



DIGITALIZACE MORAVSKÝCH KNIHOVNÍCH SBÍREK

Moravská zemská knihovna v Brně

**Zpráva za 3. rok řešení projektu
2010**

**Zpracovali:
Ing. Petr Žabička
Mgr. Pavla Švástová**

Listopad 2010, Brno

Obsah

A Konstantační část.....	3
A.1 Rešerše.....	3
A.2 Současný stav.....	3
A.3 Vstupní data.....	5
B Analytická část.....	6
B.1 Vlastní řešení.....	6
B.2 Přínos řešitele.....	7
B.3 Posun znalostí.....	7
C Návrhová část.....	8
C.1 Výsledky řešení.....	8
C.2 Závěr.....	9
C.3 Návrhy opatření.....	9
D Použití účelové podpory.....	10
D.1 Tabulky a komentář.....	10
D.2 Opravy a udržování.....	10
D.3 Cestovné.....	10
D.4 Mzdové náklady.....	10
D.5 Zákonné sociální pojištění.....	10
D.6 Zákonné sociální náklady.....	10
E Resumé a klíčová slova.....	11

A Konstantační část

A.1 Rešerše

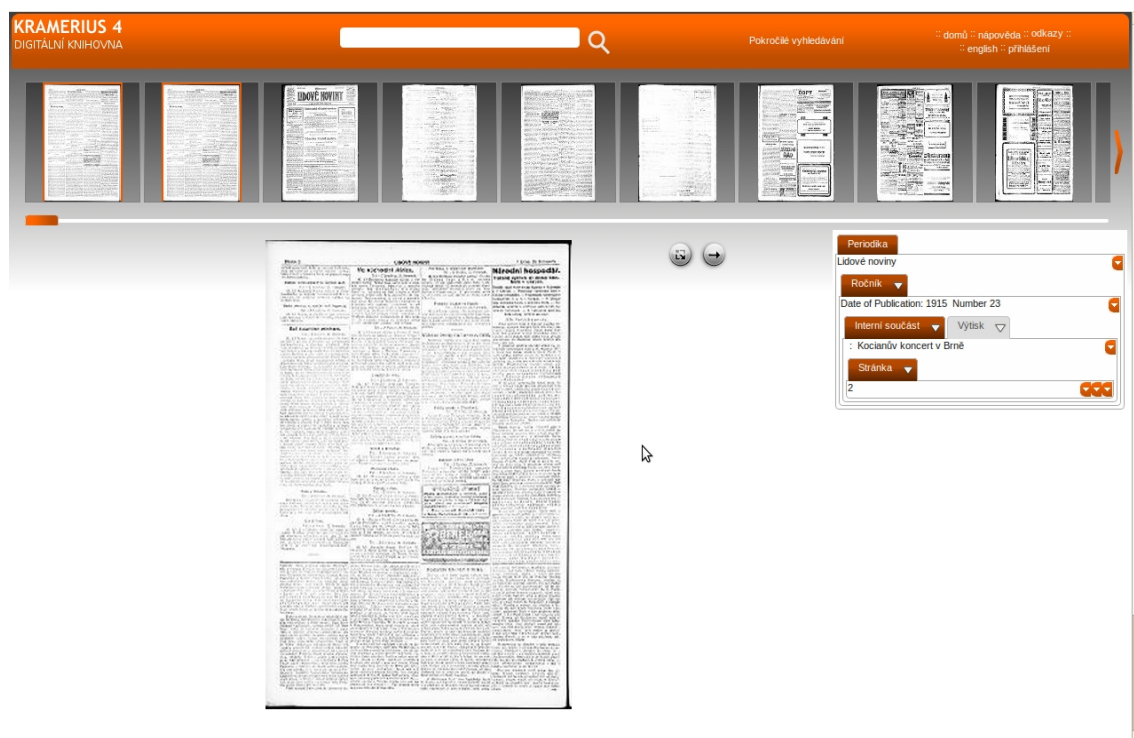
V letošním roce byly vydány dvě publikace obsahující bibliografie periodik Pásmo, Index, Moravsko-slezské revue a Moravsko-slezského sborníku:

- 1) KUBÍČEK, Jaromír. Pásmo 1924-1926. Index 1929-1939. 1. vyd. V Brně : Moravská zemská knihovna, 2010. 104 s. ISBN 978-80-7051-185-5 (brož.)
- 2) KUBÍČEK, Jaromír. PAPÍRNÍK, Miloš. Moravsko-slezská revue 1905-1923. Moravsko-slezský sborník 1918-1920. 1. vyd. V Brně : Moravská zemská knihovna, 2010.

