DML-CZ – SOUČASNOST A BUDOUCNOST

Oldřich Ulrych a Jiří Veselý

Univerzita Karlova, Matematicko-fyzikální fakulta,
Sokolovská 83, 186 75 Praha 8 – Karlín,
emaily: ulrych@karlin.mff.cuni.cz, jvesely@karlin.mff.cuni.cz

Na konci tohoto roku skončí pětiletý projekt digitalizace české matematické literatury DML-CZ (viz základní informace v [10]). Představuje další vývojovou fázi v oblasti archivace, prezentace a okamžité dostupnosti publikací z oboru matematika. Je všeobecně známo, že se vývoj v této oblasti po staletí stále zrychluje. K ilustraci si stačí připomenout několik vybraných letopočtů: Jedna z nejrozšířenějších matematických knih, Eukleidovy *Elementa* (*Stoicheia*) ze 4. stol. př. n. l., která ovlivňuje vývoj matematiky dodnes, vyšla tiskem poprvé r. 1506. Vůbec první tištěnou matematickou knihou byla patrně *Treviso Arithmetic* z r. 1478 a prvním odborným časopisem, který zveřejňoval i matematické práce, byl *Journal des sçavans*, založený r. 1665. Téhož roku začaly vycházet i *Philosophical Transactions of the Royal Society of London*. Patrně nejstarší dosud existující čistě matematický časopis *Journal für die reine und angewandte Mathematik* (*Crelles Journal*) začal vycházet r. 1826.

Pozoruhodná je vzrůstající rychlost, se kterou matematické poznatky přibývají. Již kolem r. 1850 vycházelo přibližně 1000 matematických vědeckých článků ročně a o 100 let později jich bylo ročně již zhruba šestkrát víc. Přitom matematické poznatky zastarávají daleko pomaleji než poznatky jiných věd: mezi dnes citovanými pracemi je asi polovina starých alespoň deset let a čtvrtina více než dvacet let. Jednou z cest potřebných ke zvládnutí takové informační exploze, která postihuje všechny vědecké oblasti, je digitalizace. Právě v matematice je situace o to naléhavější, že dosažené výsledky zpravidla nebývají nahrazovány novými, spíše by se dalo říci, že se vrší.

Vývoj matematiky v některém směru se někdy zastaví nebo uvázne v mrtvém bodě, aby po delší době opět ožil a nabyl na intenzitě (např. některé výpočetní algoritmy byly objeveny mnoho let před jejich praktickým využitím). Přitom vědecká práce v matematice vyžaduje snadný přístup k těmto výsledkům, přehlednou orientaci i sledování jejich vzájemné provázanosti. Databáze *MathSciNet*, která vznikla z referativního časopisu *Mathematical Reviews* (MR), obsahuje recenze z více než 1900 časopisů a v charakteristice databáze *ZMATH* se uvádí, že přináší recenze

Podporováno projektem 1ET200190513 *DML-CZ: Česká digitální matematická knihovna* podporovaného v rámci programu „Informační společnost“ Akademie věd ČR (Národní program výzkumu, 2005-2009).

z asi 3500 časopisů. Ta vznikla z referativního časopisu *Zentralblatt für Mathematik* (Zbl) a obsahuje i referáty, publikované v dříve existujícím periodiku *Jahrbuch über die Fortschritte der Mathematik*. Přitom v obou databázích jde jen o výběr článků důležitých k odborné práci v matematice. V této souvislosti poznamenejme, že nejstarší český matematický (a fyzikální) *Časopis pro pěstování mathematiky a fysiky* začal vycházet r. 1872 a v dnešní době jen těch výlučně matematických u nás vychází osm.

Příchod praktického uplatnění digitalizace a nové technologie komunikace, umožňující širokou dostupnost Internetu, daly vědecké práci v matematice novou dimenzi. Rychlejší přístup k výše zmíněným elektronickým databázím umožňuje snadnou orientaci ve výsledcích. Podstatné přitom je i to, že obě již delší dobu využívají stejnou oborovou klasifikaci (MSC), která vznikla v USA r. 1970 a která reaguje i na vnitřní vývoj matematiky. Stejně tak se stalo samozřejmostí, že obě databáze jsou neustále doplňovány o nová data a obohacovány přístupem k plným textům referovaných prací.

