

ZDOKONALENÍ VIRTUÁLNÍHO BADATELSKÉHO PROSTŘEDÍ MANUSCRIPTORIA: VÝVOJ NÁSTROJŮ PRO DALŠÍ AGREGACI A ZPŘÍSTUPNĚNÍ DIGITALIZOVANÝCH HISTORICKÝCH DOKUMENTŮ

Analýza potřeb a nástrojů – vstupní analýza projektové integrace

Obsah

Úvod.....	4
Schéma návrhu integrace nástrojů pro další agregaci a zpřístupnění digitalizovaných historických dokumentů	4
Generátor jednoznačného identifikátoru digitálního dokumentu	5
Analýza uživatelských požadavků	5
Popis navrženého řešení	5
Použité technologie	5
Prohlížeče	5
Integrace v rámci systému Manuscriptorium	5
Dostupnost aplikace	6
Soubory aplikace.....	6
Popis API rozhraní.....	6
Ukázka volání API	6
Odezva API	7
Popis uživatelského rozhraní aplikace generátoru	7
Chybová hlášení generátoru	8
Závěr	9
Validátor balíčků VISK 6	10
Analýza a specifikace uživatelských požadavků.....	10
Očekávaná struktura v ZIP souboru	11
Popis řešení	11
Použité technologie	11
Zabezpečení aplikace	11
Prohlížeče	11
Integrace	12
Popis užití validátoru a postupu kontroly.....	12
Očekávaná struktura balíčku	15
Prováděné kontroly	15
Chybová hlášení aplikace	16
Závěr	16
Zobrazování pozicování vyhledaných slov v obrazech digitálních dokumentů	18
Cíle řešení	18
Návrh řešení	18
Rozšíření správního systému Manuscriptoria	19
Rozšíření Rešeršní platformy a Webové části aplikace Manuscriptorium	19
Implementace	21

Závěr	22
Propojení katalogu NKČR a MNS	22
Cíle řešení	22
Návrh řešení	22
Implementace	23
Přístup k výsledkům řešené úlohy	23
Konvertor EoD (eBooks on Demand)	24
Specifikace funkčních požadavků	24
Popis vstupních a výstupních dat	24
Očekávané vstupy z oddělení mikrografie a digitalizace (odbor Historických a hudebních fondů) NKČR.....	25
Parametry výstupu ze zařízení PS7000C	25
Katalogizační popis.....	25
Technické parametry obrazových dat.....	25
Identifikace jednotlivého obrazu	26
Struktura názvů souborů a složek	26
Přílohy	26
Literatura	27

Úvod

Tento dokument obsahuje vstupní analýzu projektové integrace a potřeb nástrojů pro zdokonalení virtuálního badatelského prostředí Manuscriptoria (dále MNS). Předmětem analýzy integrace byly nástroje pro online validaci balíčků VISK 6, konvertoru EoD (eBooks on Demand), generátoru jednoznačného identifikátoru digitálního dokumentu, zobrazení pozicování vyhledaných slov v obrazech digitálních dokumentů a propojení katalogu NKČR s MNS.

AiP Beroun dlouhodobě spolupracuje s NK na rozvoji systému Manuscriptorium (MNS). Díky mnohaletým zkušenostem jsme mohli navrhnout optimální řešení integrace nových nástrojů a provést analýzu uživatelských požadavků. Požadavky formulované níže v tomto dokumentu predikují užití technologií a návrhy řešení jednotlivých úkolů. Validátor balíčků VISK 6, generátor jednoznačného identifikátoru digitálního dokumentu a zobrazování vyhledaných slov v obrazech digitálních dokumentů musí být řešeny jako rozšíření stávajícího badatelského systému MNS a jeho modulů, zatímco konvertor pro EoD by měl být především výkonnou aplikací zapojenou do digitalizační linky pracoviště oddělení mikrografie a digitalizace (odbor Historických a hudebních fondů) a jeho nového skeneru PS7000C. U tohoto konvertoru bude jeho integrační povaha zajištěna přesným dodržáním konvencí a standardů definovaných pro digitalizaci v programu VISK 6 a tím bude dosaženo i přímé návaznosti na systémy a nástroje užívané v NKČR.

Schéma návrhu integrace nástrojů pro další agregaci a zpřístupnění digitalizovaných historických dokumentů

Níže uvedené schéma zobrazuje rozdělení nástrojů MNS v rámci celého badatelského prostředí. Zařazení do příslušného sloupce odpovídá místu přístupu k danému nástroji a návrhu řešení integrace. Žlutou barvou jsou vyznačeny nové nástroje a moduly MNS, o kterých pojednává tento dokument.

Domovská stránka MNS / CMS	Digitální knihovna (DK)	Ostatní (samostatné) nástroje MNS
Novinky	Vyhledávání	M-Tool 2.0
Naposledy přidané dokumenty	Zobrazení dokumentu	MNS Kandidátů 2.0
News Letter	Zobrazení faksimile	Gaiji bank
Správa uživatelských účtů a skupin	Virtuální dokumenty	EOD konvertor
Validátor balíčků VISK 6	Kolekce	
Generátor jednoznačného identifikátoru digitálního dokumentu	Oblíbené položky	
	Pozicování hledaných slov	
	Propojení katalogu NK s MNS	

Nástroj pro validaci balíčků VISK 6 a generátor jednoznačného identifikátoru digitálního dokumentu jsou dle specifikace požadavků nezávislé na datech digitální knihovny, nikterak nekooperují s metadaty, faksimile, ani s obsahem vytvořeným uživateli v rámci jejich osobních účtů. Oba nástroje však budou přístupné jen registrovaným a přihlášeným uživatelům s patřičně nastavenými přístupovými právy. Z tohoto důvodu spadají oba nástroje do oblasti MNS, která zajišťuje správu uživatelů a přístupových práv. Tato část je vyvíjena nad CMS Drupal. Nástroj pro generování jednoznačného identifikátoru digitálního dokumentu mimo to přímo využívá databázi identifikátorů institucí a knihoven, které přiděluje NKČR. Tato databáze je interní systémová databáze systému MNS. Oba nástroje mají sloužit zájemcům o digitalizaci v rámci programu VISK 6 především pro ověřování vlastních výstupů z digitalizace proti definici a specifikaci VISK 6.

Naproti tomu nástroje pro pozicování hledaných slov v obrazech digitálních dokumentů a propojení katalogu NKČR s MNS jsou na datech zpřístupněných v digitální knihovně zcela závislé a přímo s nimi pracují.

Samostatným nástrojem je pak realizovaný EOD konvertor. Tento nástroj je určen pro samostatné pracoviště v NKČR.

Podrobněji se uvedenými nástroji zabývají následující kapitoly.

GENERÁTOR JEDNOZNAČNÉHO IDENTIFIKÁTORU DIGITÁLNÍHO DOKUMENTU

Cílem úlohy bylo vytvořit, zprovoznit a zpřístupnit on-line aplikaci, která umožní na základě vstupních údajů o místě uložení a signatuře či jiném označení dokumentu vygenerovat správný jednoznačný kód digitálního dokumentu ve tvaru, jak je vyžadován v definici programu VISK 6. Konstrukce identifikátoru digitálního dokumentu je shodná z konstrukcí identifikátoru fyzického dokumentu, který je používán v dalších projektech souvisejících s Manuscriptoriem.

Aplikace disponuje grafickým rozhraním pro koncové uživatele a také API rozhraním pro využití v libovolných partnerských systémech prostřednictvím protokolu http. Tímto nástrojem se výrazně zjednoduší generování příslušných údajů zejména pro další zhotovitele digitalizace. Služba generátoru bude s ohledem na schéma návrhu integrace nástrojů řešena jako modul systému pro správu obsahu (CMS), který je součástí prezentační vrstvy Manuscriptoria.

Analýza uživatelských požadavků

Na základě konzultací s vybranou skupinou přispěvatelů do MNS a detailní znalosti badatelského systému MNS jsme dospěli k formulaci uživatelských požadavků na Generátor jednoznačného identifikátoru digitálního dokumentu. Tyto požadavky vychází z praktických zkušeností dlouholetých uživatelů MNS a odborných názorů spolupracovníků při vývoji MNS.

Seznam požadavků na Generátor jednoznačného identifikátoru digitálního dokumentu:

- Aplikace musí být dostupná po internetu pomocí http protokolu prostřednictvím nejpoužívanějších prohlížečů (Internet Explorer, FireFox, Chrome).
- Aplikace bude požadovat minimální počet vstupních hodnot, tedy jen jméno instituce (místa uložení) a signatura či jiné označení dokumentu.
- Bude navrženo jednoduché a přehledné uživatelské rozhraní (UI).
- Jednoznačný kód instituce, který přiděluje partnerům NK, si zjistí aplikace sama z aktuálního seznamu.
- Aplikace bude automatizovaně vždy přistupovat do aktuálního seznamu kódů institucí fungujícího jako samostatná databáze systému Manuscriptorium.
- Aplikace kontroluje zadání obou vstupních hodnot uživatelem a v případě zjištění chyb uživatele upozorní.
- Uživatelské prostředí (UI) generátoru bude prakticky a vizuálně integrované do prostředí MNS.
- Výstup generátoru prostřednictvím UI bude zpracovatelný přes schránku počítače (Ctrl+C, Ctrl+V).
- Generátor bude dostupný přes http protokol prostřednictvím vlastního rozhraní.
- Generátor je dostupný pouze registrovaným uživatelům Manuscriptoria.