A.2 Současný stav

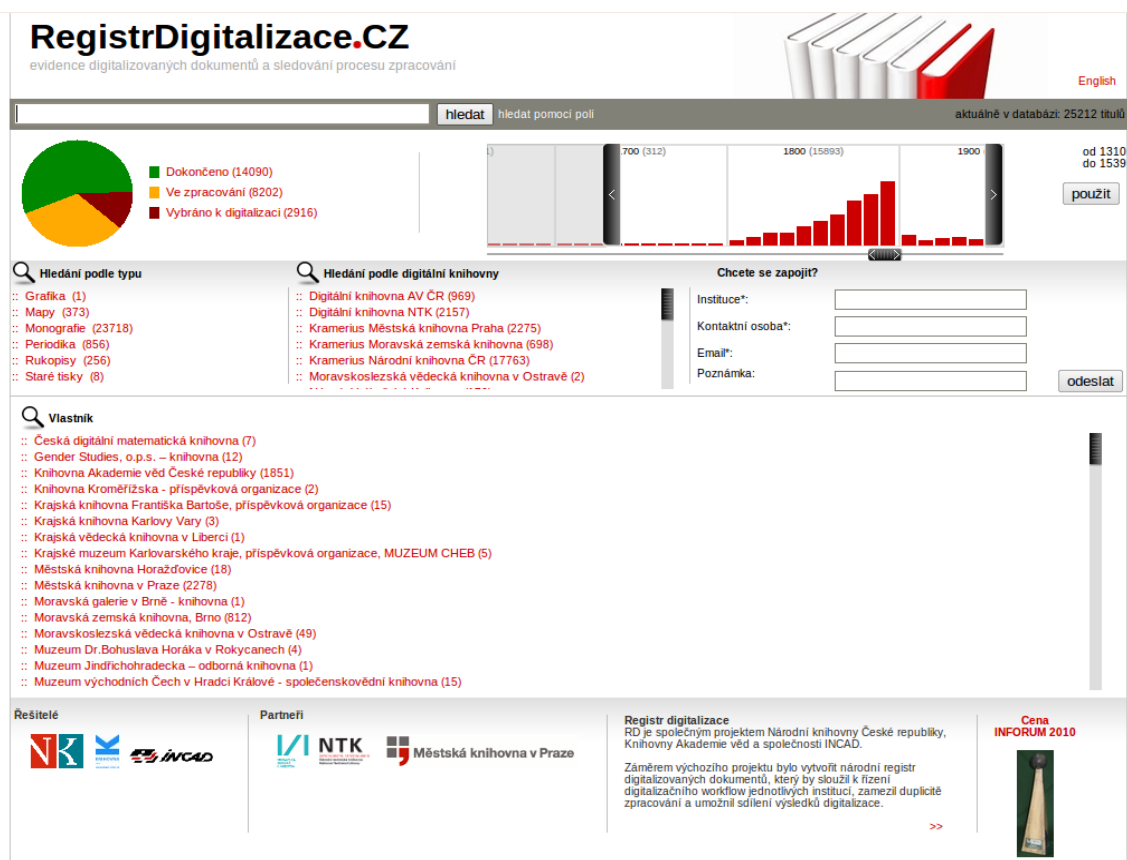
Projekt *Digitalizace moravských knihovních sbírek* má návaznost na řadu dalších projektů Moravské zemské knihovny (dále MZK) i dalších knihoven v ČR. Spolu s Národní knihovnou (dále NK) spolupracujeme na projektu Národní digitální knihovna. Projekt byl schválen, nyní jsme ve fázi příprav výběrových řízení na manažera projektu a systémového integrátora, který by měl zajistit nákup veškeré techniky, hardwaru a softwaru, jejich dodání na místo určení, zapojení, nainstalování a vzájemné integrace. Začátek provozu se v MZK předpokládá na polovinu roku 2011. Výsledky projektu *Digitalizace moravských knihovních sbírek* tak budou uloženy k archivaci v LTP systému, který bychom měli získat v rámci projektu *Národní digitální knihovna*. V současnosti probíhají analýzy a počáteční práce na konverzích do formátů specifikovaných pro LTP.