Na celém světě vznikají elektronická úložiště, budovaná jak na komerčním, tak na nekomerčním základě a je třeba říci, že i zde „královně věd“ náleží ve srovnání s ostatními vědami průkopnická role. Srovnáme-li křivolakou cestu Eukleidových *Základů* ke čtenářům s přístupností práce v takovém úložišti z kteréhokoli místa na zeměkouli vybaveného připojením k Internetu, uvědomíme si dimenzi propastného rozdílu. Poznamenejme, že jedno takové úložiště (zpravidla jich bývá několik kvůli omezení nebezpečí ztráty dat) umožní stovkám knihoven lepší archivaci, což je další nemalý a bezprostřední ekonomický přínos digitalizační strategie.

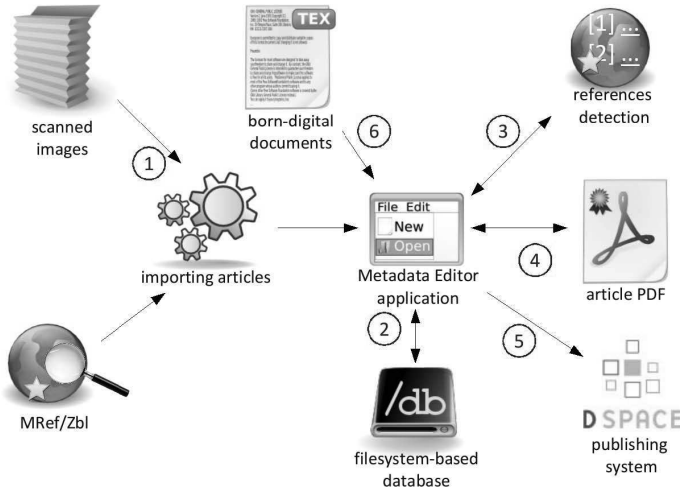
V poměrně nedávné době začaly databáze a také některé elektronické časopisy poskytovat aktivní reference. Náhled na citovaný odkaz je čtenáři k dispozici na jedno či více kliknutí – to v případě, že je nutno použít ke zprostředkování vazby jednu ze zmíněných databází. Elektronické verze časopisů jsou vytvářeny často i vydavatelem, který přístup k nim poskytuje paralelně s tištěnými verzemi nebo i místo nich. V poslední době je vznik elektronických verzí, zahrnující zpravidla i období, kdy byly publikace vytvářeny horkou a nikoli digitální sazbou, umožňován národní iniciativou. To je případ i projektu DML-CZ. Základní informace, které nebudeme opakovat, lze nalézt na webové stránce projektu [3] a v článcích [1], [8].

Co český digitalizační projekt DML-CZ přinese matematické komunitě? Všimněme si nejprve nejdůležitějšího výstupu projektu, tedy toho, co uživatel z české časopisecké a knižní produkce v oblasti matematiky na síti nalezne. Bude tam např. celý *Časopis pro pěstování mathematiky a fysiky* a časopisy navazující: *Časopis pro pěstování matematiky*, *Czechoslovak Mathematical Journal* a *Mathematica Bohemica*. A také časopisy *Acta Universitatis Palackianae Olomucensis. Mathematica*, *Acta Mathematica et Informatica Universitatis Ostraviensis*, *Aplikace matematiky* (dnes vydávaný pod názvem *Applications of Mathematics*), *Archivum Mathematicum*, *Commentationes Mathematicae Universitatis Carolinae*, a *Kybernetika*. Považovali jsme za vhodné k nim přiřadit i slovenský časopis *Mathematica Slovaca*. Nebudou chybět ani všechny sborníky velkých konferencí a zimních škol, které u nás byly organizovány, např. *Equadiff*, *Toposym*, *Winter School in Abstract Analysis* nebo *Winter School Geometry and Physics*, pokud k jejich vystavení budeme mít potřebná práva. A také některé vybrané knihy, např. stará vydání matematických prací Bernarda Bolzana apod. Později by měly přibýt i odborné pedagogické a popularizační časopisy z oblasti matematiky, které u nás vycházejí.

Pokud zmiňujeme práva k vystavení, bude u každého časopisu individuální tzv. *moving wall*. Jejich vydavatelé určí, od které doby mohou být na síti plné texty článků volně přístupné. I u sborníků konferencí je zpřístupnění plných textů závislé na vydavateli, zejména proto, že ne vždy byly vydány u nás. V této oblasti existuje spousta delikátních problémů, jejichž podrobnější popis přesahuje rozsah tohoto textu.