Popis navrženého řešení

Použité technologie

Pro řešení úkolu s ohledem na požadovanou integraci do systému Manuscriptorium (MNS) byly použity technologie již využívané tímto systémem. Aplikace Generátoru je vytvořena pomocí technologií PHP, HTML, CSS, JavaScript.

Prohlížeče

Aplikace byla testována a optimalizována pro následující prohlížeče a jejich verze:

- Internet Explorer, verze 8 – 11
- Google Chrome, od verze 38
- Mozilla Firefox, od verze 30
- Opera 30
- Safari 5.1

Integrace v rámci systému Manuscriptorium

Aplikace je řešena jako modul CMS (Drupal) prezentační vrstvy Manuscriptoria. Tento modul po instalaci do CMS vytvoří blok, tedy funkční prvek umístitelný do stránek systému MNS. Aplikace provádí automatizovaně kontrolu seznamu institucí a jejich identifikátorů na referenčním FTP umístění. Tento seznam pro objednatele spravuje a udržuje jako samostatnou databázi AiP Beroun v rámci provozu MNS.

Dostupnost aplikace

Generátor je dostupný pouze pro oprávněné uživatele na adrese: <http://www.manuscriptorium.com/en/node/5090>

Soubory aplikace

Součástí řešení je i předání zdrojových kódů popisovaného řešení. Zdrojové kódy se nacházejí v těchto souborech, které jsou umístěny na předaném CD:

- `generator_Fyzid.info` (Tento soubor obsahuje popisné informace pro CMS Drupal - jméno modulu, popis, verze)
- `generator_Fyzid.module` (Hlavní soubor aplikace se všemi skripty a funkcími aplikace)
- `generator_api.php` (Rozhraní pro programové volání funkcí generátoru přes http)

Popis API rozhraní

Pro volání funkce generátoru prostřednictvím API jsou definovány tyto parametry:

- `repository`: // celé jméno instituce s diakritikou a rozlišením velkých a malých písmen
- `shelf_mark`: // signatura nebo jiné označení dokumentu

Rozhraní je přístupné na tomto URL:

http://www.manuscriptorium.com/modules/generatorFyzID_module_Drupal/generator_api.php

Ukázka volání API

Následující ukázka kódu v JavaScriptu je příkladem naplnění parametrů a volání funkce generátoru:

```
<script>
var apiUrl = "http://www.manuscriptorium.com/modules/generatorFyzID_module_Drupal/generator_api.php";

function g2cms_callApi(params, output, err) {
    $.ajax({ type: "POST", async: true, url: apiUrl, dataType: "json", data: { parameters: params },
    success: output, error : err });
}

function logResponse(jqXHR) {
    var output;
    console.log(jqXHR);
    console.log(JSON.stringify(jqXHR));
    console.log(JSON.stringify(jqXHR.result));
}

var params = {
    repository_code:'NLOTRO',
    shelf_mark:'097978',
    repository:'National Library of the Republic of Kazakhstan'
    };
console.log(params);
g2cms_callApi(params, logResponse, logResponse);
</script>
```

Odezva API

Pro následující zadané hodnoty:

- repository: "Národní knihovna" // *jméno instituce s diakritikou a rozlišením velkých a malých písmen*
- shelf_mark: "Osek 72" // *signatura (či jiné označení dokumentu) dokumentu*

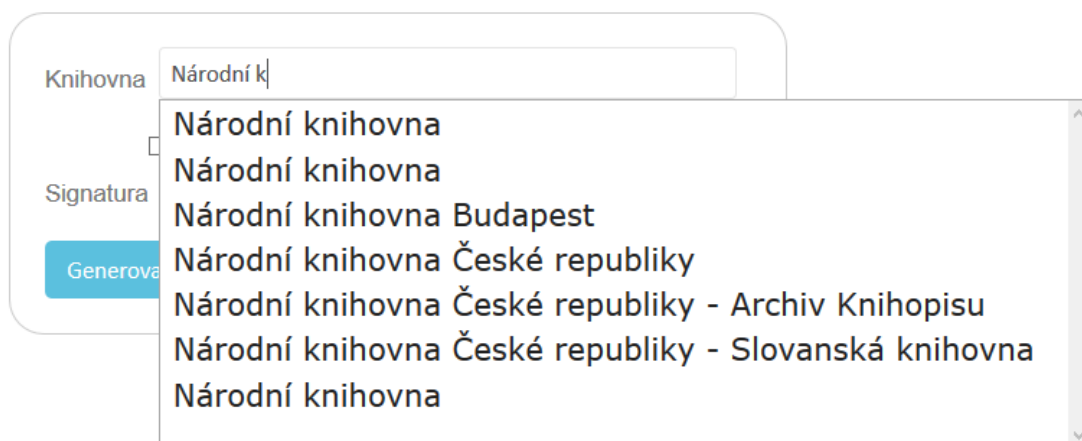
Volání funkce API vrací objekt JSON v následujícím tvaru:

```
{"knihovna": "Národní knihovna", "kodKnihovny": "NKA____", "signatura": "Osek 72", "fyzid": "NKA____OSEK_72____2ZTJY35"}
```

Výsledný identifikátor je uvozen elementem FyzID - "fyzid": "NKA____OSEK_72____2ZTJY35".

Popis uživatelského rozhraní aplikace generátoru

1. Po zobrazení stránky s formulářem generátoru uživatel vyplní položku knihovna. K položce knihovna je implementován našeptávač, který na základě zapsané části názvu instituce dohledává relevantní jména ze seznamu partnerských institucí. Nabídka jmen se zpřesněním redukuje. Našeptávač rozlišuje znaky s diakritikou, nikoliv velká písmena.



2. Dále uživatel vyplní do položky signatura buď signaturu nebo jiné označení odpovídající digitalizovanému dokumentu
3. a klikne na tlačítko Generovat. Je možné si zaškrtnout také volbu „Pamatuj knihovnu pro tuto instanci“. Tato volba usnadní práci při opětovném generování jednoznačného identifikátoru digitálního dokumentu pro více dokumentů z jedné instituce.
4. Aplikace odešle údaje na server, kde proběhne výpočet identifikátoru.

5. Výsledek vypíše do výstupní tabulky na téže stránce.

Repository	Národní knihovna České republiky
Repository code	NKCR__
Shelf-mark	VII.A.18
Identifier	NKCR__VII_A_18__4BZSZD5

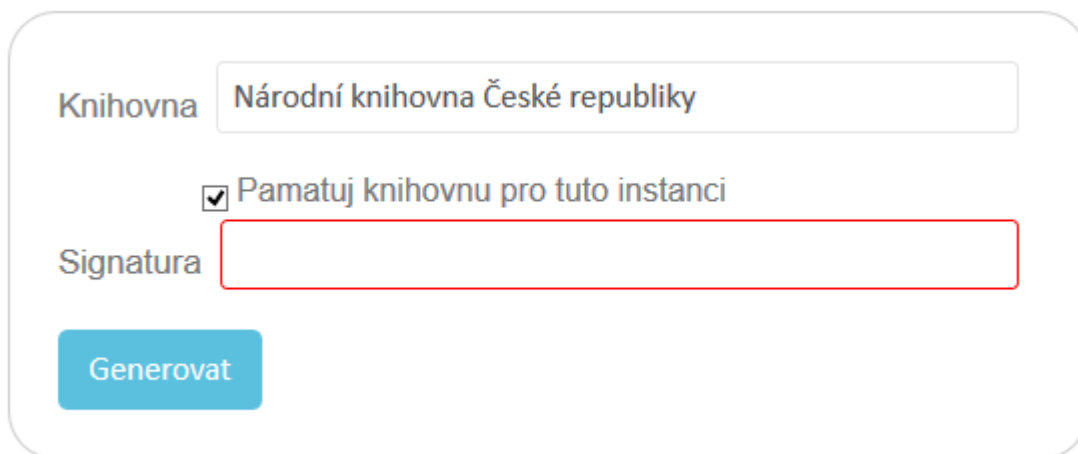
Obrázek 1: Tabulka výsledků při generování identifikátoru místa uložení. Aplikace je součástí MNS a nabízí proto UI také v anglickém jazyce.

Knihovna	Národní knihovna České republiky	Knihovna	Národní knihovna České republiky
☒ Pamatuj knihovnu pro tuto instanci		Kod Knihovny	NKCR__
Signatura		Signatura	VII.A.18
Generovat		Identifikátor místa uložení	NKCR__VII_A_18__4BZSZD5

Obrázek 2: Rozložení obou částí aplikace. Blok s výsledky se zobrazí až po kliknutí na tlačítko Generovat

Chybová hlášení generátoru

Aplikace kontroluje vyplnění vstupních polí. Pokud není jakékoliv pole vyplněné, orámování bloku vstupů se zbarví červeně. V případě chybného vyplnění jména instituce, nebo neexistujícího jména, je uživateli vysláno následující hlášení a výzva: „Nenalezeno v seznamu knihoven, obraťte se prosím na správce“.