V září byla vypuštěna ve stadiu betaverze čtvrtá verze digitální knihovny Kramerius založené na digitálním repozitáři Fedora Commons, která podporuje zobrazování článkových metadat a stane se novou prezentační vrstvou pro veškeré zdigitalizované dokumenty. Strukturální metadata jsou zde ukládána ve formátu FOXML a popisná metadata ve formátu MODS. V současnosti probíhá testování a konverze ze staré verze Krameria, kde byla metadata uložena ve formátech DTD pro periodika a DTD pro monografie. Demoverze i s článkovými metadaty Lidových novin vytvořených v projektu *Digitalizace moravských knihovních sbírek* je dostupná na adrese <http://krameriusedemo.mzk.cz>. Po prvních testech jsou v současnosti laděny problémy s rychlostí zobrazování obrazových souborů ve formátu DJVU. Plánuje se výměna obrázků DJVU za archivní JPEG, což bude zřejmě poněkud časově náročné.



Obr. č.1: Digitální knihovna Kramerius verze 4

S prezentací zdigitalizovaných dat souvisí projekt Registr digitalizace, do kterého je MZK aktivně zapojena od loňského roku. Registr digitalizace je společným projektem Národní knihovny ČR a Knihovny Akademie věd, jeho realizaci zajišťuje firma Incad. Základní myšlenkou projektu bylo shromáždit informace o dokumentech digitalizovaných na našem území na jediné místo, aby se zabránilo zbytečným duplicitám při stále narůstajících objemech digitalizovaných dat. V první řadě je tedy určen pro knihovny (a další instituce), které digitalizují nebo to mají v plánu. Zjistí zde, zda dokument, který se chystají zpracovat, nedigitalizoval již někdo jiný, s vlastníkem dokumentu se můžou domluvit, zda by neposkytl kopii pro zveřejnění ve vlastní digitální knihovně apod. Knihovny, které mají vybudovanou digitalizační linku, můžou využít Registr digitalizace jako evidenci vlastních digitalizačních aktivit a jako nástroj pro řízení workflow. Další vývoj navazuje na projekty Národní digitální knihovna a Kramerius verze 4, proběhl přechod na novou vnitřní datovou strukturu, která umožní lepší navázání na metadata z knihovních katalogů a digitální knihovny Kramerius tak, aby bylo možné většinu procesů řešit automatizovaně. Zároveň by mělo být lépe řešeno přidělování perzistentních identifikátorů novým digitálními objekty, což je důležité např. právě pro články.



Obr. č.2: Registr digitalizace

V projektu Digitalizace moravských knihovních sbírek jsou využity výsledky jiného projektu MZK – Staré mapy online. Konkrétně se jedná o implementaci image serveru IIPImage do Krameria verze 4. Image server umožňuje efektivní a rychlé zobrazování velkoformátových obrázků uložených ve velkém rozlišení i s pomalým internetovým připojením.

Vytvářet článková metadata a publikovat je v Krameriu verze 3 lze pomocí metadatového editoru vyvinutého v MZK v projektu hrazeném z VISKu 3 Obálky knih a metadatový editor. Z důvodu přechodu na Krameria verze 4 byl zahájen vývoj nové verze metadatového editoru, který bude podporovat metadatové formáty MODS a FOXML a umožní editaci a vytváření metadat právě pro tuto verzi.

