DML-CZ: Česká digitální matematická knihovna

Text projektu

1. Shrnutí současného stavu

Nezbytným předpokladem moderního výzkumu a vývoje v každém oboru je dobrý přístup k informacím o získaných poznatcích obsaženým v odborné literatuře. Vzhledem k prudce se zvětšujícímu množství těchto informací vzrůstá i potřeba informace efektivně skladovat, třídit a vyhledávat s využitím digitálních systémů a počítačových sítí.

To vše platí zejména v případě matematiky, která bývá nazývána královnou věd a bez níž se dnes téměř žádný výzkum a vývoj neobejde. Na rozdíl od řady jiných vědeckých disciplín si matematická literatura zachovává aktuálnost a užitečnost i po mnoho desítek let a některé takto „staré“ reference mohou poskytnout klíč k úspěšnému výzkumu. V matematice více než v kterékoli jiné disciplíně jsou nové poznatky založeny na odkazech na předchozí výsledky obsažené v existující literatuře s důrazem na vysokou přesnost a spolehlivost citací.

Po řadu let již proto různé skupiny matematiků po celém světě vyvíjejí aktivity směřující k vytvoření elektronických databází matematických poznatků, k digitalizaci relevantní matematické literatury a k vývoji efektivních vyhledávacích metod. Převážná část matematické literatury publikované ve 20. století je spolu s bibliografickými údaji a stručnými popisy obsahu zmapována v celosvětově rozšířených referativních časopisech Zentralblatt (navazuje na svého předchůdce Jahrbuch über die Fortschritte der Mathematik, který vycházel v letech 1868–1942) a Mathematical Reviews (vychází od r. 1940). Oba časopisy jsou nyní v převážné míře převedeny do podoby elektronických databází Zentralblatt MATH a MathSciNet, které by se měly stát důležitou součástí budovaného systému celosvětové digitální matematické knihovny. Obě databáze vykazují velmi vysokou míru využívání.

Evropská matematická společnost (EMS) na svém informačním serveru EMIS (http://www.emis.de) zřídila Elektronickou matematickou knihovnu ElibM zahrnující jistý segment časopisů a sborníků vydávaných v Evropě. V roce 2003 EMS připravila pro 6. rámcový program návrh široce pojatého projektu na vytvoření Evropské digitální matematické knihovny. Přes vynikající hodnocení návrhu nebyly pro jeho realizaci přiděleny prostředky a proto bylo dohodnuto postupovat dále „lokálně“ pomocí menších projektů na národní úrovni.

V roce 2003 se koordinace společných aktivit matematiků, vydavatelů odborné literatury, knihovníků a dalších odborníků v globálním měřítku ujala Komise pro elektronické informace a komunikaci Mezinárodní matematické unie (IMU) s cílem vytvořit celosvětovou digitální matematickou knihovnu (WDML), široce zpřístupnit vědecké a kulturní dědictví obsažené v matematických publikacích a dlouhodobě je zachovat. U nás tyto snahy sleduje Česká matematická společnost – sekce Jednoty českých matematiků a fyziků, která reprezentuje české matematiky v EMS a v IMU.

Dosud provedené kroky a získané poznatky ukazují, že tento obrovský (odhaduje se, že v celosvětovém měřítku půjde asi o 50 mil. stran textu) a komplexní cíl lze při koordinovaném úsilí v přiměřené lhůtě zvládnout. Výzkumné a realizační skupiny při

univerzitách v řadě zemí již pracují na vytváření digitální knihovny, řeší dílčí problémy, vyvíjejí metody a shromažďují poznatky, na které lze navázat a které lze přizpůsobovat specifickým požadavkům národního kulturního a jazykového prostředí. K nejvýznamnějším patří projekty NUMDAM (http://www.numdam.org) na univerzitě v Grenoble, DIEPER na univerzitě v Göttingen (http://gdz.sub.uni-goettingen.de/dieper), JSTOR (http://www.jstor.org) v USA či mezinárodní projekt EMANI (http://www.emani.org). Pravidelně jsou organizovány mezinárodní semináře a konference (např. v r. 2002 satelitní konference *Electronic information and communication in mathematics* při Světovém kongresu matematiků v Pekingu, v r. 2004 satelitní konference *New developments in electronic publishing in mathematics* při 4. evropském kongresu matematiků ve Stockholmu), jejich výsledky jsou publikovány ve sbornících.