Obrázek 3: Červeně se obarvilo pole vstupní položky „Signatura“ poté, co uživatel bez jeho vyplnění klikl na tlačítko generovat. Stejně tak je tomu u výše uvedené položky „Knihovna“.

Závěr

Aplikace byla implementována a testována na vývojových serverech řešitele a následně uvedena v poloprovoz na provozním serveru Manuscriptoria na adrese: <http://www.manuscriptorium.com/cs/node/5090>

V tomto stádiu projektu (beta verze) může s ohledem na vyhodnocování praktických zkušeností uživatelů docházet ke změnám uživatelského rozhraní a dokumentaci projektu. Již při testování byla odhalen jeden problém, který může vyvstat po vygenerování a následném použití jednoznačného identifikátoru digitálního dokumentu, a to přestože bude identifikátor vygenerován naprosto v pořádku na základě zadaných platných údajů. Právě zadání platných údajů, především signatury či jiného označení dokumentu, je v tomto případě kamenem úrazu.

V instituci vlastníci či spravující digitalizovaný dokument může dojít v době mezi vygenerováním jednoznačného identifikátoru pomocí tohoto Generátoru a vložením dokumentu do systému, který na svém vstupu generuje u každého záznamu jednoznačný identifikátor (např. Manuscriptorium nebo Registr digitalizace historických dokumentů), ke změně signatury (označení dokumentu). Může to být kvůli přestěhování a přesignování nebo jen při zpracování záznamu o dokumentu do dokumentografického systému, který vyžaduje jinou formu zápisu signatury (označení dokumentu).

Jako příklad můžeme zvolit staré tisky, které byly zdigitalizovány na digitalizačním pracovišti pod správnou signaturou (správnou pro popisný záznam ze záchranné digitalizace) a byly posléze zkatalogizovány v knihovnickém systému Aleph NKČR pod správnou signaturou (správnou pro knihovnický systém). Signatury těchto dokumentů vyexportované z knihovnického systému (např. 62 A 000127) a z metadat digitalizačního pracoviště (např. 62.A.127) se diametrálně liší a při tvorbě jednoznačného identifikátoru digitálního dokumentu není možné je automatizovaně identifikovat jako označení jednoho a téhož dokumentu. Pokud tedy následně do Manuscriptoria vstupuje záznam s jednoznačným identifikátorem vygenerovaným Generátorem na základě údajů z digitalizace a zároveň záznam s identifikátorem vygenerovaným podle označení z knihovnického systému, zůstanou v MNS jako dva rozdílné exempláře (až do jejich ručního propojení) a v Registru digitalizace historických fondů získají dva rozdílné persistentní identifikátory. V uvedeném příkladu se jednalo pouze o dva záznamy jednoho dokumentu, přičemž jen jeden má digitální kopii, ale u map již zdigitalizovaných v NKČR (některé již třikrát) hrozí vznik tří až pěti různých identifikátorů téhož dokumentu podle toho, jak se měnily signatury a zároveň byla metadata o mapách zapisována v různých formátech a systémech.

VALIDÁTOR BALÍČKŮ VISK 6

Smyslem této úlohy bylo doplnit nový prvek do systému služeb poskytovaných Národní knihovnou v souvislosti s programem VISK 6. Vznikla tak služba, která umožní formální kontrolu balíčku dokumentu VISK 6 před jeho předáním do systémů NKČR. Cílem je usnadnit tvorbu komplexních digitálních dokumentů dle platné definice VISK 6, která je normotvornou pro danou oblast.

Validátor byl na základě zadání řešen jako on-line služba, která umožní zájemcům provádět vzdálenou kontrolu jejich dokumentů. Výsledkem kontroly každého dokumentu je zpráva o formální správnosti s výčtem a popisem případných chyb. Požadavek na zřízení online služby však z kapacitních omezení neumožní nabídnout ke kontrole transportní balíček - tedy balíček, kde je obsaženo více signatur ve složkách UC a MC. Ve verzi poloprovozu se v kontrolovaném souboru ZIP může vyskytovat jen jeden digitální dokument. Maximální velikost kontrolovaného balíčku nesmí překročit 500 MB.

Analýza a specifikace uživatelských požadavků

AiP Beroun řadu let spolupracuje v rámci programu VISK 6 s NKČR na digitalizaci rukopisů a starých tisků. Na základě této mnohaleté spolupráce a praxe vznikly také specifikace požadavků na datové balíčky VISK 6. Všechny softwarové nástroje vyvinuté pro zajištění digitalizační linky AiP Beroun zcela a do detailu dodržují definice VISK 6. Díky této zkušenosti a praxi jsme definovali funkční i nefunkční požadavky na nástroj pro validaci balíčků VISK 6. Zde je jejich souhrn:

- Aplikace musí být dostupná po internetu.
- Interface aplikace bude přístupný pomocí http protokolu prostřednictvím nejpoužívanějších prohlížečů (Internet Explorer, FireFox, Chrome).
- Pro přenos balíčků (komprimovaných souborů ZIP) bude použit File transfer protocol (FTP).
- Balíček ke kontrole obsahuje UC a MC adresáře dle specifikace VISK 6, ale vždy jen pro jeden digitalizovaný dokument. Toto omezení je nutné s ohledem na objemy dat, která by se přenášela na server.
- Pro komprimaci kontrolního balíčku bude očekáván formát ZIP.
- Aplikace bude požadovat minimální počet vstupních hodnot.
- Aplikace nebude uchovávat testované balíčky.
- Výstupem kontroly bude report zobrazený v okně prohlížeče i uložitelný na lokální disk počítače.
- Report bude obsahovat všechny nalezené nedostatky.
- Aplikace bude dostupná jen přihlášenému uživateli Manuscriptoria (MNS), který má přiřazenu roli s oprávněním k užití Validátoru.
- Validátor bude provádět následující druhy kontrol:
 - 1) Je balíček formátu zip?
 - 2) Správnost struktury balíčku - zda obsahuje všechny složky
 - 3) Je v každé složce stejný počet obrázků?
 - 4) Jestli složka UC nebo MC obsahuje správnou strukturu
 - 5) Jestli jsou obrazové soubory ve složkách ve správném formátu
 - 6) Jestli složka MISC obsahuje správné podsložky a XML soubory (v MC navíc techdecs.xml)
 - 7) Jestli složka MISC v MC obsahuje data barevné kalibrace
 - 8) Jestli všechny soubory ve složce MC nebo UC existují a velikostně se shodují se záznamem uvedeným v fixity.md5
 - 9) Jestli všechny obrázky uvedené na konci XML souborů souhlasí, existují a nejsou žádné navíc
 - 10) Jestli jsou všechna XML validní podle přiložených schémat

Očekávaná struktura v ZIP souboru

Jméno_balíčku_VISK6.ZIP

- MC
 - Složka pojmenovaná podle FyzID
 - EX
 - G0
 - MISC
 - fixity.md5
 - další složky na této úrovni nejsou kontrolovány ani očekávány (kapacitní omezení!)
- UC
 - složka pojmenovaná podle FyzID
 - G0
 - MISC
 - N0
 - N1
 - N2
 - P0
 - S0
 - fixity.md5
 - další složky na této úrovni nejsou kontrolovány ani očekávány (kapacitní omezení!)

Popis řešení

Použité technologie

Pro řešení úkolu s ohledem na integraci do systému Manuscriptorium (MNS) byly použity technologie již využívané tímto systémem. Aplikace je vytvořena pomocí PHP (vyžaduje instalaci knihoven FTP a ZIP), HTML, CSS, JavaScript.

Zabezpečení aplikace

Validátor je dostupný pouze přihlášenému uživateli a to jen pokud je uživateli přidělena v systému MNS odpovídající role (role_for_validation). Pracovní prostor (adresář na serveru) pro ukládání nahrávaných balíčků ZIP je nasměrován mimo prostor WWW serveru a tomuto adresáři je přidělen limit na jeho velikost.

Prohlížeče

Aplikace byla testována a optimalizována pro následující prohlížeče a jejich verze:

- Internet Explorer, verze 8 – 11
- Google Chrome, od verze 38
- Mozilla Firefox, od verze 30
- Opera 30
- Safari 5.1

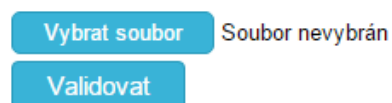
Integrace

Aplikace je vytvořena jako samostatná stránka, ale její užití je možné jen pro přihlášené uživatele systému MNS, kterému byla administrátorem přidělena role s právy na užití validátoru.