The screenshot shows the 'Metadata Editor' interface. At the top, there's a navigation bar with links like 'search', 'statistics', 'authors', 'verify', 'help', and 'admin'. Below this, a breadcrumb trail shows the current path: 'ME / serial / Powidatel aneb. Sbírka rozmanitých pamětí hodných příběhů, powídaček, důvtipných nápadů a přípořádů k poučení a obveselení / Volume 1 / Issue 0 / --- Undefined article name --- /'. The main title of the editor is '--- Undefined article name ---'. Below the title, there are two buttons: 'Save' and 'Save and Next'. The form on the left is organized into sections with expandable/collapsible icons (plus and minus signs). The sections include: 'Unique Identifier' (with fields for URNType and SICType), 'Core Bibliographic Description Periodical' (with fields for Title, Title Group, Main Title, Sub Title, Parallel Title, Sorting Title, Key Title, and Coden), 'GMD' (with fields for Publisher, Printer, and Physical Description), 'Series', 'Language', 'Subject' (with UDC and DDC fields), and 'Keyword'. Each field has a corresponding 'x' icon for deletion. On the right side, there's a preview of the document page, showing the title 'Propow/dky a podobenswji.' and the beginning of the text. The preview is labeled 'prev | next | references' and has navigation controls '<< | < | > | >>'. The page number '186' is visible in the top left corner of the preview.

Obr. č.3: Úprava článkových metadat v metadatovém editoru MZK

A.3 Vstupní data

Nadále probíhá zpracování článkových metadat. Přepis z katalogizačních lístků byl ukončen a nyní probíhá excerptce přímo ze svázaných ročníků Lidových novin a dalších periodik. Metadata jsou nadále zapisována do excelové tabulky. V letošním roce byla podrobně analyzována data z dosud vytvořených excelových tabulek, které obsahují přepisy článkových metadat z knihovních lístků a na základě této analýzy byla provedena konverze a zkušební import do digitální knihovny Kramerius. Oproti původnímu plánu z loňského roku převést metadata do formátu podporovaném Krameriem verze 3 (DTD pro periodika), bylo rozhodnuto, že budou převedena do formátu, který umožňuje uložení a zobrazení v Krameriu verze 4 (FOXML a MODS).

B Analytická část

B.1 Vlastní řešení

Jedním cílů na letošní rok bylo vytvořit program, který metadata článků z excelových tabulek převede do XML ve formátu, který lze vložit do systému Kramerius verze 4. Z analýzy vyplynulo několik problémů, které souvisí s dosavadním systémem vytváření záznamů a jejich zapisování do excelových tabulek. Každý řádek tabulky reprezentuje jeden článek a obsahuje metadata deskriptivní a strukturální. V deskriptivních metadatach není jednotně řešeno jméno autora – někdy je ve tvaru „Příjmení, Jméno“ (např. Příkryl, Ondřej), někdy je uveden pseudonym, pod kterých psal do novin a jméno ve tvaru [„Jméno Příjmení“] (např. -is {=Josef Hais?}), proto není možné jednoduše rozčlenit do jednotlivých polí v XML. Navíc dosavadní krameriovské XML neumožňuje zadat alternativní jméno autora.

Další problém je ve sloupci, který stručně popisuje, o co se ve článku jedná a slohový útvar. Z žánrů jsou nejčastěji zastoupeny recenze, fejetony, referáty, sloupky a úvahy. Data nejsou zcela jednotná, takže jsme se zatím rozhodli vložit celý obsah do prvku *anotace*. Podrobnější rozčlenění na žánry je možné, ale nepřesné, např. v řádku 6 v tabulce výše se nachází text „reakce na fejeton“, pokud by probíhalo rozčlenění na žánry podle pravidla: „pokud najdeš v textu slovo „fejeton“, přidej článku atribut „fejeton“, neodpovídalo by to v tomto případě pravdě, jelikož se nejedná o fejeton, ale pouze reakci na něj.