2. Cíle a věcný obsah projektu

Cílem navrhovaného projektu je zkoumat, vyvinout a aplikovat postupy, metody a nástroje, které umožní vytvořit infrastrukturu a podmínky pro realizaci České matematické digitální knihovny (DML-CZ) zahrnující relevantní část odborné matematické literatury vydané v českých zemích a pro její začlenění do světové matematické digitální knihovny WDML. Součástí řešení je zahájení vlastního procesu digitalizace a zpřístupnění digitálního materiálu koncovým uživatelům. V návaznosti na to bude zahájen výzkum pokročilých technologií pro vyhledávání v matematických dokumentech a začleňování materiálů vzniklých již přímo v digitální podobě (born-digital).

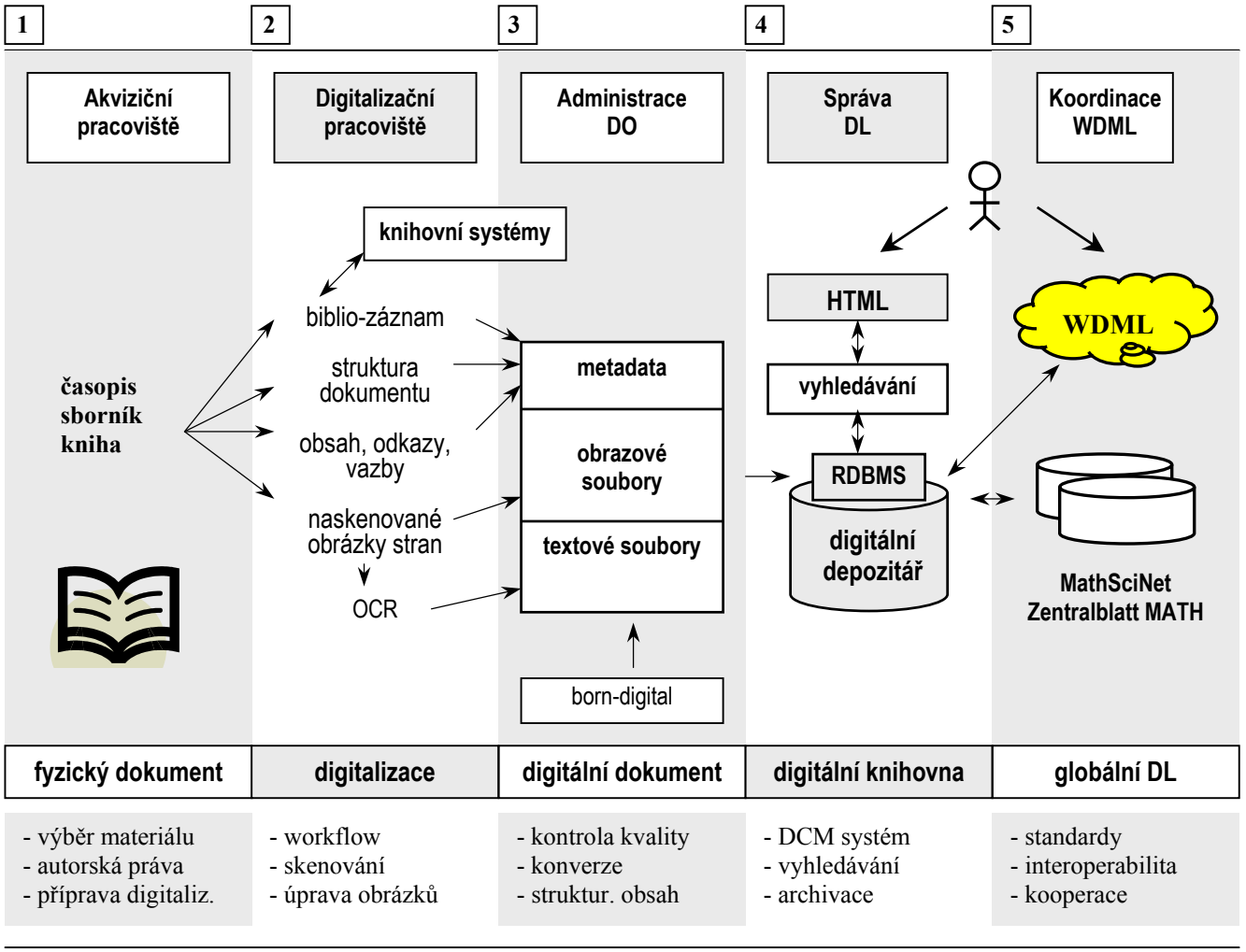
Do DML-CZ by měly být zařazeny v prvé řadě odborné časopisy mezinárodní úrovně vydávané českými institucemi, jako jsou *Czechoslovak Mathematical Journal* a *Applications of Mathematics* (vydává Matematický ústav AV ČR), *Archivum Mathematicum* (Masarykova univerzita) a řada dalších. V dalším půjde o sborníky konferencí vydané českými univerzitami a vědeckými ústavy. V poslední fázi mohou být digitalizovány vybrané monografie, učební texty, disertace, výzkumné zprávy apod. Podle předběžného odhadu by jádro DML-CZ mělo zahrnovat 200 až 300 tisíc stran textu.

