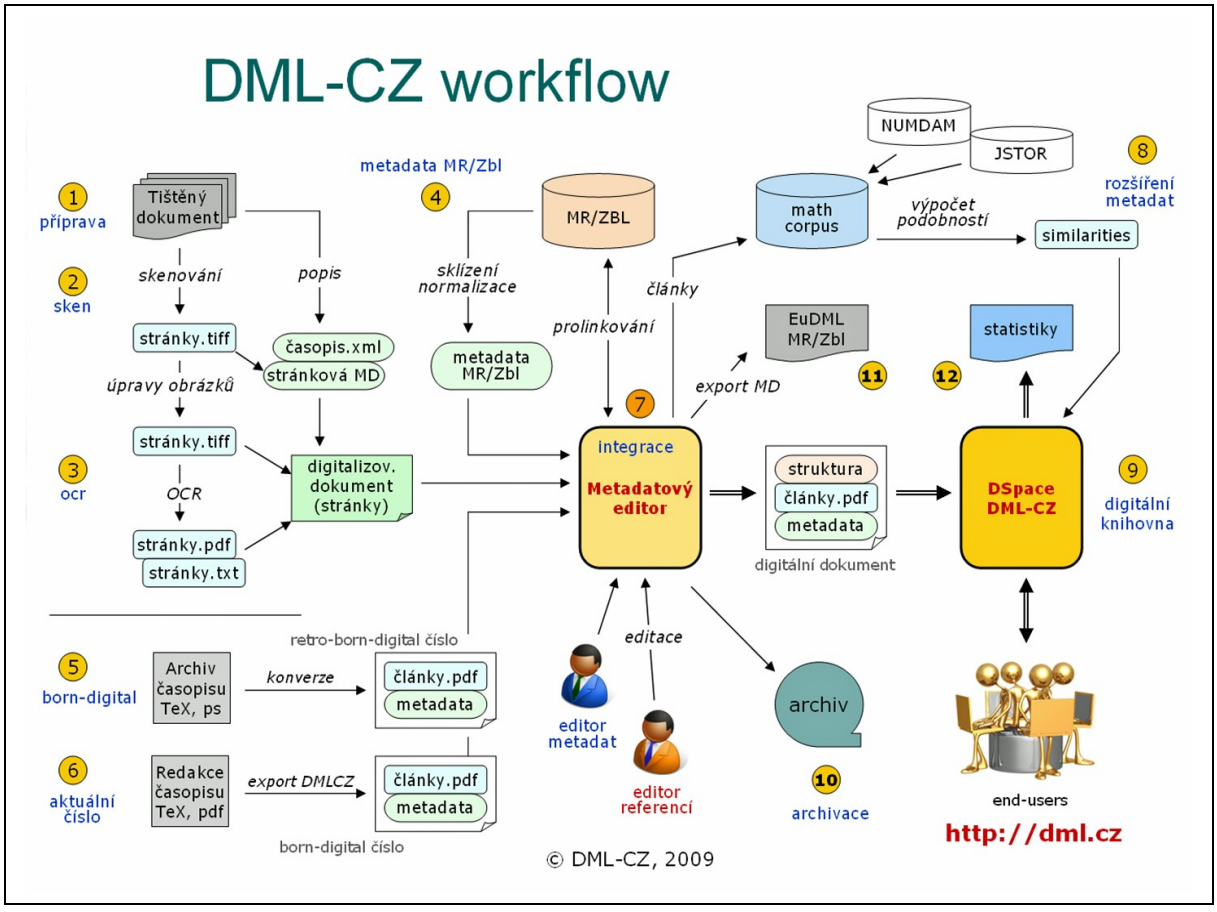
K vytvoření takové digitální knihovny je zapotřebí poměrně složitý postup zpracovávající různé druhy dokumentů a různé formy vstupních dat. Materiály zařazované do DML-CZ pochází ze tří publikačních období a jsou zpracovávány třemi rozdílnými způsoby:

- *tištěné dokumenty*: jde o časopisy, monografie a sborníky vydané většinou před rokem 1990, které existují pouze v tištěné podobě. Tyto dokumenty jsou skenovány, obrazy stránek jsou digitálně zpracovány (prahováním, odstraňováním šumu, narovnáním, rozpoznáváním textu), následně jsou stránky seskupovány do článků a jsou generována dvouvrstvá článková PDF obsahující obrazovou i textovou vrstvu, pro jednotlivé články jsou vytvářena popisná metadata a jsou zpracovávány reference. Hovoříme zde o tzv. retro-sken zpracování.
- *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 značkovacím a typografickém 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, a to i v rámci jednoho časopiseckého titulu. Takže je nezbytné konvertovat je do požadovaného jednotného tvaru. Metadata článků lze obvykle extrahovat z digitálních podkladů, je však třeba zohledňovat přitom specifika použité sazby. Tento typ zpracování označujeme termínem retro-born-digital.
- *nové digitální dokumenty*: v průběhu řešení projektu byl pro jednotlivé redakce vydávající 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 (born-digital zpracování).