Není úplně systémově řešen zápis strukturálních metadat v případech, kdy se jedná o článek na pokračování. Reálně existují dvě možnosti:

- 1.) Článek vycházel na pokračování v souvislé řadě – potom je datum ve sloupci G ve tvaru „od_data-do_data“ a číslo ve sloupci H ve tvaru „od_číslo-do_číslo“ (viz tabulka, řádek 8)
- 2.) Článek není v souvislé řadě – pak jsou uvedena všechna data oddělená čárkami ve sloupci G a k nim příslušný počet čísel atd. (viz tabulka, řádek 9)

Z těchto dat by obecně nebylo složité vyextrahovat strukturální metadata a na základě nich přiřadit deskriptivní metadata k příslušnému ročníku, číslu a stránce, uložené v digitální knihovně. Na rozdíl od DTD pro periodika Krameria verze 3, kde je v jednom XML souboru obsaženo vše od strukturálních přes popisná, administrativní a technická metadata, Kramerius 4 obsahuje XML soubory zvlášť k jednotlivým typům metadat i k jednotlivým titulům, ročníkům a číslům a vše je logicky propojeno RDF vazbami. Proto jsme vytvářeli pro každý článek zvlášť MODS pro popisná metadata (název, autor, žánr,...) a zvlášť FOXML soubor s odkazy na obrázky a vazbami na příslušné číslo, ročník a titul, ve kterém se nachází. Vše je propojeno pomocí identifikátorů, které články neobsahovaly a bylo nutné jim nějaké přidělit.

Analýzou strukturálních metadat byla zjištěna nekonzistence dat v některých řádcích tabulky. Typickým problémem je, že nesouhlasí počet stran a počet příslušných čísel u článků na pokračování, chybějící označení, zda se jedná o ranní či odpolední vydání, nejednotné oddělovače a nechtěné překlady. Chybovost je zapříčiněna ve většině případů lidským faktorem a dále ne zcela šťastným návrhem původní struktury excelové tabulky. Z analyzovaných dat jsou zpětně vytvářeny nové tabulky, které obsahují chybné nebo nejednoznačné záznamy z původních tabulek. Obecně lze však říct, že vzhledem k celému objemu všech zpracovaných článkových metadat je chybovost poměrně nízká.

První vzorek konvertovaných metadat byl importován do demoverze Krameria 4, kde jsou články označeny jako „interní součást“. (Viz obrázek č. 1). Další import je očekáván po instalaci ostré verze Krameria 4 pro MZK a replikaci všech dat z Krameria 3. Instalace již existuje, nyní se replikují data, což je časově poměrně náročná operace. Dokončení replikace a import článkových metadat se očekává do konce roku 2010.

Z vybraného vzorku periodik a monografií byly zdigitalizovány a publikovány v Krameriu verze 3 následující tituly: Lidové noviny, Povídatel, Večery a Kniha pro každého Moravana. K ostatním dokumentům jsou vytvářena metadata. Do konce roku proběhne replikace a zveřejnění v Krameriu verze 4. Díky spolupráci s řešiteli projektu pro vývoj Krameria verze 4 byla v letošním roce testována funkčnost Krameria 4 pro zobrazování a proběhl import první várky metadat k několika ročníkům Lidových novin.

B.2 Přínos řešitele

Hlavní přínos projektu v tomto roce spočíval v analýze článkových metadat a zjišťování toho, jak se která metadatová schémata s takovým popisem vyrovnávají. Vzhledem k chystanému přechodu na Krameria verze 4, který používá schémata FOXML pro strukturální a MODS pro popisná metadata, se upustilo od konverze do formátů, která využívala třetí verze Krameria. Tyto analýzy budou využity v chystaném projektu Národní digitální knihovna, s metadaty k článkům se počítá i v dalších projektech – Kramerius, ve kterém zatím článková metadata nebyla dříve publikována, metadatový editor, kde budou metadata vytvářena, Registr digitalizace atd.

Pro Kramerius verze 4 byla vytvořena specifikace formátu MODS pro článková metadata, která bude sloužit i pro potřeby popisu zdrojů v projektu Národní digitální knihovna.

Dalším přínosem projektu v letošním roce je analýza a následná implementace image serveru IIPIImage do Krameria verze 4. Image server řeší doposud problematické a poněkud uživatelsky nepřívětivé zobrazování větších formátů dokumentů, jako jsou například mapy nebo právě stránky novin. V budoucnu to může otevírat cestu k využití anotačních nástrojů vyvíjených v projektu Europeana Connect, které jsou založeny na technologii OpenLayers.

