Systém Kramerius

Nositel projektu: Národní knihovna ČR, **URL:** http://kramerius.nkp.cz/

Bc. Petr Vlček, 3. semestr (N-AP), 13. prosince 2005

Popis projektu

Projekt Kramerius je podprogram č. 7 programu VISK Ministerstva kultury ČR určený pro záchranu dokumentů ohrožených degradací papíru formou mikrofilmování a digitalizace. Systém Kramerius, vyvíjený v rámci tohoto programu, je open source software určený pro zpřístupnění digitálních kopií archivních a cenných dokumentů a jejich metadat široké veřejnosti v souladu s autorským zákonem. Systém umožňuje doplňování, nahrazování, mazání kompletních dokumentů, metadat nebo obrazových souborů, export a import dokumentů a jejich částí, replikace dat na další instance, jednoduché vyhledávání a propojení s dalšími systémy.

Historie projektu a aktuální stav

V roce 2003 zahájila Národní knihovna ČR spolu s dalšími institucemi projekt zpřístupnění digitalizovaných archiválií prostřednictvím Internetu pod názvem Kramerius. Po proběhnutí výběrového řízení byla realizace systému svěřena firmě Qbizm Technologies, a. s.. Jednou z podmínek výběrového řízení bylo, že systém musí být realizován pomocí open source technologií a být musí uvolněn pod open source licencí (GNU GPL), aby mohla být aplikace poskytnuta zdarma i dalším institucím.

Za účelem standardizace založené na struktuře XML, vytvořila Národní knihovna sadu vlastních pravidel pro popis digitalizovaných objektů ve formátu DTD a XML Schema. Mezi hlavní typy digitalizovaných objektů patří monografie, periodika a muzejní předměty. Kompletní přehled schémat se nachází na [1]. Do této doby byla do systému Kramerius zapracována podpora pro dva typy digitalizovaných objektů: periodika a monografie. Systém dále podporuje protokol OAI PMH, pomocí kterého lze z databáze získávat metadata ve formátu Dublin Core a dále v podporovaných formátech definovaných NKP. Systém podporuje zpracování obrazových souborů ve formátu DjVu a v dalších běžných obrazových formátech.

Na počátku roku 2005 byla do systému Kramerius implementována funkce statického exportu — uložení digitalizovaných dokumentů na CD-R nebo DVD média pro prohlížení off-line.

Vývoj systému stále probíhá a do budoucna se počítá s rozšířením systému o další typy digitalizovaných objektů a formátů metadat.

Cíle projektu

Cílem projektu Kramerius je usnadnit přístup veřejnosti k digitalizovaným dokumentům s možností řídit přístup k jednotlivým dokumentům v souladu s autorskými právy. Dříve byly digitalizované dokumenty poskytovány prostřednictvím CD-R médií, což s sebou přinášelo řadu praktických problémů. Dokumenty byly často rozděleny na mnoho částí a fyzická média byla náchylná k poškození. Zpřístupnění digitalizovaných dokumentů prostřednictvím lokální sítě a Internetu je pohodlnější a rychlejší. Prostřednictvím replikací je umožněno jednoduché sdílení digitalizovaných dokumentů mezi více institucemi.

Popis systému

Použité technologie

Kramerius je implementován v jazyce Java na platformě Java 2 Enterprise Edition. Systém lze provozovat v servlet kontejneru Apache Tomcat verze 4.1.x a pro uložení dat je využita databáze PostgreSQL verze 7.5.x. Ukládání dat probíhá přes abstraktní vrstvu, Hibernate, mapující relace relační databáze na objekty, proto je teoreticky možné použít libovolnou relační databázi podporovanou Hibernate bez nutnosti úprav aplikace. Pro

generování prezentační vrstvy je použita technologie Apache Struts.

Detailní popis systému

Datová vrstva

Systém Kramerius se dá logicky rozdělit na tři vrstvy (viz. Obrázek 1). Nejspodnější je *vrstva datová*, která zahrnuje úložiště obrazových souborů v souborovém systému a relační databázi, kde jsou uložena metadata obrazových souborů a data systému Kramerius.