Vytvoření digitální matematické knihovny je dlouhodobý a rozsáhlý cíl. Řešení navrhovaného projektu položí základy její české součásti a připraví předpoklady jejího úspěšného dobudování. Digitální matematická knihovna bude mít velký význam nejen pro rozvoj matematiky jako vědní disciplíny, ale i pro všechny ostatní uživatele výsledků matematického výzkumu včetně studentů a pedagogů. Obrovský objem informací související s hlubokým záběrem do historie, požadavek vysoké přesnosti a spolehlivosti a specifický způsob zacházení s matematickou literaturou představuje řadu originálních problémů v oblasti digitalizace, archivování, distribuce a prezentace informací a vzdáleného přístupu k nim. Získané poznatky však budou mít z velké části univerzální charakter a bude možné je využít k rozvoji a managementu informačních a znalostních systémů i v jiných oborech.

3. Předpokládaný metodický postup

Na základě výsledků přípravné fáze, v níž bude detailně studován aktuální stav světového poznání a trendy v dané oblasti (neobjevovat již objevené), budou zkoumána a implementována řešení zohledňující česká specifika a české prostředí na jedné straně, na druhé straně respektující světové standardy a trendy umožňující začlenění do WDML.

Základní schéma systému DML-CZ, jeho komponenty a návaznosti jsou znázorněny na následujícím obrázku:



Věcný obsah projektu zahrnuje návrh a realizaci řešení vzájemně provázaných problémových okruhů v následujících pěti oblastech (v závorce předpokládaný hlavní řešitel pro danou oblast):

1. akvizice (MÚ AV ČR ve spolupráci s MFF UK)

- zpracování právní expertízy z oblasti vlastnických a autorských práv (IPR – Intellectual Property Rights) ve vztahu k DML-CZ
- zřízení pracoviště pro získávání a přípravu materiálů k digitalizaci
- výběr konkrétních materiálů (časopisy, sborníky, monografie, …) k digitalizaci
- ošetření IPR pro vybrané materiály
- příprava materiálů k digitalizaci

2. digitalizace (KNAV ČR ve spolupráci s MÚ AV ČR a dalšími partnery projektu)

- stanovení technických parametrů digitalizace v souladu s připravovanými zásadami WDML-BPS (Best Practice Statement)
- stanovení digitalizačního workflow (příprava, popis, digitalizace, kontrola kvality, úpravy obrázků, OCR, začlenění do systému, …)