B.3 Posun znalostí

Operativním řešením problémů, které vyplynuly z analýzy článkových metadat, by byla změna v systému vyplňování excelové tabulky – bude potřeba zavést větší granularitu při vyplňování – každý údaj do zvláštního sloupce, logičtější rozdělení strukturálních metadat, apod., aby byla tato data lépe strojově zpracovatelná. Nejlepším řešením ovšem bude zapisování příslušných metadat přes metadatový editor vyvíjený MZK, díky tomu budou články okamžitě navázány na příslušný sken stránky jejich publikování v Krameriu 4. Funkcionalita pro vkládání článků byla doposud vyvíjena v metadatovém editoru, který MZK používá pro publikování digitalizovaných dokumentů v Krameriu verze 3. Nespornou výhodou zapisování článků přímo přes metadatový editor je eliminace chyb ve strukturálních metadatech. Takové chyby se v excelové tabulce vyskytují a jsou určující pro to, zda se příslušný metadatový záznam sváže s konkrétním skenem stránky, na kterém se příslušný článek nachází. Neovlivní to však překlepy a chyby v popisných metadatech, které vedou k tomu, že článek není v takovém případě vyhledatelný.

C Návrhová část

C.1 Výsledky řešení

Cíle ze zadání projektu a jejich plnění jsou následující:

- 1) *Analytické zpracování obsahu 10 titulů, zpracování souboru analytických záznamů a jejich uvedení jako metadata do vyhledávacích systémů*

Práce na konverzi metadat stále probíhají. Celkově je ke konci října zpracováno 60649 záznamů článků z celkem 10 periodik. Dále se pokračují analytické práce na Lidových novinách.

- 2) *Digitalizace vytipovaného souboru dokumentů a jejich zpřístupnění na regionálním datovém úložišti. Zpracování formálních a obsahových metadat, zpracování metodiky tvorby a zpřístupnění obsahových metadat.*

Byla provedena podrobná analýza a praktická zkouška toho, co zpřístupnění obnáší. Bylo zjištěno, že jelikož v Krameriu 3 nebylo s článkovými metadaty počítáno, není zobrazení těchto informací uživatelsky přívětivé. Vzhledem k relativně rychlému vývoji Krameria 4 bylo rozhodnuto, že budeme konvertovat článková metadata do formátu, který podporuje tato verze. V tomto směru byla provedena nová analýza a zajištěny programátorské práce, které konverze a import umožňují.

- 3) *Vyvinout jednoduchý program, který automaticky vytvoří data použitelná pro provázání záznamu článku do plného textu v digitální knihovně Kramerius*

Program byl vytvořen, otestován a první várka metadat byla importovaná k digitálním dokumentům v Krameriu 4. K vidění na adrese: <http://krameriusedemo.mzk.cz/>.

- 4) *Posouzení stávajících SW pro zpracování obsahových metadat u knih a článků z časopisů a novin, jejich dopracování a optimalizace pro ověření k vypracování postupů k provozní praxi*

V metadatovém editoru vyvíjeném v MZK lze vytvořit popisná i strukturální metadata k seriálům až na úroveň článků a publikovat je v systému Kramerius 3. Nyní probíhá vývoj nové verze editoru, který dokáže takto editovat záznamy a vytvářet články i v Krameriu verze 4.

- 5) *Přínos v aktuální problematice digitalizace knihovních fondů v paměťových institucích*

MZK se významnou částí podílí na vývoji nových SW nástrojů, které jsou využívány v oblasti digitalizace dokumentů. Jednak je to projekt Kramerius verze 4, kde byla již v loňském roce iniciována integrace článkových metadat a v letošním roce byl implementován image server IIPImage. Dále je testováno zobrazování obrázků v nestandardních velikostech a formátech. Doposud se v digitální knihovně využíval jako prezentační formát pro user copy DJVU a JPEG, v rámci projektu Národní digitální knihovna začnou vznikat JPEG2000 soubory, které dokáže IIPImage poměrně rychle zobrazit.

