Projekt MEMORIA: rukopisy a staré tisky na Internetu

Ing. Stanislav Psohlavec, AiP Beroun s.r.o.
PhDr. Zdeněk Uhlíř, Národní knihovna ČR

Článek byl zveřejněn v časopise ITlib „Informačné technológie a knižnice“ 2/04
a zde je použit se svolením redakce http://www.cvtisr.sk/itlib

MEMORIA je název sdružující iniciativy, které se zrodily v souvislosti s provozem a řešením programu Memoriae Mundi Series Bohemica (MMSB). Projekt Memoria směřuje k vybudování virtuálního badatelského prostředí pro oblast historických knižních fondů. Projekt podporuje vznik nových a využívání existujících informací, zajišťuje jejich dlouhodobou životnost a trvalou použitelnost. Zpřístupňuje výsledky dosavadní digitalizace dokumentů, výsledky detailních popisů historických dokumentů a přináší přístup k bibliografickým informacím z různých zdrojů v databázi "Manuscriptorium". Perspektivně budou zařazovány plné texty primárních dokumentů, tj. bude dostupná edice originálních historických dokumentů, i sekundárních dokumentů, tj. dokumentů zpracovaných na jejich základě. Náhledy do digitalizovaných rukopisů jsou volně přístupné, kvalitní zobrazení je licencované, přičemž pro aktivní účastníky projektu je licence zdarma .

Cíle projektu

Vycházíme z toho, že nové, stejně jako staré informace o historických dokumentech vznikly v nejlepší víře a úmyslu. Doba i podmínky v nichž vznikly se podstatně liší, průběžně se mění názory, rozšiřuje se poznání na které se navazuje Z principu nelze žádná data vzít za definitivní, správná. Jsme přesvědčeni, že každá dostupná informace je pro badatele přínosem a vodítkem pro další samostatnou práci.
Praktickým cílem programu MEMORIA je vybudování virtuálního badatelského prostředí pro oblast historických fondů formou sdíleného **otevřeného katalogu těchto fondů** na který navazuje pořizování digitálních forem těchto fondů (obrazových, textových) a jejich zpřístupňováním a tedy připojení **digitální knihovny obrazových kopií a plných textů.** Nezbytností je zabezpečení propojení k dalším datům majícím vztah k těmto fondům (studie, další popisná metadata, další elektronické dokumenty/objekty) a také poskytování informací jiným systémům standardizovanými postupy (Z39.50, OAI)
Současným výsledkem snahy o dosažení těchto cílů je databáze Manuscriptorium, která je dostupná buď přes stránku představující celý projekt MEMORIA www.memoria.cz nebo přímo přes www.manuscriptorium.com .
Tento první krok je prezentací existujících informací a nelze ho proto zatím nazvat badatelským prostředím. Tím se stane až bude umožněno tyto zveřejněné informace řízeně měnit, doplňovat, provazovat.

Rutinní aktivity

Tyto aktivity zahrnují rutinní popisy a digitalizaci vzácných originálů dokumentů, výrobu digitálních dokumentů, archivaci a ochranu digitálních dat. Tyto převážně výrobní činnosti probíhají především v rámci projektů **VISK 6**, případně dalších projektů, které koordinátor a provozovatel projektu MEMORIA iniciovali nebo se na nich podílejí.

Popisy

Popisy digitalizovaných dokumentů musí být dosti podrobné, protože jsou určeny nejen ke katalogizaci, ale také k propojení digitálních obrazů do formy elektronického dokumentu, jehož typickou podobou je virtuální kniha.
Zde je zajímavá geneze prostředků, které se pro základní popisy dokumentů využívají. Zdánlivě logické řešení popisovat digitální obrazy se v praxi neosvědčilo z více důvodů, přičemž hlavním byla absence mnoha informací, které jsou obrazem nenahraditelné a které nese jen sám originální dokument.
Na druhou stranu cílem základního popisu dokumentů v tomto projektu není tradiční způsob katalogizace, nejde primárně o náhradní slovní reprezentaci originálu, která má v tištěném prostředí nesporné oprávnění.. Jestliže je v elektronickém prostředí dostupná reprezentace originálu v podobě obrazu a očekává se připojování dalších informací, pak se slovní prezentace (až interpretace) originálu v počátku stává postradatelnou, pokud není jejím účelem poskytnutí signifikantních informací vedoucích k nalezení dokumentu a informací v něm obsažených.