- výběr, popř. adaptace a nasazení sw-systému pro komplexní podporu digitalizace
- ověření použitelnosti dostupných digitalizačních zařízení v Knihovně AV ČR a jejich případné doplnění
- digitalizace, skenování
- úprava digitálních obrázků
- pořizování metadat (u deskriptivních s využitím bibliografických záznamů v knihovním systému)
- OCR

3. digitální dokument (MFF UK ve spolupráci s ÚVT MU, FI MU a KNAV ČR)

- specifikace struktury digitálního objektu (DO)
- stanovení metadat (deskriptivní, strukturální, administrativní – včetně technických a IPR)
- persistentní globální identifikace digitálních objektů
- realizace složených DO pro různé typy dokumentů (seriál, článek, monografie)
- stanovení archivních a prezentačních formátů
- konverze mezi formáty a generování digitálních derivátů
- transformace born-digital (již vytvořených v elektronicky zpracovatelné podobě) materiálů na unifikované DO vyhovující navrženým a podporovaným DTD (definicím typů dokumentů)
- evaluace možností automatizace konverze vizuálně značkovaných OCR dat na logicky strukturované dokumenty
- strukturovaný textový obsah born-digital DO
- vzájemná provázanost DO, možnosti automatizace hypertextových odkazů
- propojení s databázemi

4. digitální knihovna (ÚVT MU a FI MU ve spolupráci s KNAV ČR a MFF UK)

- výběr, přizpůsobení a nasazení Content Management systému (CMS) pro digitální knihovnu DML-CZ (preferována budou volně dostupná sw-řešení typu DSpace)
- správa DO
- řízení přístupu k DO
- dlouhodobá archivace DO
- vyhledávání v digitální knihovně (metadata, plné texty) a prezentace výsledků uživatelům (obrazové soubory) – přístup uživatelů v prostředí www
- indexování archívu pro přesné dotazování

5. začlenění do WDML (MÚ AV ČR po koordinační stránce a ÚVT MU, FI MU po technické stránce)

- zajištění interoperability a začlenění DML-CZ v rámci WDML (např. na bázi automatizovaného sklízení metadat s využitím standardů OAI – Open Archive Initiative a protokolu OAI-PMH – Protocol for Metadata Harvesting)
- napojení digitálního obsahu DML-CZ na referativní databáze Zentralblatt MATH, MathSciNet

Řešení uvedených okruhů otázek bude vždy předcházet studium existujících metod a postupů. S ohledem na začlenění DML-CZ do WDML bude mít velký význam spolupráce a výměna poznatků se zahraničními skupinami pracujícími v oblasti digitalizace a zkoumání možností jejich využití v našem prostředí. Půjde především o skupiny

v Göttingen a v Grenoble. Důležitá bude také aktivní účast řešitelů na mezinárodních konferencích a seminářích věnovaných tématice WDML a digitalizace obecně. Bude nezbytné úzce spolupracovat s komisemi ustavenými IMU a EMS pro koordinaci WDML.

Výsledky výzkumu budou průběžně ověřovány v pilotní části projektu, v jejímž rámci bude digitalizován časopis *Czechoslovak Mathematical Journal* vydávaný od r. 1951 Matematickým ústavem AV ČR. Časopis se zejména v prvních třiceti letech vyznačoval vícejazyčností – články byly publikovány v angličtině, ruštině, němčině, francouzštině, italštině, přičemž i v rámci jednoho článku se mohlo vyskytovat více jazyků (abstrakt, seznam použité literatury). Tato složitost může mít zvláštní význam pro ověřování některých specifických postupů a výsledků (automatické rozpoznávání textu, vyváření metadat, značkování apod.). Po ověření v pilotní části projektu budou výsledky aplikovány na další digitalizovanou literaturu.

K ověřování může být využit i elektronický materiál, který byl v rámci projektu DIEPER na univerzitě v Göttingen vytvořen digitalizací *Časopisu pro pěstování matematiky* (nyní *Mathematica Bohemica*) vydávaného od r. 1872 Jednotou českých matematiků a fyziků (od r. 1952 MÚ AV ČR) a časopisu *Commentationes Mathematicae Universitatis Carolinae*, který vydává MFF UK.