2 Jiří Rákosník. DML-CZ: Česká digitální matematická knihovna. *Sborník semináře „Matematika na vysokých školách“*. Herbertov u Vyššího Brodu, Česká republika, 31.8.-2.9.2009. http://project.dml.cz/docs/herbertov2009_rakosnik.pdf

Veškeré podklady pro DML-CZ (získané kterýmkoliv z výše uvedených způsobů) jsou soustředěny a zpracovány v Metadatovém editoru, který je integračním centrem všech aktivit při vytváření digitální knihovny. Na zpracování dat v Metadatovém editoru se podílí různí partneři (včetně spolupracujících studentů) s rozdílným stupněm oprávnění k jednotlivým činnostem. Po kompletaci celého digitálního dokumentu, zkontrolování správnosti a úplnosti všech jeho částí a po doplnění vazeb na jiné dokumenty jak v rámci DML-CZ tak i mezinárodním (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 následujícím obrázku:



Obr. 1 Schéma postupu při vytváření DML-CZ

Základní kroky při tvorbě DML-CZ:

- *Příprava dokumentů k digitalizaci*: shromáždění tištěných podkladů, kontrola úplnosti a kvality předloh, vytvoření kontrolního soupisu a struktury dokumentu, ošetření autorských a vydavatelských práv (smlouva s vydavatelem).
- *Digitalizace – skenování*: skenování tištěných předloh v digitalizačním centru Akademie věd ČR v Jenštejně na knižním skeneru Zeutschel OS 7000. Předlohy jsou skenovány v rozlišení 600 dpi v režimu greyscale (4 bitová hloubka) a archivovány jako master-verze v nekomprimovaném formátu TIFF. Následně jsou digitální skeny upravovány programem BookRestorer (vyrovnání řádek textu, odstranění šumu, úprava kontrastu, ořez, různé geometrické transformace, binarizace). Současně jsou vytvořena základní metadata ve formátu XML (identifikace dokumentu, jeho struktura, čísla stran).
- *Rozpoznání textu – OCR*: upravené skeny z předchozího kroku jsou zpracovány programy na rozpoznávání textu (OCR – Optical Character Recognition), které umožní získat z bitmapových obrázků digitální text. Pro zlepšení přesnosti a pro rozpoznání matematických formulí probíhá tento proces ve dvou krocích: nejprve jsou skeny zpracovány univerzálním OCR programem ABBY FineReader; jeho výstup je následně předán k rozpoznání matematiky specializovanému matematickému programu INFTY (vyvinutému na japonské univerzitě v Kyushu). Výsledkem tohoto procesu jsou dvouvrstvé PDF-soubory jednotlivých stran dokumentu obsahující jak bitmapový obraz stránky, tak její rozpoznaný text.
- *Sklízení metadat z matematických databází*: pro zpracovávané články jsou automatizovaně získána metadata ze dvou světových matematických recenzních databází – americké MathSciNet (MR – Mathematical Reviews) a evropské Zentralblatt MATH (ZBL). Tato metadata slouží v dalších krocích zpracování jednak jako základ popisných článkových metadat, jednak k propojení článků DML-CZ na jejich recenze v databázích MR/ZBL.
- *Retro-born-digital zpracování:* starší digitální předlohy článků získané z digitálních archivů vydavatelů jsou obvykle v rozdílných formátech a verzích (někdy jsou k dispozici zdrojové soubory ve formátu TeX, které je možné dle potřeby upravit a vygenerovat výsledné digitální soubory, jindy jsou k mání pouze výsledné soubory ve formátu PDF či postcsript). Před začleněním do DML-CZ je třeba tyto digitální předlohy upravit, zkontrolovat shodu vůči tištěným originálům a převést je do jednotného formátu. K tomu je využívána sada volně dostupných a/nebo vlastních vytvořených programových nástrojů.