Aplikační vrstva a její nejdůležitější služby

Rozhraní mezi vrstvou datovou a *vrstvou aplikační* zajišťuje služba pro práci s úložištěm (repository service) a abstraktní vrstva Hibernate pro přístup k databázi, mapující relace v databázi na objekty ve vyšších vrstvách aplikace. Aplikační vrstva je hlavní částí systému Kramerius. Jednotlivé funkce jsou logicky rozděleny mezi služby, většinou běžící jako samostatné vlákno po celou dobu běhu systému. Realizace služeb jako vláken je nutná, protože jednotlivé operace jsou prováděny nad velkým množstvím digitalizovaných objektů (zpracování může trvat hodiny až dny) a je potřeba zajistit rozumnou odezvu aplikace. Složitější operace jako import nebo export objektů se tedy neprovádí ihned po obdržení požadavku, ale jsou těmito službami obsluhovány v krátkých intervalech.

Služba pro *import* se stará o uložení externích objektů do systému Kramerius. Import dokumentů probíhá z předem definovaného importního adresáře v souborovém systému. Před samotným importem je do něj třeba umístit XML soubor s metadaty a odkazy na externí soubory, obrazové soubory a případně textové soubory s obsahem obrazových předloh, určené pro fulltextové vyhledávání. Import probíhá následujícím způsobem:

1. Validace a převod XML dokumentu do ekvivalentní objektové struktury (tzv. marshalling) pomocí nástroje Castor.
2. Analýza metadat a import obrazových a textových souborů do úložiště a do databáze.
3. Uložení metadat do databáze (převod objekty-relace).

Služba pro *export* provádí opačné kroky jako služba pro import. Zvolený dokument, nebo jeho část jsou exportovány do předem definovaného exportního adresáře.

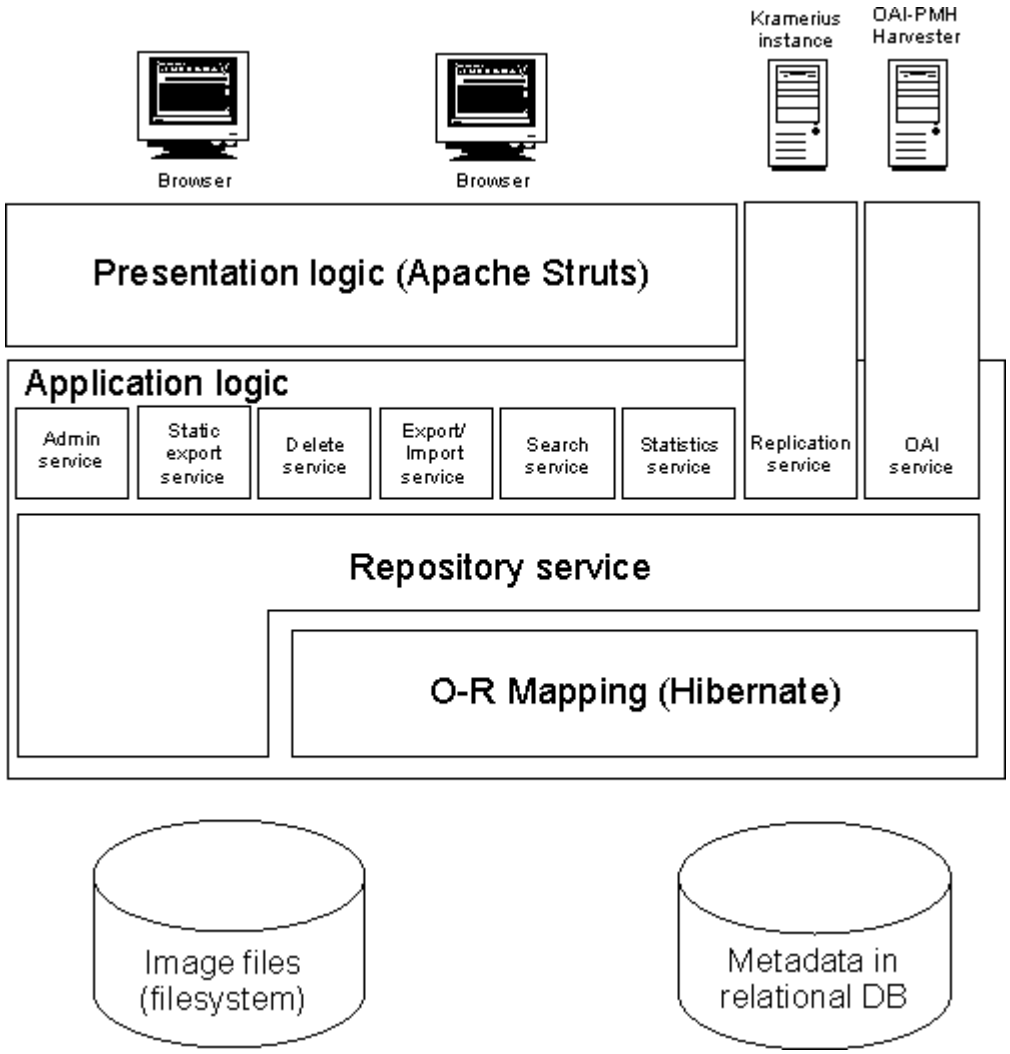
Služba *mazání* se stará o odstraňování metadat označených pro smazání z úložiště a databáze. Mazání probíhá periodicky ve zvláštním vlákně.