4. Přibližný časový rozvrh řešení

První rok řešení projektu bude věnován zejména studiu aktuálního stavu a shromažďování poznatků. Budou navázány pracovní kontakty se zahraničními skupinami pracujícími na tvorbě WDML, především v Německu a ve Francii, a s příslušnými odbornými komisemi IMU a EMS. Bude provedena identifikace relevantní literatury a stanoveny priority pro její digitalizaci a zařazení do WDML. Bude vypracována právní analýza vztahu autorských práv a prezentace digitalizovaných textů. Bude navržen workflow pro pořizování metadat, digitalizaci a archivování prvotního digitálního materiálu a připraven pilotní projekt pro ověřování získaných výsledků. Bude pořízen server pro instalaci, archivování a provozování digitální knihovny.

Ve druhém roce bude ve středisku spolunavrhovatele 4 digitalizován materiál pro pilotní část projektu (cca 30 tisíc stran). Bude instalován server s prototypovým sw-řešením. Souběžně bude zahájeno shromažďování poznatků o formátech literatury již existující v nějaké digitální podobě (digital-born), jejich vzájemné porovnávání a studium možností navrhnout univerzální efektivní nástroje a postupy pro zařazení této literatury do vytvářené digitální knihovny.

V dalších letech bude řešení pokračovat výzkumem nástrojů inteligentního vyhledávání, propojování s databázemi Zentralblatt MATH a MathSciNet a vzájemného propojování dokumentů prostřednictvím elektronických odkazů, jednoznačné identifikace a strukturování dokumentů.

Protože problematika WDML se dynamicky rozvíjí, budou jednotlivé konkrétní kroky řešení průběžně přizpůsobovány aktuální situaci a stavu poznání. Výsledky ověřování na pilotním projektu budou pravidelně vyhodnocovány. Po úspěšném ověření budou jako jádro vznikající české digitální matematické knihovny v závěrečné fázi zapojeny do WDML. Budou připraveny organizační a právní podmínky k tomu, aby bylo výsledků projektu využito k zajištění provozu, rozšiřování, zpřístupnění a dalšího vývoje knihovny

v prostorách pracoviště navrhovatele ve spolupráci s Jednotou českých matematiků a fyziků.

5. Podmínky pro řešení projektu

Pracoviště navrhovatele i všech spolunavrhovatelů jsou dostatečně vybavena základní počítačovou technikou pro řešení teoretických i praktických úkolů a jejich ověřování a mají dobrý přístup do internetové sítě. Knihovna AV ČR disponuje nově zřízeným digitalizačním střediskem s kapacitou cca 40 000 zpracovaných stran měsíčně, osazeným potřebnou moderní a výkonnou technikou a programovým vybavením:

- 1 x barevný knižní skener Digibook RGB 10000 formát A1
- 2 x černobílý knižní skener Zeutschel OS 7000 formát A2
- Server Dell PowerEdge 2650 s diskovým polem Dell PoverVault o kapacitě 1 TB určený pro zpracování a archivaci.
- Software Book Restorer pro grafické úpravy.
- Systém Sirius pro zpracovaní a archivaci digitalizovaného materiálu.
- Systém Kramerius pro zpřístupnění digitalizovaného materiálu.

Pro instalaci a provoz digitální knihovny vytvořené v rámci pilotního projektu bude třeba pořídit dostatečně dimenzovaný server a zálohovací zařízení, která budou instalována na pracovišti navrhovatele.

Knihovní depozitáře pracovišť všech účastníků projektu obsahují veškerý materiál, který by měl být v rámci projektu digitalizován.

Řešitelský tým byl sestaven tak, aby odpovídal náplni a komplexnosti navrhovaného projektu.