Nebylo by technicky příliš obtížné naskenované časopisy českého původu pouze vystavit. To ostatně již před řadou let pro některé z nich zajistilo digitalizační centrum v Göttingen; viz [5]. Prohledávání takto jednoduše vystavených materiálů bylo však prakticky nemožné, a tak mělo toto úsilí převážně archivační či zpřístupňovací roli. České vydávající instituce uvítaly tuto iniciativu a digitalizace se prováděla v souladu s jejich přáním. Některé podklady z digitalizace v Göttingen jsme měli možnost částečně využít, pokud byly naskenovány v dostatečné kvalitě. Digitalizační proces a jeho technika se s lety rychle vyvíjely a právě návštěvy a konsultace v göttingenském digitalizačním centru byly pro nás velice užitečným vodítkem pro stanovení základního pracovního postupu. Poznamenejme na okraj, že technické předpoklady pro digitalizaci u nás byly vytvořeny prakticky až po r. 2002 v souvislosti s povodněmi, které zničily spoustu odborné literatury.

Digitalizace v oblasti matematiky úzce souvisí s myšlenkou vytvoření světové digitální matematické knihovny (WDML), kterou iniciovala Mezinárodní matematická unie. Dostatečně přesný projekt její realizace se vlastně teprve vytváří. Na druhé straně však bylo kde se inspirovat: Matematici již zřejmě znají úspěšný a francouzskou vládou podporovaný projekt NUMDAM (viz [7]). Na celoevropské úrovni se o podobný projekt dlouho usilovalo: uvážíme-li jeho rozsah, je zřejmé, že realizace potřebuje i silné (evropské) finanční zázemí. Je potěšitelné, že projekt EuDML navrhující propojení existujících digitálních matematických repozitářů byl letos přijat a bude se v příštích třech letech naplňovat. Do projektu jsme se zapojili a DML-CZ by se měla stát součástí budované Evropské digitální matematické knihovny.



Schematické znázornění postupů při tvorbě DML-CZ

Všimněme si nyní některých vybraných aspektů projektu DML-CZ. Tvorba digitální knihovny je složitý komplex činností, který je schematicky znázorněn na předcházejícím obrázku. Podstatná část řešení projektu spočívá ve vytváření a zdokonalování specializovaných počítačových programů a nástrojů, jejichž účelem je získat potřebná digitální data v požadované kvalitě a co možná omezit úmornou ruční práci, které se tak jako tak nelze zcela vyhnout. Základní kroky spočívají v pořízení digitálních verzí dokumentů (1 a 6 v obrázku), v získávání základních metadat (3), ve zpracování získaných dat v Metadatovém editoru a v ukládání všech dat ve strukturované databázi (2), v sestavení odpovídajících dat do souborů typu pdf (4) a ve zveřejnění výsledku ve zvoleném prezentačním systému (5).

Základní princip spočívá v tom, že obrázky stránek daného článku (nebo kapitoly) jsou prezentovány v jednom souboru typu pdf. Popis článku je tvořen jeho metadaty, která obsahují standardní bibliografické informace: titul článku, jeho autora či autory, jazyk, ve kterém je napsán, název časopisu, ročník, číslo, stránky atd. Jde o data, která jsou většinou obsažena v databázích spojených s časopisy MR a Zbl, pokud v nich ovšem byla recenze článku uveřejněna. Data však nejde mechanicky převzít, často se navzájem liší a v některých případech dokonce nesouhlasí se skutečností. Jména autorů bývají zkomolena, titul článku v recenzi je uveden v jazyce recenze apod. Není-li původním jazykem článku angličtina, mohou existovat i dvě různé verze překladu názvu. Proto se vše pečlivě kontroluje a přebírá se zachováním principu zabránit ztrátě jakýchkoli (správných) relevantních dat. Některé články mají v našich metadatech titul až ve třech jazykových mutacích, nikdy však nesmí chybět anglická verze pro efektivní vyhledávání. Metadata zároveň obsahují propojení s databázemi *MathSciNet* a *ZMATH*, pokud jsou v nich články zachyceny.

