



Národní digitální knihovna

... v cílové rovině

Luděk Rašek (ludek.rasek@cgi.com)

7.10.2014



EVROPSKÁ UNIE
EVROPSKÝ FOND PRO REGIONÁLNÍ ROZVOJ
ŠANCE PRO VÁŠ ROZVOJ



Tento projekt je spolufinancován ze strukturálních fondů EU (Evropského fondu pro regionální rozvoj) prostřednictvím IOP

Agenda

Seznámení s projektem

Digitalizace

Dlouhodobé uložení

Zpřístupnění

Přehled projektu

Cíle

- Digitalizace
 - Vytvoření linky pro masovou digitalizaci knih
 - Digitalizace 26 000 000 stran za 2 roky provozu (cca 54 000 stran denně)
- Vytvořit dlouhodobé úložiště (LTP) dle modelu OAIS
 - Uložit digitalizovaná data
 - Příprava LTP na certifikaci
- Zpřístupnění dat
 - Export digitalizovaných dat do digitálních knihoven (Kramerius)
 - Indexace dat a vytvoření vyhledávacího portálu (Vufind)

Harmonogram

- polovina 2011 – výběrové řízení
- 2012 – implementace systému, 2 měsíce pilotního provozu
- 2013, 2014 – běžný provoz, stabilní denní výkon více než 54 000 stran (~220knih)

Organizace

- CGI – systémový integrátor
- 2 digitalizační pracoviště – Praha (NK), Brno (MZK)
- Jedno datové centrum (Praha)

Digitalizace



CGI

Experience the commitment®

Digitalizační proces – řízený dig. workflow

Řezání

- Rozřezání kníží vazby pro digitalizaci na dokumentovém skeneru

Příprava knih

- Přejímka knih určených k digitalizaci
- Výběr knih schopných digitalizace
- Vkládání knihy do digitalizačního workflow
- Skenování obálek (Praha)

Skenování

- Provedení digitalizace
- 6 druhů skenerů
- Výběr skeneru dle vhodnosti pro danou knihu
- Možnost skenovat knihu na více skenerech

Práce s fyzickou knihou

Zpracování obrazu

- Otáčení
- Rozdělování stránek
- Ořez stránek
- Primárně využíván Scantailor
- Možno využít libovolného nástroje

Editace metadat

- Revize výsledku ořezu a následných automatických kroků
- Zadání struktury knihy
- Rozdělování periodik
- Doplnění typu strany
- Úprava metadat

Výstupní kontrola

- Finální kontrola celé entity
- Kontrola kvality obrazu
- Kontrola metadat
- Finální schválení

Práce s digitálními daty

Skenování - hardware

Prague

- Full auto
 - 1x 4DigitaBooks DL Mini
 - 1x 4DigitalBooks DL3003 (velké formáty)
 - 2x Treventus Scan Robot 3 (včetně plochého skeneru)
 - 1x Canon DR-X10C – document scanner
- Semi auto
 - 1x 4DigitalBooks Scan2Page (velké formáty)

Brno

- Full auto
 - 1x 4DigitaBooks DL Mini
 - 1x 4DigitalBooks DL3003 (large formats)
 - 2x Treventus Scan Robot 3 (includes flatbed scanner for cover scanning)
 - 1x Canon DR-X10C – document scanner
- Semi auto
 - 1x 4DigitalBooks Scan2Page (large formats)
- Manual
 - 1x i2s-digibook suprascan (velmi velké formáty)

Manuální zpracování

- **Zpracování obrazů**
 - Ořez, natočení, kontrola výstupu skeneru
 - Scantailor – opensource upravený podle požadavků NDK
 - Možno využívat/integrovat libovolný SW
- **Editace metadat**
 - Kontrola výstupů
 - zpracování obrazu
 - automatického generování obrazů
 - Doplnění strukturálních metadat (typy stran, čísla stran)
 - Dělení periodik na ročníky a čísla
- **Manuální OCR**
 - V případě vysoké chybovosti OCR je možno v tomto kroku provést revizi
- **Kontrola**
 - Finální kontrola kvality

