Zpráva ze zahraniční služební cesty

Jméno účastníka cesty	**Jan Hutař**
Pracoviště – instituce, adresa	ODO
Pracoviště – zařazení	1.6
Důvod cesty	**návštěva NK Španělska – studium procesů digitalizace a dlouhodobého uložení dig. dat**
Místo – město	**Madrid**
Místo – země	**Španělsko**
Datum (od-do)	**31.8.-2.9.2010**
Podrobný časový harmonogram	31.8.2010 let Praha- Madrid 1.9.2010 celodenní jednání v NK Španělska 2.9.2010 jednání a odlet do Prahy
Spolucestující z NK	PhDr. Bedřich Vychodil, Mgr. Tomáš Foltýn, PhDr. Jiří Polišenský
Finanční zajištění	IOP-NKP
Cíle cesty	viz důvod cesty
Plnění cílů cesty	splněno
Program a další podrobnější informace	viz níže
Přivezené materiály	
Tištěné přílohy a elektronické dokumenty	
Datum předložení zprávy	10.9.2010
Podpis předkladatele zprávy	

Biblioteca Digital Hispanica (BDH) - návštěva NK Španělska

1-2.9.2010
Jan Hutař

- Biblioteca Nacional (dále BN) má 9mil. knih ve fondu, 60% z nich je mimo hlavní budovu ve skladu
- mají také 1 mil. fotografií a 100.000 videí
- v digitální knihovně BDH mají 32.000 dokumentů rozdělených do sbírek
- NK neřeší žádné otázky koordinace – zaměřuje se jen na své věci, otázky koordinace na národní úrovni řeší Ministerstvo kultury, někdy ve spolupráci s NK

Digitalizace

- rozpočet na vybudování BDH je od firmy Telefonica – 10 mil. Euro na 5 let do r. 2012 (cílem je zdigitalizovat 200.000 děl, 25 milionů stran)
- digitalizace hist. novin – 6 mil. stran – nejsou v digitoolu, ale v proprietárním systému Pandora – jako vícestránkové PDF, chtějí to dostat do DigiToolu
- v rámci PDF mají kapitoly (bookmarks) nebo články, strukturální metadata nemají/nepotřebují
- během zavádění digitalizace se ukázalo, že jedním z hlavních problémů byly organizační změny

- Po skenovaní se materiál běžně nepůjčuje /pouze na speciální vyžádání)i, v budoucnu budou upravovat a zpřísňovat možnost výpůjček originálního materiálu/

- propracovaný systém výběru dokumentů k digitalizaci – odborní knihovníci z různých oddělení (hudební, rukopisy apod.) navrhují dokumenty + na základě UDC a dotazníků

- výběr dokumentů pro digitalizaci probíhá ze všech dostupných kopií v NK

- na základě desetinného třídění z katalog. záznamu vytvářejí sbírky v digitální knihovně

- cílem není zdigitalizovat moderní dokumenty, ale staré a jedinečné, moderní věci musí udělat jiné knihovny, např. městská (do NK si přeci nikdo nechodí půjčovat poslední bestsellery, tak proč je digitalizovat)

- **www.bdn.bne.es**

Workflow

- současná rychlost je 12000 stránek za den

- Chtějí zvýšit produkci

- údaj o tom, že byl dokument zdigitalizován se ukládá do záznamu v katalogu do pole 899 MARCu

digitalizaci zajišťuje externí firma, NK jí předá dokumenty a přebere výsledek , nemusí tedy řešit workflow digitalizace, ale jen manipulaci s fyz. dokumenty

- výpůjčka dokumentu je provedena na „umělého“ uživatele, což je digitalizační firma – není tam informace o osobě, co digitalizuje, ale pouze o firmě, která digitalizuje\

- výpůjčka je na 50? dní

- firma drží MC 2 roky po předání NK, pak je maže, za uložení předaných dat je zodpovědná NK samozřejmě

- **Naskenuji obě stránky najednou a pak rozdělí na dvě**

- používají viditelný watermark na spodní částí scanu, pracné, vkládá se ručně ve Photoshopu

výstup z digitalizace

- MC
- UC – jde do digitoolu
- XML – jde do digitoolu – jen základní bibliografický popis
- DC – pro europeanu (OAI-PMH)
- PREMIS object – na naší úrovni – uloženo s MC
- OCR – nenabízí uživateli , je jen součástí PDF a pro vyhledávání

nemají žádný identifikátor typu URN:NBN, nemají ani NBN

pro noviny mají METS záznam, ovšem jen pro strukturu, IE = titul (1x METS) nebo ročník (1x METS) – uvádějí, že je to příliš velké a nedá se s tím pracovat

Quality control

- Na obrazových datech - Firma kontroluje zaměstnanci očima, žádný SW. Kontroluji se MC jeden po druhém!