Popis užití validátoru a postupu kontroly

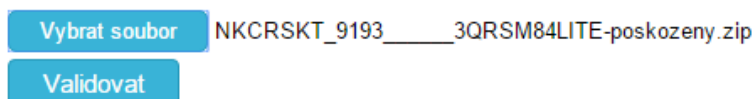
1. Uživatel vybere balíček pro kontrolu kliknutím na tlačítko Vybrat soubor (ve Firefox je popisek tlačítka „Procházet“).

Validátor balíčku VISK6



Obrázek 4: Výchozí stav validátoru

Validátor balíčku VISK6



Obrázek 5: Ovládací prvky validátoru po výběru balíčku.

2. Po kliknutí na tlačítko Validovat je zahájen přenos souboru na server.



Obrázek 6: U horního okraje validátoru se v průběhu nahrávání souboru na server vykresluje červená linka odpovídající průběhu procesu.

3. Uživatel během kontroly a nahrávání balíčku může akci zastavit kliknutím na tlačítko „Přerušit“.
4. V průběhu nahrávání a kontroly je uživatel informován o činnosti validátoru.



Validátor balíčku VISK6

Vybrat soubor

NKCRSKT_9193_____3QRSM84LITE-poskozeny.zip

Přerušit



Kontroluji balíček

Obrázek 7 Průběh kontroly balíčku.

5. Aplikace nahraje soubor do pracovní složky na serveru.
6. Validátor extrahuje všechny soubory z balíčku.
7. Po úspěšném rozbalení smaže původní komprimovaný balíček.
8. Validátor zahájí kontroly extrahovaných souborů.
9. Validátor vypíše uživateli chyby a stav balíčku



Validátor balíčku VISK6

Vybrat soubor

NKCRSKT_9193_____3QRSM84LITE-poskozeny.zip

Validovat

Vynulovat

Stáhnout výsledky kontroly

Balíček není validní

Kontrola MC

Počty obrázků ve složkách nejsou shodné

▶ Kontrola přítomnosti souborů podle fixity.md5

▶ Kontrola Obrázků v XML souboru T_9193_____3QRSM84_CZ.XML

▶ Kontrola Obrázků v XML souboru T_9193_____3QRSM84_EN.XML

▶ TECHDESC.XML

Kontrola UC

Počty obrázků ve složkách nejsou shodné

▶ Kontrola přítomnosti souborů podle fixity.md5

▶ Kontrola Obrázků v XML souboru T_9193_____3QRSM84_CZ.XML

▶ Kontrola Obrázků v XML souboru T_9193_____3QRSM84_EN.XML

Obrázek 8: Výsledek kontroly, po kliknutí na modré šipky se rozbalí detail výsledku dané sekce reportu.

10. Dále má uživatel možnost validátor uvést do výchozího stavu kliknutím na tlačítko „Vynulovat“,
11. nebo si uložit výsledek kontroly kliknutím na tlačítko „Stáhnout výsledky kontroly“

Souhlasí	MISC/Schemas/XML.XSD
Nesouhlasí velikost	MISC/TECHDESC.XML
Souhlasí	MISC/T_9193____3QRSM84_CZ.XML
Souhlasí	MISC/T_9193____3QRSM84_EN.XML
Souhlasí	EX/T_9193____3QRSM84EX0001R.JPG
Souhlasí	EX/T_9193____3QRSM84EX0001V.JPG
Souhlasí	EX/T_9193____3QRSM84EX0002R.JPG
Souhlasí	EX/T_9193____3QRSM84EX0002V.JPG
Souhlasí	EX/T_9193____3QRSM84EX0003R.JPG
Neexistuje	EX/T_9193____3QRSM84EX0003V.JPG
Neexistuje	EX/T_9193____3QRSM84EX0004R.JPG
Neexistuje	EX/T_9193____3QRSM84EX0004V.JPG
Souhlasí	EX/T_9193____3QRSM84EX0005R.JPG
Neexistuje	EX/T_9193____3QRSM84EX0005V.JPG

Obrázek 9 Detail výsledku kontroly.

Balíček není validní

Kontrola MC

Počty obrázků ve složkách nejsou shodné



Kontrola přítomnosti souborů podle fixity.md5

Stav	Soubor
Souhlasí	G0/T_9193____3QRSM84G00001R.JPG
Souhlasí	G0/T_9193____3QRSM84G00001V.JPG
Souhlasí	G0/T_9193____3QRSM84G00002R.JPG
Souhlasí	G0/T_9193____3QRSM84G00002V.JPG
Neexistuje	G0/T_9193____3QRSM84G00003R.JPG
Neexistuje	G0/T_9193____3QRSM84G00003V.JPG
Neexistuje	G0/T_9193____3QRSM84G00004R.JPG

Obrázek 10 Ukázka vygenerovaného a uloženého reportu v HTML.



Validátor balíčku VISK6

Procházet...

NKCRSKT_9193_____3QRSM84LITE-poskozeny.zip

Validovat

Vynulovat

Stáhnout výsledky kontroly

Obrázek 11: Ovládací prvky validátoru se nachází všechny v menu u horního okraje. Pouze v prohlížeči Fire Fox je ignorováno stylování tlačítka pro výběr souboru balíčku.

Očekávaná struktura balíčku

Balíček musí pro zahájení kontroly splňovat tyto podmínky:

1. Balíček je ve formátu ZIP.
2. Balíček ZIP obsahuje správnou strukturu, tedy rozbalená složka obsahuje přímo v sobě složky MC, UC
3. Složka MC obsahuje složku nazvanou shodně s jednoznačným identifikátorem dokumentu
4. Tato složka pak dále obsahuje povinně složky EX, G0, MISC a soubor fixity.md5
5. Složka UC obsahuje složku nazvanou shodně s jednoznačným identifikátorem dokumentu
6. Ta pak obsahuje povinně složky G0, MISC, N0, N1, N2, P0, S0 a soubor fixity.md5

V UC jsou povinné tyto kvality UCI

Parametry odvozených digitálních kopií obrazů pro uživatelské kvalitativní hladiny, optimalizované pro zpřístupnění v Manuscriptoriu a Europeaně	
G0	thumbnail/gallery kvalita (JPG) pro malý náhled, výška 100pix
P0	preview kvalita (JPG) větší náhled, šířka 200 pix
N0	low kvalita (JPG), cca 95DPI, individuálně volená míra komprese, velikost cca 200kB, šířka cca 800pix
N1	normal kvalita (JPG), cca 200 DPI, individuálně volená míra komprese, velikost cca 400kB, s čitelnými i nejmenšími znaky
N2	high kvalita (JPG), 300DPI, uložen ve formátu JPG, (faktor kvality 90%).
S0	bitonální obraz (GIF) černobílý obraz, velikost cca 150kB, rozměry stejné jako N1 s tím, že jsou zachovány detaily ve vysokých jasech i v hlubokých stínech

Provádění kontroly

Dle specifikace programu VISK 6 a výše uvedené analýzy jsou na vybraným balíčkem prováděny tyto kontroly:

- a) Je balíček formátu zip?
- b) Správnost struktury balíčku, zda obsahuje všechny složky
- c) Jestli je v každé složce stejný počet obrázků
- d) Jestli složka UC nebo MC obsahuje správnou strukturu
- e) Jestli jsou obrazové soubory ve složkách ve správném formátu
- f) Jestli složka MISC obsahuje správné podsložky a XML soubory (v MC navíc techdecs.xml)
- g) Jestli složka MISC v MC obsahuje data barevné kalibrace

- h) Jestli všechny soubory ve složce MC nebo UC existují a velikostně se shodují se záznamem uvedeným v fixity.md5
- i) Jestli všechny obrázky uvedené na konci XML souborů souhlasí, existují a nejsou žádné navíc
- j) Jestli jsou všechna XML validní podle příložených schémat

Chybová hlášení aplikace

Z přehledu chybových hlášení aplikace je také patrné, jaké kontroly validátor provádí.