Popis předcházející digitalizaci má výhodu mimo jiné i v tom, že při popisu dokumentu se současně provádí kontrola, zda je rukopis bez rizik způsobilý pro digitalizaci. Zavedli jsme princip pevně strukturovaného popisu **DOBM** (Digitization of Old Books, Manuscripts, and Other Documents) využívajícího SGML, který je vytvářen před digitalizaci. Nástroje jsou volně k dispozici a jsou dosud používány pro svou jednoduchost, nepokládáme je však již za perspektivní.
Pevná strukturace popisu ve formě DOBM má sice výhodu v tom, že je jednoduchá a usnadňuje tudíž rutinní až industriální práci. Nevýhodou však je to, že využívá pouze tvrdě strukturovaných dat na jedné nebo vůbec nestrukturovaných dat na druhé straně: bibliografické údaje a údaje o některých snadno typizovatelných vnějších znacích jsou ve formě tvrdě strukturovaných dat, zatímco ostatní údaje včetně údajů o intelektuálním obsahu originálního dokumentu jsou v podobě volného textu. V případě podrobného (a tedy rozsáhlého) popisu se tak klade překážka jak prosté orientaci v zobrazeném záznamu, tak sofistikovanějšímu vyhledávání. Ukázalo se, že pevná struktura záznamu je pouhým mezistupněm k takové formě záznamu, která bude využívat také dat semistrukturovaných.
Vytvořením pokročilejší formy zápisu dat se zabýval evropský **projekt MASTER** (Manuscript Access through Standards for Electronic records), jehož řádným partnerem byla i Národní knihovna ČR. **Standard MASTER** (nejprve na bázi SGML, posléze XML), který byl výsledkem této aktivity, umožňuje vytváření a využívání zejména semistrukturovaných dat, tzn. je přizpůsobivější jak variabilitě popisovaného materiálu, tak orientaci při zobrazení a postupům při vyhledávání. Je založen na strukturaci obsahových elementů do hloubky i na relativně volném využívání funkčních elementů, které se mohou vztahovat k různým horizontálním i vertikálním místům ve struktuře celého popisu. Pevná jsou pouze pravidla syntaxe. Ve standardu MASTER lze tudíž pořizovat jak zcela jednoduché, informačně minimálně nasycené záznamy, tak záznamy jdoucí do hloubky popisovaného originálního dokumentu. To znamená, že jeho praktické využití je velice široké a flexibilní, adaptovatelné pro různé účely i různou míru znalostí o materiálu, aniž je to na překážku využití v informačním systému.

Zavedení popisů ve formě XML v rámci projektu MASTER vedlo k prvotnímu popisu dokumentů ve volné a badatelským potřebám přizpůsobenější struktuře MASTER a k následnému přepisu dat do pevné formy DOBM. Nyní se dokončují prostředky využívající jen XML. I tyto prostředky budou volně dostupné. Původně podstatnou většinu popisů dokumentů zajišťoval dosti rozsáhlý kolektiv spolupracujících odborníků. Významné množství dokumentů si nyní popisují jejich majitelé sami. K úspěšné spolupráci je nutné jen velmi krátké a jednoduché zaškolení, jaká je nutno dodržovat formální pravidla. Přesto mnoho partnerů využívá možnosti nechat popsat dokumenty současně s digitalizací, protože spolupracujeme s uznávanými odborníky garantujícími dlouhodobě včasnost a kvalitu popisů.

Digitalizace

Technologické vybavení má ve své historii několik „nej".
První digitální kamera KODAK 460 RGB byla první kamera tohoto typu v ČR a v tzv. východních zemích. Také první digitální kamera BetterLigth 6000 byla první v Česku, a to v době, kdy tato kamera byla ještě takřka neznámá. Byla vybrána na základě mimořádné kvality produkovaných obrazů a disponuje rozlišením až 48 milionů pixelů (neaproximované RGB).
Současnou špičkou je Special BookScanner 145 CRUSE. Jde o modifikací scannerů určených především pro snímání obrazů a map. Vznikl v přímé spolupráci firmy CRUSE a AiP Beroun, která se na jeho vývoji přímo podílela. Je ve všech ohledech optimalizován pro náročnou digitalizaci vzácných historických dokumentů. Ke vzniku tohoto zařízení přispělo paradoxně nerovnoměrné a nejisté financování projektů, které způsobilo, že bylo nezbytné rychle reagovat na neočekávané zvýšení požadavků na digitalizaci. Zařízení vzniklo během tří měsíců a bylo pořízeno firmou AiP Beroun na leasing.
Poslední zařízení založené na kameře BetterLigth, které nahradilo kameru KODAK, je již plně vyvinuto firmou AiP Beroun. Přináší zúročení dosavadních praktických zkušeností a výrazně snižuje investiční náklady oproti dřívějším nákupům univerzálních zařízení. Toto zařízení bude instalováno také v Univerzitní knihovně v Bratislavě.
Všechna zařízení jsou optimalizována s ohledem na bezpečnost dokumentů, snadnou manipulaci s nimi a jejich ochranu před UV a IR zářením. Tyto primární požadavky neovlivňují vysokou produktivitu a kvalitu na úrovni špičkové studiové práce.