- Na obrazových datech - V knihovně se kontrolují namátkově "očima" v případě poškození se kontaktuje firma, která dodá preskenovane nebo uložené nepoškozené soubory. **Firma skladuje 1 rok nebo dva data.**

- Kontrola na popisnych metadatech UNICON \kontrola duplicit v ramci knihovny\

- Xml

Standards and Formats

MC – TIFF
UC - JPEG a PDF s OCR

- dříve omezení PDFka na 20MB, dnes 50 MB max. pak se dělí na dvě nebo vice PDF

Neimplementovali UC JP2 on demand /zadny image server/

- Využívají JP2 pouze jako UC na stáhnutí z webu \podle jich slov se JP@ neosvědčil, nechtějí využívat plug-ins, používají pouze jako UC\

- Maji pdf streaming software pro zpřístupnění, jinak jsou PDFka obrovská a z toho důvodu nedělají nic v barvě – snad kromě rukopisů…

Mají více paralelních projektů. Každý projekt produkuje "různé MC, různé nastaveni, různá kvalita.

Digitool system

- používají DigiTool se search enginem Autonomy (Velká Británie)
- Digitool není propojen na katalog, což považují za hlavní problém
- v digitoolu míchají staré i moderní dokumenty

- vytvořili na DT svoje vyhledávání – personalised field mapping, search analytics, indexují to přes OAI-PMH

- oracle – v něm mají metadata

Legal Deposit

- Čekají na nový "deposit legal act" asi příští rok, bude zahrnovat i elektronické online dokumenty, zatím jen fyzické nosiče

- Producenti budou povinni poslat produkci v digitální formě pres internet. Bez jakéhokoli hesla nebo klice

- Poslali draft, návrh, jak by to melo vypadat na MK

Archivace webu

mají nově smlouvu s Internet Archive
ten pro ně dělá sklizně domény .es

- IA jim pak data dodává, NK je archivuje
- přístup z NK
- IA data nebude nabízet přes své rozhraní
- vlastníkem dat je NK

LTP systém a ochrana dat

spolupracují s NK Francie, chtějí od nich získat open source LTP systém SPARC, který je již hotov od 2009

- zatím ukládají ve file systému mastery, podobně jako my – mají databázi kde co je a co je to za titul, včetně technických metadat (omezeně)
- Některé MC jsou uloženy na externích HDD, na serverech
- Jméno souboru "MC" reprezentuji "call numbers" – tj. systémové číslo, problém je, že se může měnit > tj. vytvořili databázi, kde každému filu dají item ID a hotovo (v dtb je systém. číslo, bibl. záznam, tech. metadata a item ID a místo uložení)
- Nemají pořádné identifikátory mezi MC a bibliografickými metadaty. Skladuji zde pouze technická metadata.
- Neplánují geograficky oddělené úložiště

na komerční systém nemají a asi nebudou mít prostředky, tj. hledají řešení open source

- při masové digitalizaci nedělají mikrofilmy

LTP debaty a uložení dat se účastní 3 oddělení

- preservation dept.
- Digital library services
- IT

Ostatní služby digitální knihovny

Interactive books

- interaktivní prezentace nejvýznamnějších děl NK
- prohlížení, linky na reálie, seznam všech edic, historie knihy, období, social networks, hudba, manuální přepis textu apod.
- Nejde o OCR jde o prepis textu
- Technologie jako google ,tiles, pyramidova struktura JPEGs
- Daji nam link na demo "Don Quijote de la Mancha"

Print on demand

- http://bne.bubok.com
- založeno na pdf
- smlouva s firmou bubok – hlavní cíl je být konkurencí pro komerční firmy v době elektr. čteček apod.
- Zkoušeli par firem na knihy, ale mají špatné zkušenosti, špatná kvalita, celkem drahé, trvá 14 dni od zadosti. Amazon

- Vidí potenciál v reprodukcích map a plakátů. Tisknou z PDF. Vypadají velmi pěkně. Tisknuto na jakoby papírovou folii, která je odolnější než papír.