Balíček neobsahuje předepsanou strukturu adresářů - pokud se nepodaří nalézt MC a UC

Balíček není v očekávaném formátu (ZIP) - pokud vybraný soubor není ZIP

Neexistují XML soubory v MISC - pokud nejsou nalezeny očekávané XML

Neexistuje dostatek XML souborů - pokud je málo XML

Neexistují všechny šablony pro XML v MISC/Schemas - chybí šablony ze Schema

Neexistuje soubor TECHDESC.XML - když není nalezen soubor TECHDESC.XML

Počet obrázku se špatným formátem - pokud v kontrolované složce jsou nalezeny neočekávané soubory

Ve složce UC/S0 nejsou obrázky - když nejsou nalezeny obrázky

Počty obrázků ve složkách nejsou shodné - pokud je rozdíl počtu ve složkách UC

UC Není složka nebo neexistuje - když není nalezena složka UC

G0 Není složka nebo neexistuje - když není nalezena složka G0

MISC Není složka nebo neexistuje - když není nalezena složka MISC

N0 Není složka nebo neexistuje - když není nalezena složka N0

N1 Není složka nebo neexistuje - když není nalezena složka N1

N2 Není složka nebo neexistuje - když není nalezena složka N2

P0 Není složka nebo neexistuje - když není nalezena složka P0

S0 Není složka nebo neexistuje - když není nalezena složka S0

MC Není složka nebo neexistuje - když není nalezena složka MC

EX Není složka nebo neexistuje - když není nalezena složka EX

G0 Není složka nebo neexistuje - když není nalezena složka G0

Neexistuje šablona pro validaci TECHDESC.XML - nepovedlo se najít šablonu pro validaci TECHDESC

Neexistuje šablona pro validaci - není šablona pro validaci XML souborů

Nalezeno chyb - počet nalezených chyb

Neexistuje soubor fixity.md5 - chybějící soubor kontrolního součtu

Nesouhlasí počty odkazů uvedené v XML a v složce G0 - odlišné počty obrázků v XML a složce G0

Nebyl vybrán vstupní soubor ZIP - pokud není vybrán zip a potvrdí se validace

Neexistují JPG předlohy pro kalibraci barev - chybějící soubory s kalibračními tabulkami

Neexistuje soubor ICC - chybí ICC profil

Balíček je validní - balíček byl shledán validní

Nepodařilo se zapsat do XML - pokud selže korekce hlavičky TEI

Závěr

Aplikace byla implementována a testována na vývojových serverech řešitele a následně uvedena v poloprovoz na provozním serveru Manuscriptoria na adrese: <http://www.manuscriptorium.com/apps/validatorVisk6/index.php>

Validátor je dostupný pouze po předchozím přihlášení a přidělení práv k užití. Testovací účet byl předán se zdrojovými kódy.

Pro online verzi budou v podstatě existovat vždy limity přenosových rychlostí a výpočetního výkonu serveru. Budou se vyskytovat nároky na řízení současného přístupu více uživatelů a požadavky na dávkové zpracování produkce digitalizačního pracoviště. Proto navrhujeme tuto online verzi ponechat jako kontrolní nástroj pro menší digitalizační projekty, nebo pro počáteční kontroly při nastavení digitalizačních linek. Systémově správné řešení pro velké objemy dat a hlavní produkci balíčků VISK 6 by bylo vytvoření kontrolního modulu ve workflow Národní digitální knihovny.

ZOBRAZOVÁNÍ POZICOVÁNÍ VYHLEDANÝCH SLOV V OBRAZECH DIGITÁLNÍCH DOKUMENTŮ

Cíle řešení

V Manuscriptoriu jsou agregovány komplexní digitální dokumenty relativně mladší provenience, které obsahují také automaticky rozpoznané plné texty (technologie OCR). V drtivé většině se jedná o dokumenty z projektu Národní knihovny České republiky „Hromadná digitalizace historických a vzácných dokumentů ve spolupráci se společností Google“.

Jde o staré tisky různých oborů, tj. kromě církevní literatury jsou to tisky z oblasti přírodních věd (matematika, fyzika, astrologie, astronomie, medicína, botanika, alchymie, počátky chemie), dále právo, geografie, cestopisy, romány, hry, ale i gramatiky, slabikáře, kuchařky, zemědělské příručky apod.“

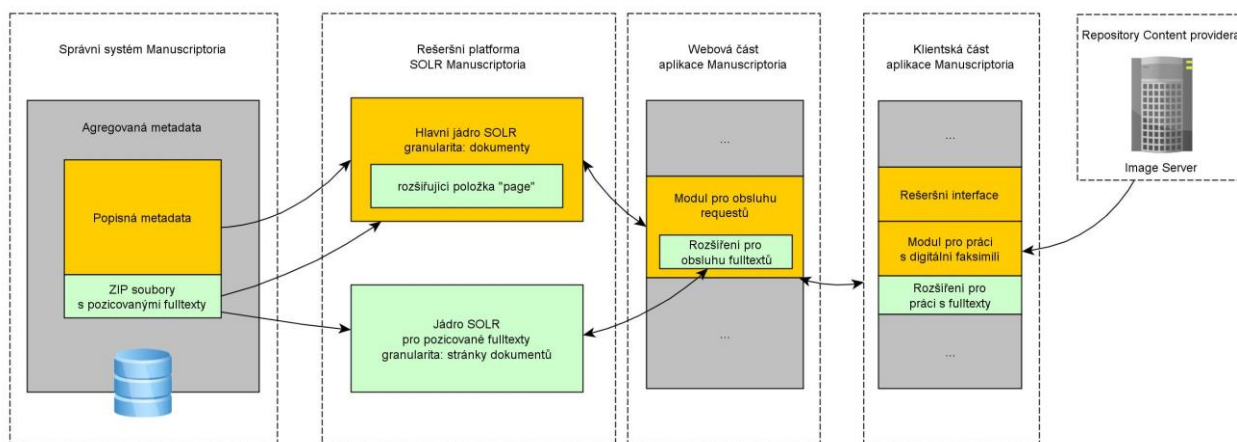
Z hlediska jazykového převažují latinské, německé a české tisky, jsou však zastoupeny i jiné evropské jazyky.

V nejbližší době bude agregováno cca 40 000 takových dokumentů.

Digitální kopie předmětných dokumentů obsahují mimo jiné metadata o pozicích rozpoznávaných slov v obraze stránky. Smyslem řešení úlohy je najít způsob, jak převést daná metadata do vnitřních struktur Manuscriptoria a pracovat s nimi při prezentaci digitální kopie uživateli v koncovém uživatelském rozhraní.

Návrh řešení

Následující schéma popisuje navržené a implementované řešení. Barevně jsou odlišeny součásti systému, které je potřeba nově vytvořit.



Obrázek 12: Schéma řešení

Princip řešení:

- Uživatel zadá v aplikaci rešeršní dotaz, a pokud nespecifikuje konkrétní vyhledávací položku, systém nově prohledá nejen deskriptivní metadata dokumentu, ale i jeho plný text. Ve výsledcích hledání se objeví relevantní dokumenty.
- Následně se systém dotáže nového jádra dedikovaného pro plné texty a zjistí, na kterých stránkách a ve kterém místě obrazu se vyskytují libovolná slova, která tvořila původní dotaz.
- Systém tato slova vhodným způsobem zobrazí.

Rozšíření správního systému Manuscriptoria

Správní systém Manuscriptoria bude rozšířen o podporu ZIP souborů s pozicovanými plnými texty ve formátu HTML ve formě, jak jsou produkovány v projektu *Hromadná digitalizace historických a vzácných dokumentů ve spolupráci se společností Google*.

Při pre-processingu dat a metadat ke zpřístupnění systém bude pracovat s těmito ZIP soubory, extrahovat z nich potřebné informace a vyrábět z nich polotovary pro vstupní rozhraní Rešeršní platformy SOLR Manuscriptoria.

V procesu budou využívány dedikované XSL transformace pro převody mezi formáty, což je vhodné pro usnadnění budoucích modifikací v případě zavádění podpory variantních vstupních formátů.

Rozšíření Rešeršní platformy a Webové části aplikace Manuscriptorium

Rešeršní platforma dosud umožňovala vyhledávat dokumenty dle informací uvedených v deskriptivních metadatech. Za tím účelem systém využívá SOLR jako rešeršní platformu a dosud si vystačil s jediným jádrem.

V souvislosti s podporou plných textů bude toto jádro rozšířeno o položku *page*, ve které je indexován rozpoznaný text jednotlivých stránek. Položka *page* bude indexována do položky „Slova kdekoliv“.

```
<doc>
<field name="dcIdentifier">AIPGGL-NKCR__1C000001T1__178IR86-cs</field>
<field name="dcTitle">Allgemeines Gelehrten-Lexicon, Darinne die Gelehrten aller Stände sowohl männ- als weiblichen Geschlechts, welche vom Anfange der Welt ...</field>
<field name="dcSource">Národní knihovna České republiky</field>
<field name="dcSource_code">NKCR__</field>
<field name="batchTimestamp">1447381981</field>
<field name="dcCreator">Christian Gottlieb Jöcher</field>
<field name="dcContributor">Johann Friedrich, Gleditsch</field>
<field name="dcFormat">codex</field>
<field name="dcDate">1750</field>
<field name="dctermsCreated">1750</field>
<field name="dctermsProvenance">na titulním listu příspěk: Bibliotheca Majori Collegii Academici Soc. JESU Praga ad S. Clementem inscriptus Ao 1758</field>
<field name="dctermsSpatial">Leipzig</field>
<field name="dctermsSpatial">Lipsko</field>
<field name="dctermsSpatial">Německo</field>
<field name="msidno">1 C 000001T1</field>
<field name="thumbs">http://mscripthq.nkp.cz/documentrepository/manuscriptorium_googlebooks/6/NKCR__1C000001T1__178IR86/N2/1C000001T1__178IR86N20001P.JPG</field>
<field name="origDateFrom_sort">1750-01-01T00:00:00Z</field>
<field name="origDateTo_sort">1750-12-31T23:59:59Z</field>
<field name="origDateFrom">1750-01-01T00:00:00Z</field>
<field name="origDateTo">1750-12-31T23:59:59Z</field>
<field name="origDateFrom">1750-01-01T00:00:00Z</field>
<field name="origDateTo">1750-12-31T23:59:59Z</field>
...
<field name="page">1- Vorrede. der Gelehrten-nach alphabetischer Ordnung in einen kurzen Begriff zu bringen und da durch den Liebhabern dieser Arc der hiforifthen Wiffentchaftem die Bemühung zuer liehtern. Es würde das Werck ohne Zweifel gleich ....</field>
<field name="page">Vorrede. für Augen hat ohne die dritte anzufehen in welcher bereits das merke fo ihm an foßig itig geändert und gebeth worden. Es läßt sich diefe Nachlässigkeit defloweni ger entfahldigen da er gedachte dritte ...</field>
<field name="page">B Vorrede. Unwiffenheit in der Lehre von der hiforifthen Glaubwürdigkeit befehludigem wenn ich fo dächte oder frähe. dunckel. fo zweudeutig fo zweifelhafte fo voller Widenruch gegen andere. die einer leh Gefcbichte ...</field>
<field name="page">Vorrede. groffen Sammlungen befinden. Es erparen dergleichen Anmerkungen Mühe und Nachfchlagen. und es find derowegen die großen Wercke forfgaltig durchgegangen worden, damit man andere diefer Mühe ...</field>
...
</doc>
```