Výroba digitálních dokumentů

Digitální obrazy jsou spojovány s dříve připravenými popisnými daty, do formy dokumentu, který jednak obsahuje všechny popisné a technické informace s využitím standardů MASTER (dříve DOBM) a navíc vygenerované HTML soubory svazující obrazy do formy dovolující prohlížení dokumentu běžně dostupnými internetovými prohlížeči.

Archivace a ochrana digitálních dat

Archivace dat prodělala zhruba tři období:
1. V počátcích, koncem devadesátých let nebyla jiná levná možnost archivace větších objemů dat než CD-R disky. Pracovníci AiP Beroun zajistili pro bezpečnou archivaci na CD-R technologii měření kvality záznamu, zajistili vyhodnocení stárnutí vytvořených medií (viz programy Věda a výzkum). Strategie archivování byla založena na existenci sice drahých, ale kvalitních a stabilních medií (KODAK Ultima Gold).
2. Boom využívání CD-R způsobil řádové zlevnění CD-R medií, a tím odstranil z trhu drahá media vhodná pro archivaci. Kvalita medií velmi poklesla, protože konkurenční boj si vynutil snižování nákladů na výrobu, a to i za cenu nižší kvality. Přitom měřicí technika pro kontrolu kvality zůstávala stále velmi drahá a její cena dále stoupala. To nedávalo dlouhodobou perspektivu pro využívání CD-R jako archivačního media.
3. Rok 2003 přinesl radikální novinku. Pokrok ve vývoji vypalovacích jednotek přinesl další generaci IC obvodů, dovolující monitorovat činnost signálového procesoru.To dává reálnou možnost s akceptovatelnými náklady zajistit a ověřit kvalitu produkovaných disků s vypálenými daty a monitorovat jejich stárnutí. V případě signifikantního poklesu jejich kvality lze informace přepsat včas na nová media. Kvalita medií se dalším vývojem opět stabilizovala, a navíc AiP Beroun spolupracuje s výrobcem medií. To dovoluje dále využívat CD-R jako archivační medium. Samozřejmě se připravuje paralelní uložení všech těchto vzácných dat na vznikajících hromadných datových úložištích v Národní knihovně ČR.

Výzkumně vývojové aktivity

V rámci těchto aktivit jsou řešeny programové projekty výzkumu a vývoje
- národní, viz http://digit.nkp.cz/projekty/ProjektyVaV.htm
- mezinárodní, viz http://digit.nkp.cz/Projects/index_cz.htm
V průběhu řešení dosavadních úkolů vznikly mnohé prostředky pro speciální oblast historických fondů, vhodné ke zpracování a prezentaci souvisejících informací. Prakticky jsme ověřili postupy dovolující vytvoření katalogu otevřeného pro jakákoli existující data, nezávisle na jejich původní formě.
V krátké době budou na stránkách projektu MEMORIA volně zpřístupněny prostředky pro export/import dat do UNIMARC/MARC 21 a jejich konverzi do formy MASTER.

Z 39.50

Připravuje se poskytování dat prostřednictvím protokolu Z39.50. K systému bude zrealizován Z39.50 server podporující Bath profil Functional Area A level 1. Data budou předávána ve formátu UNIMARC.

OAI

Pro výměnu dat mezi kooperujícími systémy bude použit protokol Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH). Data budou poskytována a přebírána ve formátu XML ve struktuře podle DTD OpenM (NK Praha).

Druhy dat

V databázi Manuscriptorium se setkávají data různého typu.

Data zastupující originály

Digitální obrazy

Tato data si kladou za cíl co nejvěrněji nahradit kontakt s originálními dokumenty a tím originály nejenom ochránit, ale také rozšířit možnosti jejich využívání. Při pořizování obrazových dat je úzkostlivě dbáno na úplnost a věrnost vznikajících informací, včetně zachování informací o barevné kalibraci.
V současnosti je zdigitalizováno 1200 dokumentů, z nichž je nyní **zpřístupněno 1059.** To představuje **přes 500.000 obrazů**.