Automatické činnosti zpracování

- Prováděny univerzálním Transformačním modulem
 - Flexibilní výkonná komponenta, která řídí masivně paralelní zpracování dat digitalizace
 - cca 4 mil. operací denně, ~20TB/den zpracovaných dat
- Mezi jednotlivými manuálními kroky probíhá automatické zpracování
 - Získání popisných metadat z Aleph
 - Předzpracování pro zpracování obrazů
 - Provedení operací po zpracování obrazů
 - Generování výstupních formátů
 - Charakterizace obrazů na vstupu i výstupu
 - Průběžná evidence prováděných kroků pro záznam do Premis
 - Dělení svazků na jednotlivé výstupní entity (typicky čísla periodik)
 - Tvorba výstupních formátů – METS, Premis, MIX



Integrace s dalšími systémy

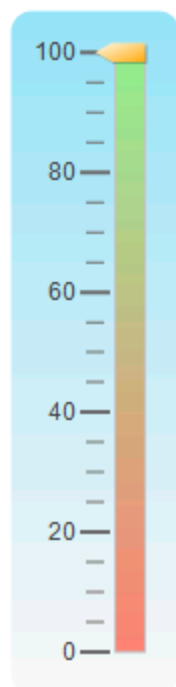
- Následující systémy jsou plně integrovány
 - Registr digitalizace – aktualizace stavu
 - URB NBN Registr – získání a registrace identifikátorů
 - Aleph – získání metadat a aktualizace záznamů
 - Kramerius 4 NK – import digitalizovaných dat
 - Kramerius 4 MZK – import digitalizovaných dat
 - LTP – import SIP balíků pro dlouhodobé uložení



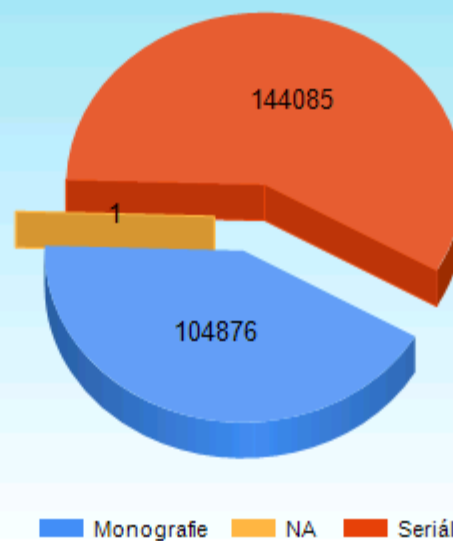
Výkonnost systému

- Denní nominální výkon linky
 - ~ 220 svazků
 - ~ 54 000 stran
 - ~ 1.35TB vstupních dat
 - nekomprimovaný TIFF
 - cca 25MB/stránku
- Proudové zpracování
 - Průchod svazku linkou
 - cca 3 dny probíhá zpracování zahrnující lidské zásahy
 - do 10 dnů od vstupu do linky je kniha uložena v LTP a publikována v Kramerovi





KB3 Celkový počet zdigitalizovaných entit (Celkem)



Objekty čekající ve frontě na zpracování 1, 3, 5, 7 a nad sedm dnů

KB4 - Postup zpracování (aktuální stav k datu 26.11.2014 7:55:09)

Fáze	Počet svazků/entit	
Příprava (svazek)	413	10 129 43 231
Skenování (svazek)	905	26 325 3 49 502
Postprocessing (svazek)	675	9 200 12 49 405
Editace metadat (svazek)	360	23 151 28 39 119
Transformační modul (svazek)	154	86 47 8 6 7
Kontrola (svazek)	128	5 102 6 15
Zpracování IE (entita)	15077	15 11 7420 7613
Ukončování IE (entita)	774	158 476 104 5 31
Digitalizováno (entita)	248806	1282 1798 876 244071
Ukončování (svazek)	594	85 319 172 7

Rozložení zpracování po skenerech

Skener	Svazků	Celkem	Napoprvé	Opakovaně	Resken	Dosken
DL3003	7488	2682970	2658339	24631	13203	11428
DL Mini	9683	2645416	2607016	38400	29372	9028
Canon DR10	58822	12128995	11999652	129343	76159	53184
Suprascan A0	13205	36731	33265	3466	2873	593
Plustec	72804	1825795	1775836	49959	25931	24028
Scan2Page	5937	540873	494509	46364	19496	26868
Treventus	20708	5895356	5828757	66599	43008	23591
Virtuální	959	406711	386611	20100	19117	983



Řídící a dohledové centrum

Aktuální stav digitalizace

Lokalita	Skutečnost		Plán		Cíl	
	svazků	stránek	svazků	stránek	svazků	stránek
MZK	48920	12166968	48143	12432540	50340	13000000
Národní knihovna ČR	52502	13045847	48143	12432540	50340	13000000
Celkem	101433	25223771	96285	24865079	100680	26000000

K datu 26.11.2014 08:03 zdigitalizováno: 25 223 771 stránek.