Obrázek 13: Ukázka struktury polotovaru pro hlavní jádro

Dále bude zavedeno pomocné jádro fulltextů, jehož granularita bude odpovídat jednotlivým rozpoznávaným stránkám. Toto jádro bude dotazováno při zobrazení detailu záznamu. Pro plnění tohoto jádra budou vznikat polotovary vzniklé zpracováním ZIP souborů s pozicovanými texty ve formátu HTML. Opět budou využity XSL transformace pro větší flexibilitu při provádění budoucích adaptací na nové vstupní formáty.

```
<doc>
  <field name="docId">AIPGGL-NKCR__1C000001T1__178IR86-cs</field>
  <field name="fyzId">NKCR__1C000001T1__178IR86</field>
  <field name="pageNum">00000019</field>
  <field name="pagePos">19</field>
  <field name="pagesCount">1184</field>
  <field name="pageId">AIPGGL-NKCR__1C000001T1__178IR86-cs-00000019</field>
  <field name="page">Vorrede. x I. E WM / ndlih tritt der erlie Band der langfi verprochenen- auch begin-ig enwar teten neuen Ausgabe des Gelehrten-Lena an das Lichn und ich fihreibe * die Vorrede zu ...</field>
  <field name="pageDim">0_0_2324_3036</field>
  <field name="word">534_954_553_970</field>
  <field name="word">Vorrede. 923_911_1326_1042</field>
  <field name="word">x 443_1171_445_1173</field>
  <field name="word">I. 1112_1112_1141_1150</field>
  <field name="word">E 348_1218_390_1273</field>
  <field name="word">WM 346_1386_385_1415</field>
  <field name="word">/ 405_1270_421_1288</field>
  <field name="word">ndlih 481_1222_597_1274</field>
  <field name="word">tritt 615_1225_701_1270</field>
  <field name="word">der 717_1224_784_1270</field>
  <field name="word">erlie 804_1221_894_1280</field>
  <field name="word">Band 909_1222_1042_1271</field>
  <field name="word">der 1062_1225_1130_1271</field>
  <field name="word">langfi 1146_1220_1274_1280</field>
  <field name="word">verprochenen- 1283_1221_1596_1279</field>
  <field name="word">auch 1623_1222_1719_1274</field>
  <field name="word">begin-ig 1737_1222_1901_1282</field>
  <field name="word">enwar 1922_1234_2055_1272</field>
  <field name="word">teten 507_1298_603_1337</field>
  <field name="word">neuen 623_1300_749_1337</field>
  <field name="word">Ausgabe 766_1289_961_1346</field>
  ...
</doc>
```

Obrázek 14: Ukázka struktury polotovaru pro pomocné jádro fulltextů

Dotaz klientské části aplikace Manuscriptorium na pomocné jádro se bude skládat z těchto informací:

- id dokumentu,
- id stránky,
- jednotlivá slova posledního platného dotazu.

Odpověď serverové části bude struktura JSON s informací o stránkách relevantního dokumentu, na kterých se nachází jedno nebo více slov, která tvořila poslední platný dotaz.

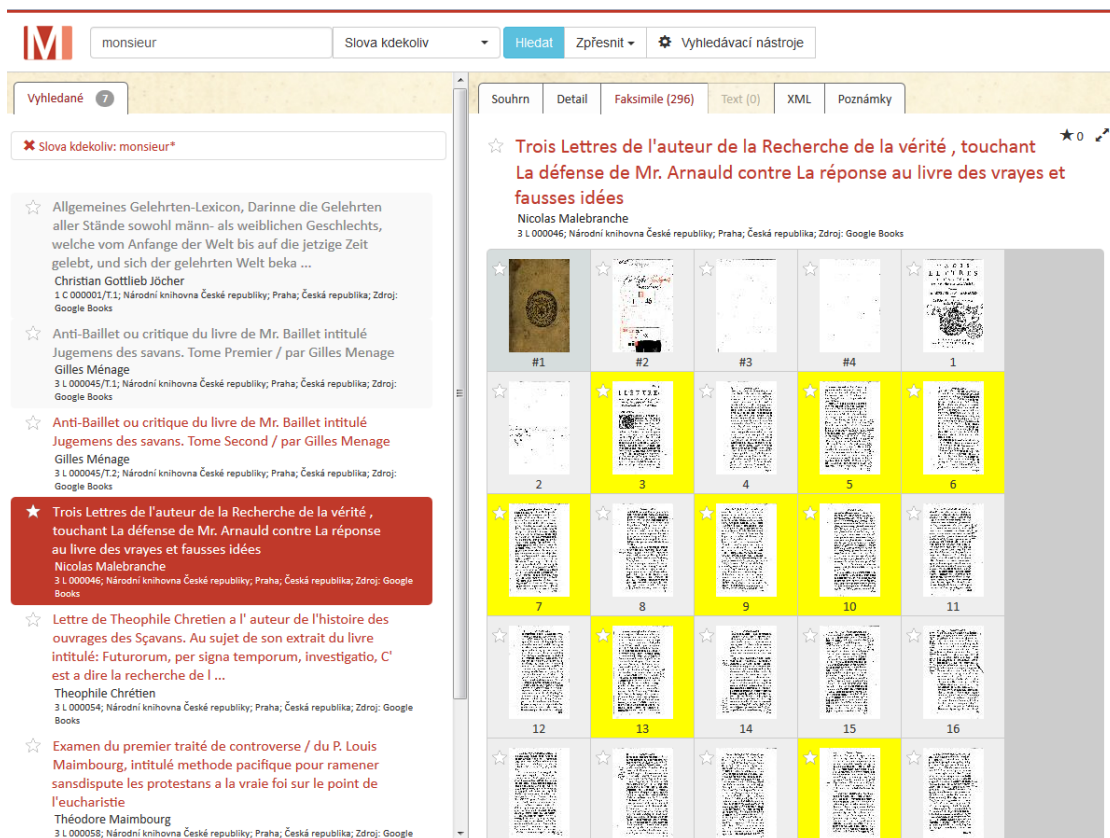
Odpověď bude tvořit seznam stránek, pro každou stránku bude uveden její referenční rozměr (nutné pro správné pozicování) a dále seznam unikátních slov, která odpovídají dotazu a na stránce se vyskytují. Pro každé takové slovo bude konečně uvedena každá pozice, ve které se slovo na stránce vyskytuje. Ukázka zamýšlené struktury odpovědi:

```
"fttH1":{
  "AIPGGL-NKCR__3_L_000046__2CU3JLC-cs-00000007":{
    "pageDim":"0_0_1097_1687",
    "words":{
      "MONSIEUR;":["151_522_382_569", "451_522_782_569"]
    }
  },
  "AIPGGL-NKCR__3_L_000046__2CU3JLC-cs-00000009":{
    "pageDim":"0_0_1097_1687",
    "words":{
      "Ou\u00ef,Mon\u017fieur,":["540_1107_828_1152"]
    }
  }
  ...
}
```

Tato struktura pak bude použita v klientské části aplikace pro zvýraznění daných stránek a pro vykreslení slov. Vyhledané stránky jsou zvýrazněny v galerii stránek faksimile a jednotlivá slova jsou dále zobrazena v každé otevřené stránce.

Implementace

Nástroj je v době vzniku tohoto dokumentu dostupný v režimu poloprovozu na <http://dbase.aipberoun.cz/manu3/apps>. Bude převeden do ostrého režimu v při update Manuscriptoria na verzi 3.1, která je plánována na přelom roku 2015/2016.