Plné texty dokumentů

Předpokládáme, že plné texty budou využívat standardu TEI a budou zpřístupňovat zpravidla pragmatické edice historických dokumentů reprezentovaných v otevřeném katalogu historických fondů, případně i v podobě digitální kopie a texty sekundárních dokumentů, vztahující se k originálům.
Tento postup se zatím připravuje a zkouší.

Popisná data

Tato data popisují dokumenty a jsou nezbytná pro rešeršní účely,pro nalezení přístupu k originálům, respektive k datům zastupujícím originály. Jsou to především:

Popisná data vzniklá v souvislosti s digitalizací

Popisy dokumentů v MASTER
Dokumenty vzniklé na základě projektu MASTER představují data, která jsou svým způsobem vzniku a použitým formátem blízká s předchozími. Tato data představují nyní již **cca 5000 záznamů**.

Další data

Existuje množství dalších dat, která mají charakter od podrobných popisů až po stručné inventární seznamy. Nyní je dostupno cca **23.000 záznamů** z těchto zdrojů.

Všechna výše uvedená data mají společného jmenovatele - formát MASTER, ve kterém vznikají nebo do kterého jsou do značné míry převoditelná.

Způsoby zpřístupňování dat

Základem jsou doposud CD-R disky, které jsou použitelné bez instalace speciálních programů jako dostatečná elektronická náhrada přístupu k obsahu originálních dokumentů pro velkou většinu badatelů. Na CD-R discích jsou nyní uchovávány také archivní kopie dat. Stále větší význam si získává **zpřístupnění na Internetu**, kde jsou digitální dokumenty zpřístupňovány v rámci elektronického on-line katalogu – databáze Manuscriptorium. Katalog je opatřen výkonnými vyhledávacími nástroji, které jsou uzpůsobené specifikám oboru.

Výběr digitalizovaných dokumentů

Díky projektu VISK6 a podpoře Ministerstva kultury ČR je možnost digitalizace a zpřístupnění využívána opravdu mnoha institucemi.Výběr digitalizovaných dokumentů je poznamenán množstvím nezávislých přispěvatelů, růzností jejich odborných specializací i motivací k digitalizaci.
Častou motivací je ochranná digitalizace – nahrazení přístupu k příliš využívanému originálu přístupem k jeho digitálnímu obrazu. U institucí, které již zahájily spolupráci s projektem MEMORIA, je však zřejmá snaha o postupné úplné zpřístupnění lokálních kolekcí významných dokumentů i o vytváření kolekcí nadinstitucionálních.
Příznivě se zde projevuje, že návrhy na zařazení do digitalizace schvaluje jmenovaná komise odborníků, tvořící poradní orgán Ministerstva kultury ČR. Celostátně řízený výběr dokumentů evidentně přispívá ke vzniku tematicky souvisejících kolekcí.
Přestože jsou hranice tematických kolekcí již viditelné, rozhodli jsme se pro začátek ponechat jediný kompletní digitalizovaný fond. Rozsáhlé rešeršní nástroje vyhledávacího systému dovolují snadno dospět ke specifické kolekci a konkrétním dokumentům.
Již nyní jsou zřetelné snahy některých badatelů ovlivnit další postup digitalizace konkrétními požadavky na **doplňování vznikajících kolekcí**. Vyhovět těmto potřebám je jedním z úkolů projektu MEMORIA . Je pravděpodobné, že využíváním databáze Manuscriptorium se tento trend prohloubí.

Řízení přístupu k digitálním obrazům

Projekt umožňuje snadné a bezplatné zpřístupnění veškerých dat přinášejících informace o existenci historických dokumentů. Na druhé straně vznikají data, která není obvyklé poskytovat v plné míře zcela volně a bezplatně. Proto bylo rozhodnuto a schváleno Radou projektu, že část informací bude zpřístupněna na základě přidělování nebo prodeje licencí