- *Nová born-digital čísla časopisů:* Redakční postupy časopisů zařazených v DML-CZ byly upraveny a doplněny novými automatizovanými postupy tak, aby při vytvoření nového časopiseckého čísla vznikla vedle podkladů pro tiskárnu současně i kompletní digitální verze připravená pro import do DML-CZ. Nová časopisecká čísla (metadata i plné texty článků) jsou tak zařazována do DML-CZ v postatě automatizovaně – s minimální pracností a minimálním časovým zpožděním.
- *Metadatový editor:* centrem všech aktivit při tvorbě DML-CZ je Metadatový editor – rozsáhlý webový systém vytvořený speciálně pro potřeby DML-CZ. Zde se soustřeďují výstupy všech předchozích kroků zpracování a zde také editoři integrují a vytváří finální obsah DML-CZ. V případě digitalizovaných dokumentů jsou seskupovány stránky do článků; u článků jsou finalizována popisná, strukturální a administrativní metadata; jsou vytvářeny seznamy referencí; jsou identifikovány a propojeny související články; je provedena série kontrol na úplnost a konzistenci všech data a metadat v DML-CZ.
- *Výpočet podobných článků:* data soustředěná v Metadatovém editorů jsou předávána ke specializovanému zpracování některým dalším subsystémům. Jedním z nich je systém pro strojovou analýzu textu a výpočet podobností mezi články. Na základě tří různých metod (LSI, RP, TFIDF) jsou identifikovány nejvíce podobné články, které jsou nabídnuty jako rozšiřující materiál uživatelům ve výsledné digitální knihovně.
- *Zpřístupnění uživatelům:* Úplná a zkontrolovaná data připravená v Metadatovém editoru jsou importována do uživatelské digitální knihovny DSpace, kde jsou zpřístupněna koncovým uživatelům. Open source systém DSpace byl upraven pro potřeby DML-CZ tak, aby nabízel snadnou a přehlednou navigaci, vyhledávání a prezentaci výsledků.
- *Návaznosti systému na okolí*: Metadatový editor a uživatelská digitální knihovna DSpace/DML-CZ poskytují řadu dalších služeb a funkcí pro komunikaci s okolím. Mezi ty nejdůležitější patří automatizované vytváření archivů a záložních kopií, generování statistik o obsahu systému a jeho využívání, poskytování metadat spolupracujícím systémům/projektům prostřednictvím technologie OAI-PMH a další.

Provoz a další rozvoj DML-CZ

Po ukončení projektu koncem roku 2009 byla digitální knihovna DML-CZ převedena do rutinního provozu, který zajišťuje Ústav výpočetní techniky Masarykovy univerzity (vlastníkem DML-CZ je Matematický ústav AV ČR). Denně ji navštíví kolem 500 uživatelů z celého světa. Knihovna se průběžně dále rozvíjí a doplňuje. Jsou

zařazována nově vyšlá čísla časopisů, pracuje se na digitalizaci nových časopiseckých titulů, začleňovány jsou nové typy dokumentů a sbírek (viz např. výše zmiňovaný digitální archiv Otakara Borůvky).

Zkušenosti a nástroje získané při práci na DML-CZ využívají řešitelé i v dalších projektech – v současnosti jsou to především projekty FFdigi a EuDML. FFdigi je projekt digitální knihovny Filozofické fakulty MU; jeho cílem je postupně digitalizovat a zpřístupnit formou digitální knihovny většinu publikací vydaných Filozofickou fakultou MU od jejího založení až do současnosti. Skenování tištěných dokumentů probíhá nízkonákladovým způsobem na jednoduchých stolních skenerech, ke zpracování a vystavení dat je využíván upravený Metadatový editor a systém DSpace. EuDML http://eudml.eu je tříletý evropský projekt (2010-2012) usilující o vytvoření evropské digitální matematické knihovny integrující obsah z národních matematických knihoven a od vydavatelů matematické literatury. Na projektu se podílí 14 partnerů z 9 evropských zemí, mezi nimi i řešitelé DML-CZ z Masarykovy univerzity a Matematického ústavu AV ČR. Digitální knihovna DML-CZ přispěje jako poskytovatel obsahu. První prototyp evropské digitální knihovny by měl být zpřístupněn v závěru roku 2011.

Projekt EuDML spolu s národními digitálními matematickými knihovnami představují významný milník na cestě k naplnění vize *světové digitální matematické knihovny* poskytující všem snadný a bezbariérový přístup ke kvalitní matematické literatuře. Současně ale mohou posloužit i jako příklad a inspirace pro další vědní obory.

Literatura

1. Miroslav Bartošek. Česká digitální matematická knihovna. *Zpravodaj ÚVT MU*. ISSN 1212-0901, roč. 20, č. 3, únor 2010, s. 11-13. http://www.ics.muni.cz/zpravodaj/articles/636.html
2. Jiří Rákosník. DML-CZ: Česká digitální matematická knihovna. *Sborník semináře „Matematika na vysokých školách“*. Herbertov u Vyššího Brodu, Česká republika, 31.8.-2.9.2009. http://project.dml.cz/docs/herbertov2009_rakosnik.pdf