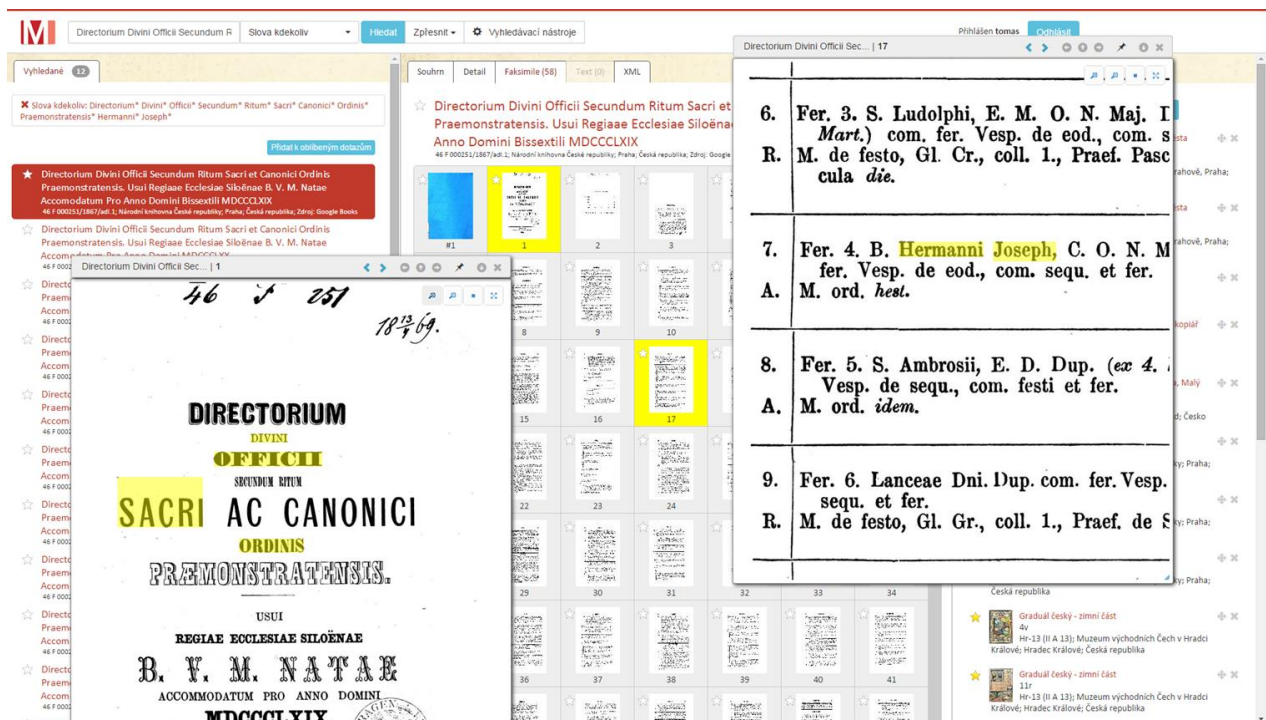


The screenshot displays the Manuscriptoria search interface. At the top, there is a search bar with the word "monsieur" entered. Below the search bar, there are tabs for "Vyhledané" (7), "Souhrn", "Detail", "Faksimile (296)", "Text (0)", "XML", and "Poznámky". The search results are listed on the left, and a grid of document thumbnails is shown on the right. The grid contains 16 thumbnails, with some highlighted in yellow. The highlighted thumbnails correspond to the search results listed on the left.

Search Results:

- ☆ **Trois Lettres de l'auteur de la Recherche de la vérité , touchant La défense de Mr. Arnauld contre La réponse au livre des vrayes et fausses idées**
Nicolas Malebranche
3 L 000046; Národní knihovna České republiky; Praha; Česká republika; Zdroj: Google Books
- ☆ **Anti-Baillet ou critique du livre de Mr. Baillet intitulé Jugemens des savans. Tome Premier / par Gilles Menage**
Gilles Menage
3 L 000045/T.1; Národní knihovna České republiky; Praha; Česká republika; Zdroj: Google Books
- ☆ **Anti-Baillet ou critique du livre de Mr. Baillet intitulé Jugemens des savans. Tome Second / par Gilles Menage**
Gilles Menage
3 L 000045/T.2; Národní knihovna České republiky; Praha; Česká republika; Zdroj: Google Books
- ☆ **Trois Lettres de l'auteur de la Recherche de la vérité , touchant La défense de Mr. Arnauld contre La réponse au livre des vrayes et fausses idées**
Nicolas Malebranche
3 L 000046; Národní knihovna České republiky; Praha; Česká republika; Zdroj: Google Books
- ☆ **Lettre de Theophile Chretien a l' auteur de l'histoire des ouvrages des Sçavans. Au sujet de son extrait du livre intitulé: Futurorum, per signa temporum, investigatio, C' est a dire la recherche de l ...**
Theophile Chretien
3 L 000054; Národní knihovna České republiky; Praha; Česká republika; Zdroj: Google Books
- ☆ **Examen du premier traité de controverse / du P. Louis Maimbourg, intitulé methode pacifique pour ramener sansdispute les protestans a la vraie foi sur le point de l'eucharistie**
Théodore Maimbourg
3 L 000058; Národní knihovna České republiky; Praha; Česká republika; Zdroj: Google Books

Obrázek 15: Výsledky vyhledávání s otevřeným dokumentem a zvýrazněnými stránkami, na kterých se vyskytuje slovo z dotazu



Obrázek 16: Otevřené obrazy stránek se zvýrazněnými nalezenými slovy

Závěr

Podle našeho názoru je vhodné pokračovat ve vývoji práce s textovým obsahem stránky a to zejména v souvislosti s implementací vlastního Image serveru a zavedením podpory IIIF v systému Manuscriptoria tak, aby bylo možno v systému pracovat přímo s výstřižky a tyto výstřižky perzistentně odkazovat (viz funkcionality Google Books).

Propojení katalogu NKČR a MNS

Cíle řešení

V Manuscriptoriu jsou agregovány komplexní digitální dokumenty vytvářené a spravované různými institucemi. Jedním z nejvýznamnějších přispěvatelů je Národní knihovna České republiky (NKČR). Katalogizační systém NKČR má dvě dílčí složky: knihovni informační systém Aleph a správní systém Oddělení rukopisů a starých tisků pro dokumenty TEI (SSORST). Cílem je umožnit propojení všech systémů tak, aby bylo možno volně procházet mezi katalogizačním systémem NKČR objednatelů a uživatelským prostředím Manuscriptoria a tak usnadnit využívání digitalizovaných dokumentů uživateli katalogu NKČR. Výsledkem řešení úlohy má být předpis pro konstrukci přímých odkazů - do Manuscriptoria bude doplněno vhodné rozhraní umožňující přímé odkazování na základě URL generovaných v katalogu NKČR a v něm dostupných metadat.

Návrh řešení

Národní knihovna provozuje systém Aleph sází STT (Staré tisky). Obsah databáze Manuscriptoria a báze STT se do určité míry překrývá. V STT jsou katalogizovány všechny dokumenty a systém obsahuje jejich popisná metadata (MARC záznamy). V Manuscriptoriu jsou záznamy často o stejných dokumentech.

Každý systém pracuje s vlastními systémovými čísly (a jinými kontrolními čísly). Na počátku konzultací s pracovníky NKČR se nabízelo řešení propojit systémy jednoduše tak, že Manuscriptorium umožní přístup k záznamu či dokumentu na základě systémového čísla mateřského zdroje záznamu či dokumentu.

Ukázalo se, že uživatelským požadavkem ve skutečnosti není propojit rodičovský záznam v uživatelském prostředí STT s agregovaným záznamem v uživatelském prostředí Manuscriptoria. Jak vyplynulo z konzultací, cílem úlohy je totiž umožnit uživateli STT přímý vstup do existující digitální kopie, pokud tato v Manuscriptoriu existuje.

Z toho vyplynuly tyto dílčí úkoly k řešení:

- Digitální kopie dokumentů jsou do Manuscriptoria ingestovány z různých zdrojů, tyto zdroje nemusí znát (a v praxi skutečně neznají) systémové číslo STT (případ inkunábulí a starých tisků pořízených v minulosti v rámci VISK 6). Propojení pomocí tohoto čísla tedy není možné.
- Obohacování agregovaných záznamů o systémové číslo STT je problematické: nelze tak činit na straně Manuscriptoria (správcem komplexního digitálního dokumentu (*KDD*) není Manuscriptorium), obohacení by mělo proběhnout nad *primárními metadaty* (tj. u zdroje dokumentu) a to správcem obsahu.

Což by v případě NKČR bylo nyní problematické s ohledem na to, že produkce VISK 6 není (pokud je nám známo) spravována v rámci systémů NKČR jinak, než že KDD jsou umístovány a organizovány v souborové databázi Centrálního datového úložiště a manipulace s nimi je proto poměrně problematická.

- I kdyby bylo propojení realizovatelné, je potřeba řešit potřebu propojovat takto jen dokumenty, které mají v systému Manuscriptorium digitální kopii. Báze STT přitom takovou informaci nemá.