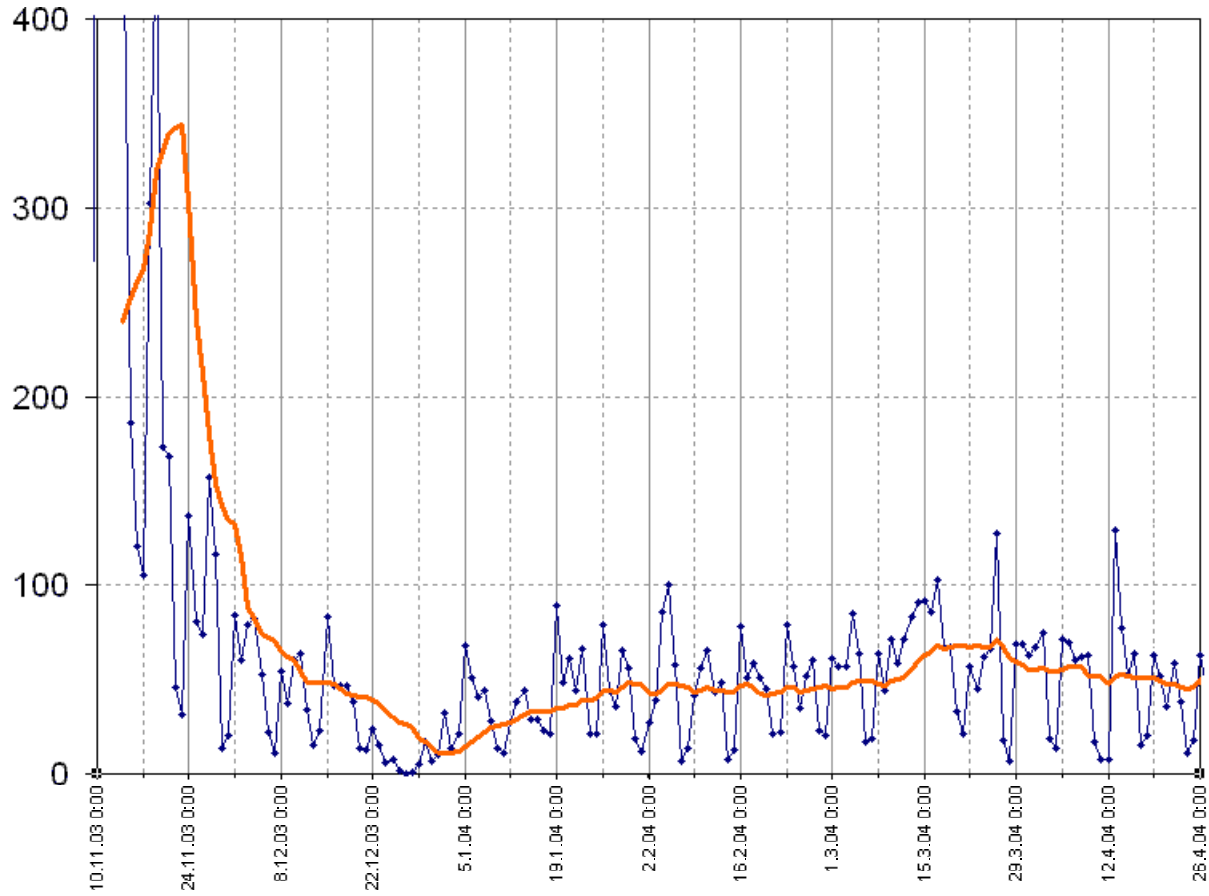
Volný přístup
Projekt MEMORIA předpokládá volné zpřístupnění veškerých dostupných dat, nesoucích informaci o existenci dokumentu a jeho základní popis. U dokumentů, u nichž existuje digitální kopie, je volně dostupný kompletní náhled do celého dokumentu v kvalitě nezbytné pro orientaci v dokumentu.

Licencované zpřístupnění
Podrobnější informace, zejména obrazy vyšší kvality a v budoucnu tzv. plné texty dokumentů, budou zpřístupněny na základě udělení licence, opravňující k jejich využívání. Licence reguluje nakládání s těmito informacemi v zájmu projektu MEMORIA.

Bezplatně jsou zpřístupněny plné digitální obrazy všem aktivně spolupracujícím subjektům.
Hledá se také možnost poskytnutí hromadné licence pro registrované knihovny a odborné školy.
Příjmy z prodeje licencí budou používány na podporu dalšího rozvoje projektu a na podporu digitalizace, zejména na aktivní doplňování kolekcí.

Přístupy z Internetu

Projekt MEMORIA a prezentace jeho výsledků na internetu v databázi MANUSCRIPTORIUM prošla v roce 2003 rychlým vývojem. Množství návštěv www stránek překročilo naše očekávání, typicky se denně připojuje různých 50 pracovišť. Významná část těchto návštěv je ze zahraničí..



Věříme, že projet bude zajímavý a užitečný pro badatele i v době, kdy ztratí punc novosti a stane se nástrojem běžné práce.