Již v tomto stadiu si čtenář patrně bude umět představit, že jde o složitý a jen obtížně algoritmizovatelný proces pořizování potřebných informací. Pro účely zachování úplné informace o článku a pro potřeby citací musí být původní název článku součástí metadat. I když se skeny zpracovávají účinnými a v rámci projektu upravenými a zdokonalenými OCR programy (*optical character recognition*), je třeba automaticky získané podklady pečlivě zkontrolovat a upravit. Neurčí-li OCR program správně jazyk textu, připomíná jeho rozeznání „rozsypaný čaj“. Z hlediska vyhledávání a zpřístupnění informace širokému spektru uživatelů má však např. český či ruský titul omezenou hodnotu. Nejen z tohoto důvodu je nutno tituly článků přeložit do angličtiny. Matematické symboly je nutné přepsat v TEXu. Zvážíme-li obtíže s transkripcí např. slovanských jmen, není překvapující, že vyhledání odkazů na databáze je také spojeno s nemalým množstvím „ruční práce“. Jde jen o letmý pohled na to, jak mnohostranná a časově náročná je práce s metadaty článků, ale snad to k ilustraci postačí.

Vlastní digitalizace na výkonných kvalitních skenerech probíhala v Digitalizačním centru Knihovny AV ČR a na zpracovávání se podílely týmy pracovníků z více zúčastněných pracovišť. Samotný proces skenování, na první pohled vcelku přímočarý, také vyžaduje značnou dávku ruční práce. To je však jen začátek. Skeny se automaticky čistí a rovnají pomocí programu *BookRestorer*, kontroluje se jejich kvalita, pořadí a úplnost a zpracovávají se pomocí OCR programů. Často je ovšem nutno obrázky stránek dodatečně čistit individuálně v grafických programech, zejména u starších výtisků, které jsou méně zachovalé nebo nedokonale vytištěné. Pak se vytváří článková struktura ročníku časopisu: stránky se seskupují

do článků, vakáty se vynechávají apod. Jestliže článek nezačíná na nové stránce, je nutno některé stránky „klonovat“, a tak tentýž obrázek stránky je součástí dvou článků. Takové množství práce bylo možno zvládnout mj. díky práci některých studentů MFF UK v Praze a MU v Brně.

Teprve v tomto stadiu se zpravidla začíná s propojením digitalizovaných článků s metadaty. Podle povahy časopisu se někdy podaří určitou část metadat „sklidit“ z databází. Ta se spojí s články a zbytek je nutno doplnit ručně. Zároveň se rozbíhá proces kontroly přesnosti a úplnosti všech získaných metadat, doplňují se spojení do databází a neanglické tituly se překládají. Články se také dělí do skupin podle typů: matematické, fyzikální, redakční, historické, politické apod. Každý *matematický* článek je opatřen pětimístnými kódy klasifikace oboru (MSC). Nerozlišujeme mezi hlavní a vedlejší klasifikací, protože uvádíme kódy MSC získané ze všech dostupných zdrojů, jak kódy přidělené článku jeho autory, tak i ty, které doplnili recenzenti obou databází; chybějící kódy je samozřejmě opět třeba doplnit v rámci tvorby knihovny. V rámci projektu byl vytvořen program (viz [9]), který analyzuje obsah článku a po srovnání s referenční bází asi 5000 článků se spolehlivě určeným kódem nabízí pravděpodobnou klasifikaci; v jednotlivých případech je třeba ji opět kontrolovat. Za zmínku stojí fakt, že i přidělení klasifikace recenzentem je velmi často subjektivní. Stává se třeba, že MSC kódy přidělené v každé z obou databází jsou vesměs různé a neshodují se ani s kódy přidělenými autorem. Zmíněný program také umožnil nabídnout uživatelům digitální knihovny zajímavou službu: u vybraného článku je pod odkazem *Similar articles* uveden seznam článků z DML-CZ, které s daným článkem vykazují určitý stupeň obsahové podobnosti.

Velmi efektivním nástrojem pro komplexní zpracovávání metadat je již zmíněný Metadatový editor, který vytvořili účastníci projektu z Masarykovy univerzity v Brně (viz [2]), a který umožňuje současnou práci několika operátorů dálkovým přístupem prostřednictvím sítě.