Služba pro údržbu *statistik* udržuje přehled o počtu obrazových souborů pro jednotlivé úrovně metadat. Je možné ji spouštět pomocí časovacího mechanismu operačního systému (např. cron).

Služba pro *statický export* provádí export metadat a obrazových souborů zvoleného dokumentu do předem definovaného adresáře na souborovém systému. Výsledkem statického exportu je statická HTML prezentace užívající stejné rozhraní, jako on-line prezentace systému Kramerius. Výsledek lze uložit na CD-R, DVD, nebo jiná paměťová média. Typ média je možné před započetím statického exportu určit a v případě, že velikost exportu překračuje možnosti zvoleného média, je výsledek rozdělen na příslušný počet médií.

Pomocí služby pro *replikaci* je možné sdílet jednotlivé dokumenty mezi více instancemi systému Kramerius. Komunikace probíhá přímo pomocí protokolu HTTP.

Prostřednictvím služby *OAI* systém poskytuje přístup pro vytěžování databáze metadat prostřednictvím protokolu OAI-PMH. Systém Kramerius nabízí metadata ve formátu Dublin Core a ve formátech NKP (monografie, periodika), implementovaných v aplikaci.



Obrázek 1: Architektura systému kramerius

Prezentační vrstva

Systém poskytuje webové rozhraní pro práci implementované pomocí frameworku Apache Struts. Rozhraní má část veřejnou, dostupnou všem uživatelům, a část neveřejnou, přístupnou pouze správcům systému.

Přístupová práva

Důležitým požadavkem na systém bylo, že informace by měly být zveřejňovány v souladu s autorským zákonem. Může se stát, že v systému Kramerius mohou být uložena i díla s datem vzniku po roce 1880, která by neměla být volně dostupná. Za tímto účelem lze pro všechny úrovně metadat nastavit atribut přístupnosti ve dvou úrovních: veřejný a neveřejný. Neveřejné dokumenty lze prohlížet pouze v sídle instituce, která systém provozuje a navenek jsou zpřístupněna pouze metadata. Veřejné dokumenty lze prohlížet kdekoliv na Internetu.

Systém také umožňuje označit jednotlivé dokumenty atributem viditelnosti. Tato možnost je v systému z ryze praktických potřeb. XML dokumenty s metadaty importované do aplikace mohou být velké a je proto nutné importovat objekty po více částech. V aplikaci lze poté data slučovat. Při zviditelnění dokumentu se kontroluje, zda již není zviditelněn dokument se stejným identifikátorem (ISBN, ISSN).

Přístupová práva jsou samozřejmě řízena i službami pro replikaci a OAI-PMH.

Zhodnocení

Se systémem Kramerius jsem měl možnost pracovat i jako vývojář a proto mohu konstatovat, že autoři systém implementovali kvalitně. I přesto však mám pár výhrad. Svázání s druhem poskytovaných metadat se mi zdá příliš těsné a prorůstající celý systém, že v případě rozšíření o další druh poskytovaných metadat by se jednalo

o poměrně pracný úkol.

I přesto, že je systém šířen pod licencí GNU GPL i se zdrojovými kódy, není dostatečně využita myšlenka open source a kromě osob z institucí provozujících systém Kramerius, neexistuje komunita přispívající do systému. Přispět k implementaci je obtížné, protože hlavní úložiště zdrojových kódů není přístupné vývojářské komunitě.

Reference

1. *Národní program zpřístupnění vzácných dokumentů*
2. *Systém Kramerius (Národní knihovna ČR)*
3. Uživatelský portál systému Kramerius
4. PhDr. Jiří Polišenský. *Role systému Kramerius v oblasti tvorby a zpřístupňování digitálních dokumentů*.