Zbývá zdigitalizovat: 776 229 stránek.

1.0.4707.32332

Stav plnění cílů

- Počty svazků
 - Na konci listopadu bylo dosaženo cílového indikátoru v počtu naskenovaných svazků
 - Lze předpokládat, že do konce roku bude dosaženo více než 100%

Již nyní je možné prohlásit digitalizační část projektu za úspěšnou !



Dlouhodobé úložiště LTP



CGI

Experience the commitment®

Long term preservation archive

- Navrženo jako homogenní úložiště pro objekty z oblasti knihovnictví
 - Objekty s dlouhodobou ochranou informací
 - Data z digitalizace (METS+JPEG2000+OCR)
 - Nově ePub a PDF z projektu NAKI
 - Objekty pouze s ochrana dat
 - Data z webarchivu
 - Jiná data
- Samostatný systém budovaný v souladu s OAIS
 - Příjem informací
 - Dlouhodobá ochrana informací
 - Zpřístupnění informací



Architektura systému

- **SAFE LTP**
 - Flexibilní indexace uložených dat
 - Správa procesů dle OAIS
 - Samostatné univerzální workflow pro interní operace
 - Volná integrace s digitalizačním workflow a digitálními knihovnami
 - Volná vazba na objektové úložiště dat
- **Transformační modul**
 - Samostatná instance
 - Řízen prostřednictvím LTP workflow
 - Vysoká míra škálovatelnosti a otevřenost změnám
- **Objektové úložiště dat**
 - IBM Information Archive
 - Čistá dostupná kapacita cca 650TB na LTO páskách
 - Možno rozšiřovat až o 300% dodáním dalších pásek
 - Geografická distribuce Hostivař (robot)-Klementinum(robot)-Brno(jen pásky)



Aktuální stav

- Rutinní provoz vstupu dat
 - z digitalizace
 - z webarchiv
- Rutinní správa systému a běh procesů ochrany
 - Ověřování čitelnosti
 - Opravy metadat z první fáze digitalizace
- Dokončování připravenosti na certifikaci
 - Závislost na procesech správy a provozu ICT organizace (NK), které dopadají na LTP v oblasti správy infrastruktury
 - Finalizace dokumentace potřebné pro certifikaci



Zpřístupnění



CGI

Experience the commitment®

Architektura

- Vufind
 - Frontend, který poskytuje vyhledávání ve všech podporovaných zdrojích
 - Nezpřístupňuje přímo obrazy
 - Digitalizace je publikována skrze Kramerius
- Backend
 - SOLR full text index
 - Infrastruktura pro sklízení různých zdrojů
 - Digitalizace
 - Aleph NK/MZK
 - Manuscriptorium
 - WebArchiv
 - Publikace obsahu pomocí OAI (Europeana)

Aktuální stav

- V současné době probíhá finální testování a příprava na pilotní provoz a následně otevření pro přístup veřejnosti



Otázky & odpovědi



CGI

Experience the commitment®

Our commitment to you

We approach every engagement with one objective in mind: to help clients succeed



Luděk Rašek
ludek.rasek@cgi.com

CGI

Experience the commitment®

Záložní slajdy



CGI

Experience the commitment®

Rozřezání knih

- Používá se pro knihy, kde existuje dostatek kopií
- Po rozřezání jsou desky a listy volné
- Volné listy je možné skenovat vysokou rychlostí na dokumentovém skeneru
- Dokumentový skener
 - Je nejrychlejším digitalizačním zařízením v lince (10 0
 - Poskytuje data, která zpravidla není nutné dále upravovat
- Po digitalizaci jsou rozřezané knihy likvidovány jako odpad
- **Zkušenosti**
 - Výkon digitalizačního pracoviště je natolik vysoký, že v rutinním provozu je třeba dobře plánovat tak, aby byl dostatek knih pro destruktivní digitalizaci