Obecně lze říci, že „mladší“ časopisy s kvalitnějším tiskem se zpracovávaly lépe, než ty starší. Jedním z podstatných faktorů při zpracování byl čas: zkrátíte-li zpracování jednoho *článku* časopisu či sborníku třeba jen o sekundu, celkově ušetřený čas pro DML-CZ představuje skoro jednodenní práci jednoho pracovníka na plný pracovní úvazek. Na druhé straně některé méně čitelné skeny *stránek* si vyžádaly i desetiminutovou péči, aby se vůbec daly číst.

Výsledky digitalizace provedené v rámci pilotní části projektu bylo možné si prohlédnout na Internetu. Ne všechny – na vybraných částech jsme museli testovat funkčnost budoucího zpřístupňování. Také jsme přihlíželi k připomínkám pracovníků redakcí časopisů a prováděli nezbytné úpravy prezentace. Od 1. 1. 2010 budou všechny digitalizované časopisy, sborníky a vybrané knihy přístupné na adrese

`http://dml.cz/`

již v celistvosti. Pro prezentaci na síti jsme se rozhodli použít program *DSpace*. Je velmi pružný a vyhovoval nejlépe nárokům, které jsme na něj kladli, i když si jeho přizpůsobení potřebám DML-CZ vyžádalo relativně dost práce. Viz [6].

Bylo by nežádoucí, aby s ukončením projektu byla data zakonzervována a po nějakém čase snad dokonce vypnuty i servery, na nichž je možné digitalizovaný materiál efektivně využívat. Bylo nutno vyřešit provoz v budoucnu, tedy dohodnout, kde budou servery umístěny, kdo se bude o ně dál starat a spoustu dalších věcí

technického charakteru v podstatě nezávislých na vystavovaném materiálu. Servery vyžadují běžnou údržbu po hardwarové i softwarové stránce, s čímž jsou spojeny i jisté náklady. Další provoz bude zajištěn ve spolupráci Matematického ústavu AV ČR, který bude provozovatelem DML-CZ po ukončení projektu, s pracovníky Ústavu výpočetní techniky MU v Brně. Tím bude zajištěno, že vzniklé materiály budou nadále dostupné. A protože jsme si vědomi toho, že přes veškerou péči bude i po obsahové stránce jistě co zlepšovat nebo opravovat, je u každého článku uvedena nabídka *Feedback*, umožňující uživateli zaslat správci knihovny podnět.

Druhý okruh problémů, které jsou spojeny s budoucností DML-CZ, je spojen s přirozenou nutností dalšího rozšiřování a aktualizace digitální knihovny. Je hezké, že můžeme vzdáleně přistupovat k podstatné části matematické literatury, která byla u nás dosud vydána, ale svět se vyvíjí a redakce všech časopisů postupně vydávají další ročníky časopisů, konají se další konference apod. Je tedy potřeba zajistit, aby se i tyto nové informace v digitální knihovně objevovaly, a to co nejdříve (při současném respektování pravidel stanovených jednotlivými redakcemi); srv. [4]. Aby to bylo možné, bylo nutno vypracovat a uvést do života nástroje a postupy, které vkládání nových ročníků časopisů či sborníků konferencí již připravovaných v digitální podobě, umožní. Přitom tento postup musí být na jedné straně (z pohledu redakce) jednoduchý, na druhé straně musí garantovat, že nová data budou konsistentní se současným obsahem digitální knihovny.

Proto byly ve spolupráci s technickými pracovníky jednotlivých redakcí navrženy a ověřeny úpravy pracovních postupů tak, aby jednotlivé redakce mohly během přípravy tisku současně vytvářet bez další dodatečné práce potřebná data pro vystavení v DML-CZ. Ta se pak budou do jisté míry automaticky bez úprav, jen po nezbytných kontrolách celistvosti a kompatibility, vkládat do digitální knihovny přímo.

V redakcích časopisů znamenala příprava výstupů pro DML-CZ při zpracování čísla většinou dobře zvládnutelné úpravy. Redakce totiž generují již řadu výstupů podobného charakteru pro jiné účely, např. pro své webové stránky, pro referativní databáze, apod. Redakční pracovníci si mohou technickými prostředky ověřit již během přípravy čísla, že data, která se generují při zpracování článků a čísla časopisu, jsou ve správném tvaru a zda jejich vložení do digitální knihovny proběhne bez problémů. Zkušenosti však ukázaly, že úplné automatizování tak složitého procesu, závisejícího na mnoha neustále se měnících parametrech, je prakticky nemožné, takže některé věci bude nutno řešit i v budoucnosti. Musí tedy být k dispozici někdo, kdo celý systém zná hlouběji než jenom uživatelsky a může být nápomocný při vyřešení problémů.