Skupina navrhovatele z MÚ AV ČR bude koordinovat činnost všech spolunavrhovatelů a jejich kontakty se zahraničními týmy, součinnost s provozovateli databází Zentralblatt MATH a MathSciNet, bude zajišťovat výběr literatury pro digitalizaci a řešení otázek copyrightu. Vytvoří podmínky pro provoz a zpřístupnění vytvořené digitální knihovny a pro její další rozvoj. Uplatní přitom své zkušenosti z počítačové sazby a vydávání odborné matematické literatury (časopisy, sborníky konferencí, monografie) a z tvorby přehledových databází matematické literatury a z dosavadní mezinárodní spolupráce v rámci aktivit směřujících k WDML. Navrhovatel je vědeckým pracovníkem MÚ AV ČR, místopředsedou České matematické společnosti JČMF, vedoucím české redakční skupiny referativního časopisu Zentralblatt a databáze Zentralblatt MATH a účastnil se přípravy v úvodu zmíněného projektu EMS na vytvoření evropské digitální matematické knihovny. Mgr. H. Severová je vědeckou tajemnicí MÚ AV ČR a výkonnou redaktorkou časopisu *Czechoslovak Mathematical Journal*, která zajišťuje i jeho počítačovou sazbu.

Spolunavrhovatel 1 (M. Bartošek) je zkušeným odborníkem a pedagogem v oblasti informatiky se zaměřením na digitální knihovny, který se podílel na tvorbě několika systémů digitálních knihoven a řešení řady výzkumných projektů v dané oblasti, včetně projektů na evropské úrovni. Jeho skupina bude řešit především otázky realizace vlastní digitální knihovny pro správu a zpřístupnění digitálního obsahu, propojování objektů, vytváření digitálních depozitářů atd. Pracoviště spolunavrhovatele úzce spolupracuje

s Fakultou informatiky MU a díky tomu rovněž disponuje velkým potenciálem studentů, kteří mohou být v rámci svého studia zapojeni do řešení konkrétních dílčích úkolů.

Spolunavrhovatel 2 (P. Sojka) je zkušeným odborníkem a pedagogem v oblasti počítačové sazby, digitalizace, zpracování přirozeného jazyka, textových informačních systémů atd. Při řešení projektu se zaměří především na identifikaci a řešení problémů značkování textů a konverze born-digital dokumentů do jednotného vybraného formátu, výběr standardů metadat a použitých technologií, na problémy digitální sazby z hlediska použití v WDML Pracoviště spolunavrhovatele rovněž disponuje velkým potenciálem studentů, kteří mohou být v rámci svého studia zapojeni do řešení konkrétních dílčích úkolů.

Spolunavrhovatel 3 (O. Ulrych) je odborným a pedagogickým pracovníkem v oblasti matematiky, informatiky a informačních technologií, počítačových sítí a počítačové sazby matematických textů. V rámci přípravy návrhu projektu získal cenné poznatky o postupech digitalizace uplatňovaných na univerzitě v Göttingen a bude při řešení projektu m.j. důležitým spojovacím článkem mezi matematiky (autory publikací a uživateli WDML) a informatiky (tvůrci struktur, postupů a nástrojů WDML). Doc. J. Veselý je vědeckým a pedagogickým pracovníkem se zkušenostmi v oblasti knihovnictví a elektronické přípravy matematických textů. Bude spolupracovat při řízení projektu, výběru a přípravě digitalizované literatury a řešení koncepčních otázek digitální knihovny.

Spolunavrhovatel 4 (M. Lhoták) je vedoucím oddělení informačních technologií, zodpovídá za provoz digitalizačního střediska KNAV ČR. Je odborníkem v oblasti elektronických databází a knihovních systémů. Jeho úkolem bude pořízení digitálních obrazů příslušné literatury, jejich propojení s odpovídajícími metadaty a prvotní archivace. Bude se podílet na projektu zejména v oblasti vývoje a úprav stávajících systémů používaných na digitalizačním pracovišti KNAV ČR s ohledem na potřeby projektu. Ing. Martin Duda je odborník na OS Linux a aplikace s ním související. Dříve byl zaměstnán ve společnosti SuSE CR, s. r. o., zabývající se vývojem a distribucí OS SuSE Linux. V současné době se zabývá administrací systému Kramerius, který slouží ke zpřístupnění digitalizovaných materiálů v KNAV. V rámci projektu bude především provádět úpravy a testy v tomto systému.

V Praze, 20.8.2004