Příprava knih

- Knihy jsou vypůjčeny z běžných nebo konzervačních fondů (s výjimkou knih pro destrukci)
- Knihy jsou prověřeny
 - Fyzický stav - rozhodnutí, zda jsou nebo nejsou vhodné pro digitalizaci
 - Ověření metadat
 - Ověření stavu v registru digitalizace pro koordinaci s ostatními digitalizacemi v ČR
- Když v kniha vyhovuje, je zavedena do digitalizačního workflow
- Knihy jsou zpracovávány po dávkách (zefektivnění při předávání)
- Vybavení pracoviště:
 - Praha
 - 4 pracovníci (zajišťují rovněž skenování obálek a desek)
 - 4 PC + 2 ploche skenery
 - Brno
 - 1 pracovník (pouze zavedení kniho do workflow)
 - 1 PC, bez skeneru

Skenování software

- Skenování je řízeno nativním SW každého skeneru
- Integrace do digitalizačního workflow
 - Skenování řídí SW skeneru
 - Integrační komponenta – Logica Skenovací Aplikace (LSA)
- LSA integruje
 - Specializováno na obsluhu skeneru a řízení skenování
 - Zobrazuje data z digitalizačního workflow
 - Slouží k zadávání údajů do workflow
 - Zajišťuje spuštění přenosu dat z lokálního disku do datového centra, kde se obrazy zpracovávají