Z výše uvedeného vyplynulo a domluveno bylo následující řešení technicky akceptovatelné oběma systémy:

1. Odkazy na digitální kopii dokumentu budou obohaceny bibliografické záznamy STT.
2. Odkaz na digitální kopii dokumentu získá STT těžením Manuscriptoria pomocí OAI-PMH.
3. Aby byly na straně STT takto obohaceny jen dokumenty, které skutečně mají digitální kopii, bude na straně Manuscriptoria vyroben speciální set dokumentů „digitalizované dokumenty, jejichž fyzické exempláře spravuje NKČR“
4. Párování záznamů STT a natěžených odkazů z Manuscriptoria proběhne na základě signatur, které v natěženém setu představují jednoznačnou párovací informaci.

Implementace

Na straně Manuscriptoria byla implementována možnost otevírat dokumenty přímým odkazem (pomocí předaného identifikátoru).

Dále byl zaveden nový metadatový formát, který předává jen elementární informace.

Nakonec byl zaveden set, do kterého jsou při pre-procesingu agregovaných informací zařazeny jen dokumenty umístěné v Národní knihovně české republiky, které současně mají v Manuscriptoriu prezentovanou digitální kopii.

Přístup k výsledkům řešené úlohy

Název setu	CZ-PrNK
Název metadatového prefixu	czprnk
URL pro těžení	http://www.manuscriptorium.com/oai/?verb=Identify
Příklad konkrétní odezvy	http://www.manuscriptorium.com/oai/?verb=ListRecords&metadataPrefix=czprnk&set=CZ-PrNK
Přímý link na dokument z příkladu odezvy	http://www.manuscriptorium.com/apps/index.php?direct=record&pid=AIPDIG-NKCR_12_J_237_3XAEKLE-cs

KONVERTOR EOD (EBOOKS ON DEMAND)

Cílem je vytvoření desktopového softwarového nástroje, který bude zajišťovat import produkce oddělení mikrografie a digitalizace (odbor Historických a hudebních fondů) a jeho nového skeneru PS7000C (na kterém jsou typicky digitalizovány bohemikální staré tisky - digitalizace místo mikrofilmování a staré tisky - pro EOD). Tuto produkci je třeba transformovat dle specifikace VISK 6 pro uložení do úložiště NK (CDÚ) a zpřístupnění v Manuscriptoriu (MNS).

Specifikace funkčních požadavků

Specifikace požadavků vychází z definice standardů programu VISK 6, dlouholeté zkušenosti na digitalizačním pracovišti AiP Beroun a v neposlední řadě z požadavků pracovníků objednatelů.

Přehled požadavků:

- dávkové zpracování - aplikace bere za zdroj složku, která obsahuje podadresáře (pro každý dokument je jeden podadresář)
- generování odvozené kvality obrazů (z vstupní kvality (např. N3, nebo EX > N2 > G0)) podporované kvality jsou: N0, N1, N2, P0, S0, G0
- V každém adresáři naskenovaného vstupního dokumentu konvertor očekává:
 - Podadresář N3 s obrázky
 - Adresář MISC, kde se nachází soubor popisu s dohodnutým stálým jménem (jméno je nedůležité, neboť tam nic jiného nebude – např.: POPIS.ISO) se strukturou ISO2709.
- generování strukturálních meta-dat v XML ve formátu TEI P5
- generování výstupních balíčků pro dvě místa (MC a UC) – sdílené síťové disky, definované cestou souborového systému
- generování „předávacích“ log-souborů jako .csv, které obsahují tyto položky: "InputId"; "DocId"; "DocName"; "Signatura"; "Pocet medii"; "Pocet obrazu"; "Typ"
- jednoduché uživatelské grafické rozhraní s minimem ovládacích prvků
- maximální automatizovanost procesu zpracování
- aplikace pro MS Windows

Popis vstupních a výstupních dat

Vstupem je adresář, kde se nachází složky naskenovaných knih, může být umístěn libovolně, cestu k němu si nastavuje obsluha v prostředí aplikace.

Katalogizační popis ve formátu MARC ISO2709 nebo XML MARC dodává NK do podadresáře /MISC, kde je očekáván konvertorem.

Výstupem jsou dva balíčky (MC a UC) dle specifikace VISK6 a soubory XML TEI P5 pro MNS

- MC - balíček TAR obsahuje (hlavní složka a také FyzId) v ní: N3 (beze změn), G0, MISC, fixity.md5
- UC - (hlavní složka a také FyzId) v ní: N2 (300dpi, kvalita uložení 90% jpg), G0 (výška 100pix, kvalita uložení 90% jpg), MISC, fixity.md5

Očekávané vstupy z oddělení mikrofotografie a digitalizace (odbor Historických a hudebních fondů) NKČR

Parametry výstupu ze zařízení PS7000C

- knihy budou digitalizovány tak, aby výstup byl použitelný pro EOD a zároveň aby vzniklá data měla potenciál být zpřístupněna v Manuscriptoriu.
- vždy bude digitalizována celá kniha (svazek) i v případě, že pro EOD je třeba např. jen jedno dílo z konvolutu
- v procesu digitalizace jsou vždy pořizovány snímky (digitální obrazové soubory) všech součástí knižní předlohy: vazba, desky, přideštiny, všechny listy (viz tabulka *Identifikace jednotlivého obrazu*)
- pro každou stránku originálního dokumentu vznikne jeden digitální obraz, který zachycuje i okraj listu, ořez digitálního obrazu bude tedy prováděn vně listu

Katalogizační popis

- fotooddělení obdrží spolu s knihou ke skenování také katalogizační popis z databáze STT ve formátu MARC21, který bude vyexportován ze systému Aleph jako soubor ISO2709
- tento soubor s katalogizačním popisem bude vložen do adresáře naskenovaného dokumentu do podadresáře MISC

Např.: signatura I.A1 -> název adresáře I_A_1

podadresář s obrazovými soubory: I_A_1\N3

podadresář s katalogizačním popisem: I_A_1\MISC

Technické parametry obrazových dat

Základní parametry digitálních obrazů pro nejvyšší výstupní kvalitu:

- barevnost (RGB),
- rozlišení 360 DPI (optické rozlišení zařízení)
- neinterpolovaný obraz,
- barevná hloubka 24 bit,
- uloženo JPG (volba FINE na daném zařízení)

Identifikace jednotlivého obrazu

Jméno obrazu je vytvořeno dle výše uvedených pravidel, přičemž pět znaků identifikujících stránku (FFFFF) musí být generováno dle těchto pravidel:

Typ předlohy obrazu (česky)	Typ předlohy obrazu (anglicky)	Jméno souboru ze scanneru	jména obrazového souboru (příklad)
běžný list – foliace		nnnnn.JPG	0001R.JPG, 0001V.JPG
běžný list – paginace			0001.JPG, 0002.JPG
Vložený list	Enclosed Sheet	ESnnn.JPG	ES01V.JPG
Zpevňovací proužek	Reinforcing Strip	RSnnn.JPG	RS01V.JPG
Hřbet	Spine	SP.JPG	SP.JPG
Horní ořízka	Head Edge	HE.JPG	HE.JPG
Boční ořízka	Side Edge	SE.JPG	SE.JPG
Dolní ořízka	Bottom Edge	BE.JPG	BE.JPG
Přední desky	Front Cover	FC.JPG	FC.JPG
Přední přidešť	Front End-Sheet	FS.JPG	FS.JPG
Zadní desky	Back Cover	BC.JPG	BC.JPG
Zadní přidešť	Back End-Sheet	BS.JPG	BS.JPG
Římské číslování přední – foliace	Front Roman Page	Frrrr.JPG	F001R.JPG, F001V.JPG
Římské číslování přední - paginace			F001.JPG, F002.JPG
Římské číslování zadní – foliace	Back Roman Page	Brrrr.JPG	B001R.JPG, B001V.JPG
Římské číslování zadní – paginace			B01.JPG, B002.JPG

kde

nnnnn

v případě paginace

0001 ... 9999

v případě foliace

0001R ... 9999R a 0001V ... 9999V

rrrr

je třímístné číslo převedené z původních římských číslic
plus znak R či V u foliace

Struktura názvů souborů a složek

Hlavní složka: adresář pojmenován dle signatury knihy (čárky, lomítka, či jiné oddělovače nahradit _)

Např.: signatura I.A1 -> název adresáře I_A_1

podadresář s obrazovými soubory: I_A_1\N3

podadresář s katalogizačním popisem: I_A_1\MISC

Přílohy

Popis programu EOD Konvertor, AiP Beroun s.r.o., 2015

Literatura

- [1] kolektiv: Definice digitálního dokumentu pro potřeby zpřístupnění a trvalého uložení v podprogramu VISK6, AiP Beroun, 2014
- [2] ing. František Šibrava a kol.: Uložení digitálních kopií rukopisů a starých tisků na datovém úložišti Národní knihovny ČR, Zpráva k Dodatku č. 8 ke Smlouvě o spolupráci ve výzkumu a vývoji, AiP Beroun, 2010
- [3] Jiří Vrbický: Vytváření a kontrola MD5 součtů, NK ČR, 2010
- [4] ing. František Šibrava: Popis programu M_Copy, AiP Beroun, 2012
- [5] ing. František Šibrava: MANUSCRIPTORIUM - Program M-Gen2 - Technická dokumentace, AiP Beroun, 2012