V projektu Registr digitalizace jsme se podíleli v letošním roce hlavně na přechodu na novou vnitřní datovou strukturu, která byla předtím vystavěna hlavně pro potřeby Národní knihovny a Knihovny akademie věd. Snažili jsme se o to, aby byla více univerzální a počítala i s jinými, poněkud nestandardními, typy dokumentů – mapami, grafikami, starými tisky a rukopisy, hudebninami apod. V návaznosti na toto se řeší možnosti zapojení nových knihoven.

Praktické poznatky z tohoto projektu též uplatňujeme při plánování procesů v projektu Národní digitální knihovna. Týká se to především analýz a testování různých typů metadatových schémat a z nich plynoucích návrhů na specifikace formátů apod.

- 6) *Tištěné publikační rejstříky osmi excerpovaných časopisů z 18.-20.století*

Tištěné rejstříky časopisů byly vydány a zároveň publikovány i v elektronické podobě v digitální knihovně Kramerius MZK.

C.2 Závěr

Ve 3. roce projektu se pokračovalo na zpracování metadat z knihovních lístků a jejich přepisu do excelové tabulky. Zároveň byly provedeny analýzy, které by tuto činnost měly urychlit a co nejvíce zjednodušit a zároveň zamezit výskytu chyb zapříčiněných ruční prací. Byl vytvořen program, který převede metadatové záznamy z tabulky do XML ve formátu, který podporuje Kramerius 4 – FOXML pro strukturální a MODS pro popisná metadata. Byl proveden úspěšný import první várky článkových metadat do testovací verze Krameria 4 a do konce roku se plánuje import zbylých metadat do ostré instalace pro MZK.

Do Krameria verze 4 byl implementován IIPIImage – image server vyvíjený v MZK v projektu Staré mapy online. To umožní efektivnější zobrazování a tím pohodlnější prohlížení větších formátů zdigitalizovaných stránek periodik.

C.3 Návrhy opatření

Během řešení se ukázalo, že je potřeba posílit softwarový vývoj a analytiku namísto ruční práce. S tím souvisí plány na příští rok – testování publikování přes nový metadatový editor do Krameria, analýza a detailnější prověření možností schématu METS/ALTO. Dále analýza možnosti provázání na jmenné autority a možnosti přivázání článkových metadat z databáze ANL na dokumenty digitalizované v příštích letech v projektu Národní digitální knihovna.

Uplatnění dalších výsledků bude záviset na skutečné podobě pokynů k poskytování výsledků pro RIV v průběhu jara 2011. Vzhledem k tomu, že v průběhu jara 2010 se tato pravidla několikrát změnila, očekáváme podobný vývoj i letos a nemůžeme tedy předpovídat, které výsledky bude možné uplatnit.

V příštím roce budeme žádat, pokud to bude možné, o uznání výsledku typu Z_{tech} – Ověřená technologie za Implementaci image serveru v Krameriu verze 4 a novou verzi metadatového editoru implementovaného do Krameria 4.

D Použití účelové podpory

E Resumé a klíčová slova

Resumé:

V projektu Digitalizace moravských knihovních sbírek je řešena problematika získávání článkových metadat z historicky významných periodik a jejich následné automatické zpracování. V letošním roce se zkoumalo hlavně zpřístupnění získaných a konvertovaných metadat a efektivní propojení na již existující data v digitální knihovně Kramerius. K tomu byl v letošním roce do Krameria verze 4 implementován image server IIPIImage, který umožní efektivnější zobrazování stránek novin v lepších rozlišeních a nových formátech.

Projektu je přínosný i v dalších aktivitách českých i zahraničních knihoven v oblasti digitalizace – projektech Národní digitální knihovna, Kramerius 4 a Registr digitalizace. V současnosti se řeší zejména specifikace metadatových formátů, identifikátory digitálních objektů, problematika dlouhodobého uchovávání digitálních dokumentů a teoretické poznatky z tohoto projektu jsou tak přínosem i v projektech souběžných či navazujících.

Klíčová slova:

Digitalizace, digitální knihovny, systém Kramerius verze 4, informační technologie, knihovní sbírky, bibliografické texty, metadata, Registr digitalizace, Národní digitální knihovna,