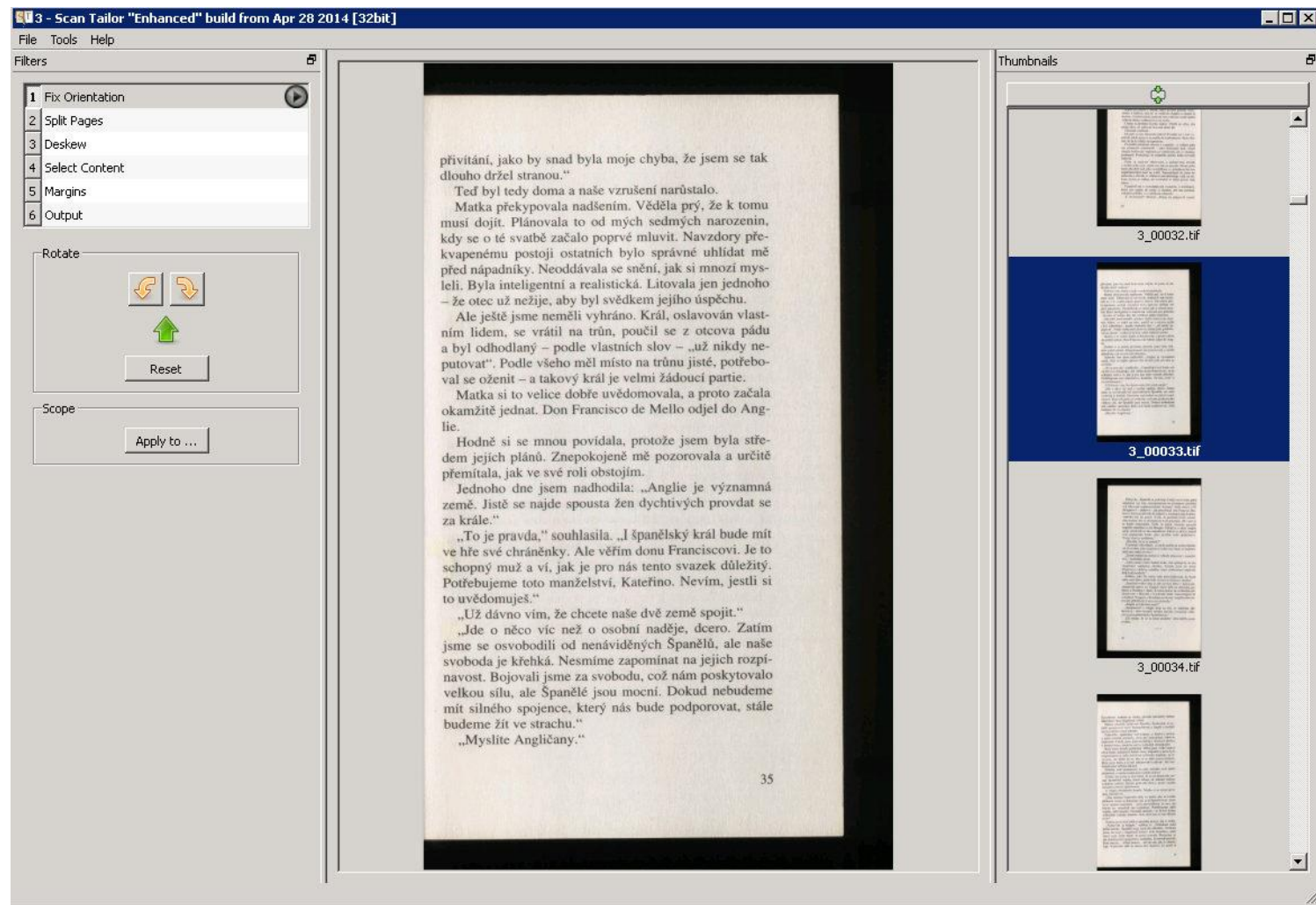


Zpracování obrazu

- Obrazy jsou automatizovaně předzpracovány
- Manuální & poloautomatické opravy
- Manuální zpracování v případě, kdy selže automatizmus
- Vybavení pracoviště
 - PC, které přistupuje na centrální farmu VDI – virtualizované stanice pro kompletní zpracování dat
 - Stanice (fyzické i virtuální) jsou sdíleny s pracovištěm editace metadat
- Software
 - Hlavním nástrojem je Scantailor
 - Založeno na opensource
 - Úpravy provedené CGI v rámci projektu pro lepší fungování dle požadavků NDK



Scantailor



Editace metadat

- Lidská kontrola výstupních obrazů (po ořezu a automatických konverzích)
- Nastavování typů stran
- Číslování stran
- Rozdělování periodik (ročník, číslo)
- Doplnění metadat
- Výbava
 - PC, které přistupuje na centrální farmu VDI – virtualizované stanice pro kompletní zpracování dat
 - Stanice (fyzické i virtuální) jsou sdíleny s pracovištěm editace metadat
- Software
 - Aplikace vyvinutá v rámci projektu na míru požadavkům NDK



Metadata editor

hvedit01.nkp.cz:9090/ndk-em/em.html?pubsid=14-08-275QPLF0V&cdnId=42627ae0-22b6-11e4-8e0d-005056827e51÷=false&suite=false&type=BK&activity=EM#0-start

procrun mode exe

Stránkování-Dělení Mets Mods Validate Nastavení O aplikaci Obnovit zálohu Uživateli: rasekl CDM: \\hvdigfsc102\CDMP_02\LINK_TARGETS\CDM_42627ae0-22b6-11e4-8e0d-005056827e51

Uložit Presun Generator Smazat Kontrola OCR

Navigace

Zrušit výběr Filtr P Zrušit filtr Index: Přejít na index

1-18 of 198

☐	I	Sken	Info	Typ	Strana	OCR
☐	1	1		Normální strana		ABBY Latinka
☐	2	1		Normální strana		ABBY Latinka
☐	3	1		Normální strana		ABBY Latinka
☐	4	1		Normální strana		ABBY Latinka
☐	5	1		Normální strana		ABBY Latinka
☐	6	1		Normální strana		ABBY Latinka
☐	7	1		Normální strana		ABBY Latinka
☐	8	1		Normální strana		ABBY Latinka
☐	9	1		Normální strana		ABBY Latinka
☐	10	1		Normální strana		ABBY Latinka
☐	11	1		Normální strana		ABBY Latinka
☐	12	1		Normální strana		ABBY Latinka
☐	13	1		Normální strana		ABBY Latinka
☐	14	1		Normální strana		ABBY Latinka
☐	15	1		Normální strana		ABBY Latinka
☐	16	1		Normální strana		ABBY Latinka
☐	17	1		Normální strana		ABBY Latinka
☐	18	1		Normální strana		ABBY Latinka

Log operací smazat

22:17:38 Načtení svazku OK
22:17:38 Test integrity dat. OK
22:17:38 Test povinných údajů. OK
22:17:38 Vytvoření history token 0-start

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16

1_image_0003_R.tif

Svazek

ID: MODS_VOLUME_0001

Název: Řád volební do obcí Republ

Podnázev: doplněný novellou k volební

Číslo části:

Rok vydání: [1919]

Poznámka:

► Popisné údaje
► Identifikátory

Zapsat do všech vybraných

Kontrola

- Finální kontrola celé entity
- Kontrola kvality obrazu
- Kontrola metadat
- Finální schválení



Zkušenosti s fyzickou manipulací

- Dostupnost vozíků na knihy
- Je nutno vhodně volit logistiku pracovišť, zkrátit vzdálenosti, usnadnit převozy
- Zpracování po dávkách
 - Při potřebě reskenu je pracné vyhledat znovu konkrétní knihu
 - Dokud nejsou plně dokončeny všechny knihy z dávky, není možné dávku uzavřít