Složitějším problémem bude začlenění dalších časopisů, sborníků a monografií do DML-CZ. Jde o operaci náročnou na práci lidí, kteří mají hlubší znalosti struktur a funkcí digitální knihovny (zvláště, pokud by se část dat musela získávat skenováním tištěných podkladů). Pro čtenáře *PMFA* bude zajímavá informace, že se v budoucnu plánuje začlenění tohoto časopisu do DML-CZ s vhodnou moving wall tak, jak tomu bude i u jiných časopisů určených širší odborné komunitě pracovníků a učitelům a žákům všeobecně vzdělávacích škol.

Dalším okruhem problémů, které přesahují současný rámec DML-CZ, je začlenění české digitální knihovny do vyšších celků a její provázanost s jinými subjekty: k těmto otázkám patří zlepšování interakce s referativními databázemi, vznik ev-

ropské či celosvětové matematické digitální knihovny a další. Prvním krokem bude zmíněný projekt EuDML na evropské úrovni, na kterém se budou někteří účastníci projektu DML-CZ podílet spolu s více než deseti zahraničními institucemi a firmami. Evropská matematická společnost iniciuje přípravu rozsáhlého projektu velké výzkumné infrastruktury, který by měl být také financován Evropskou komisí a který bude zaměřený na další rozvoj EuDML spolu s databází *ZMATH* a dalšími nástroji vědecké komunikace. V zahraničí měly výsledky práce na DML-CZ velmi kladnou odezvu a lze předpokládat, že získané zkušenosti naleznou i širší mezinárodní uplatnění. Podobný vývoj lze očekávat i v „nematematických“ oborech: v tomto směru byl projekt DML-CZ do jisté míry pilotní. Samotná oblast digitální prezentace se stále poměrně rychle vyvíjí a utváří, a tak i zde bude potřeba nemalé úsilí a množství práce při vylaďování sdílení dat s okolím.

Ze stručného popisu je snad dobře patrné, že projekt DML-CZ bude trvalým přínosem k celosvětovým trendům a že digitální knihovna, která v rámci něj byla vytvořena spolu s mnoha nástroji pro její budování a údržbu, bude poskytovat i v budoucnosti aktuální informace široké matematické veřejnosti.

Literatura

[1] Bartošek, M., *Česká digitální matematická knihovna*, INFORUM 2008: 14. konference o profesionálních informačních zdrojích Praha, 28.-30. 5. 2008, pp. 1–11.

[2] Bartošek, M., Kovář, P., Šárfy, M., *DML-CZ Metadata Editor Content Creation System for Digital Libraries*, Towards Digital Mathematics Library. Proc. of a workshop held in Birmingham, UK, July 27th, 2008, Masaryk University, Brno, 2008, pp. 139–151.

[3] *DML-CZ: Česká digitální matematická knihovna*, `http://projekt.dml.cz`.

[4] Ewing, J., *Twenty Centuries of Mathematics: Digitizing and Disseminating the Past Mathematical Literature*, Notices Amer. Math. Soc. **49** (2002), 771–777.

[5] *Göttinger Digitalisierungszentrum*, `http://gdz.sub.uni-goettingen.de`.

[6] Krejčíř, V., *Building the Czech Digital Mathematics Library upon DSpace System*, Towards Digital Mathematics Library. Proc. of a workshop held in Birmingham, UK, July 27th, 2008, Masaryk University, Brno, 2008, pp. 117–126.

[7] *Numérisation de documents anciens mathématiques*, `http://www.numdam.org/`.

[8] Sojka, P., Rákosník, J., *From Pixels and Minds to the Mathematical Knowledge in a Digital Library DML-CZ*, DML 2008 Towards Digital Mathematics Library. Proc. of the workshop held in Birmingham, UK, July 27th, 2008, Masaryk University, Brno, 2008, pp. 17–27.

[9] Sojka, P., Řehůřek, R., *Classification of Multilingual Mathematical Papers in DML-CZ*, Proc. of First Workshop of Recent Advances in Slavonic Natural Language Processing RASLAN 2007, Masaryk University, Brno, 2007, pp. 89–96.

[10] Ulrych, O., Veselý, J., *Digitalizační projekt DML-CZ*, PMFA **52** (2007), 260–261.