Books under copyright

- lze najit a procházet par stran, pak se dá zakoupit, musí být na trhu, prodávají elektronickou verzi

<u>Dopňující poznámky Vychodil</u>

Využití grafických formátů v Biblioteca Digital Hispanica Digital library:

- MC - TIFF
- UC - JPEG
- UC - PDF s využitím OCR

Historical press \v budoucnu plánuje spojení s Digital Library dnes se jedna o oddělenou aplikaci\
Drive omezeni 20MB, dnes 50 MB max pak se dělí na dvě nebo vice PDF

- Jakou verzi PDF využívají? /neví přesně/
- Využívají vodotisk

SW - DigiTool software
Tisk na přání (Print on Demand)

- **Books under copyright** \dá se najit a procházet par stran, pak se dá zakoupit, musí být již na trhu, prodávají elektronickou verzi\
- Zkoušeli par firem na knihy, ale mají špatné zkušenosti, špatná kvalita, celkem drahé, trvá 14 dni od žádosti.
- Amazon
- Atd.
- Vidí potenciál v reprodukcích map a plakátu.
- Tisknou z PDF. Vypadají velmi pěkně. Tisknuto na jakoby papírovou folii, která je odolnější než papír.

Prezentace

- Youtube Květen 2009
- Facebook od roku 2008, 67000 fanoušku
- eBooks

Digitalizace

- Po skenovaní se materiál běžně nepůjčuje /pouze na speciální vyžádání, v budoucnu budou upravovat a zpřísňovat možnost vypůjčování originálního materiálu/

National Library of Latin American Library
Workflow

- Skenují okolo 12000 stránek\den
- Plánují navýšení produkce
- Naskenuji obě stránky najednou a pak rozdělí na dvě
- Využívají vodotisk (viditelný na spodní částí skenu, pracné, vkládá se ručně v Photoshopu)

Kontrola kvality (Quality control)
- Na obrazových datech - Firma kontroluje zaměstnanci očima, žádný SW. Kontroluji se MC!
- Na obrazových datech - V knihovně se kontroluji namátkově "očima" v případě poškození se kontaktuje firma, která dodá preskenované nebo uložené nepoškozené soubory. **Firma skladuje 1 rok nebo dva data.**
- Kontrola na popisných metadatech UNICON \kontrola duplicit v rámci knihovny\
- Xml

Co skladuji:
MC, UC, XML, Doblin Core, PREMIS /object entity/
UC a bibliographic metadata v XML jdou do Digitoolu
Ingest PDF do Digitool 800 titulů / 2 weeks (report z posledního týdne)
Při digitalizaci se kniha půjči tomu kdo digitalizuje \není tam informace o osobě, co digitalizuje, ale pouze o firmě, která digitalizuje\
- Landing book \50 dnu na půjčeni\
- 40 až 80 knih na den

Standardy a formáty
obsahuje tabulku v prezentaci
Neimplementovali UC JP2 on demand /žádný image server/
- Využívají JP2 pouze jako UC na stáhnutí z webu \podle jich slov se JP2 neosvědčil, nechtějí využívat pluginy, používají pouze jako UC\
- Maji pdf streaming software pro zpřístupnění

www.bdn.bne.es

Interaktivní knihy (Interactive books)
- Nejde o OCR jde o přepis textu
- Technologie jako google ,tiles, pyramidova struktura JPEGs
- Dají nám link na demo "Don Quijote de la Mancha"

Digitool system

Maji vice paralelních projektu. Kazdy projekt produkuje "různé MC, různé nastaveni, různá kvalita.
- Preservation Depatment \zde jsou všechny MC uloženy\

- Mass Digitization
- Historic Press

Některé MC jsou uloženy na externích HDD, na serverech.
Jméno souboru “MC” reprezentuji „call numbers”
Nemají poradně identifikátory mezi MC a bibiliografickymi metadaty. Skladuji zde pouze technická metadata.
Neplánuji geograficky oddělené úložiště

Legal Deposit

- Zálohuji pouze “.es”
- Čekají na novy “deposit legal act” asi příští rok
- Producenti budou povinni poslat produkci v digitální formě pres internet. Bez jakéhokoli hesla nebo klice
- Poslali draft, návrh, jak by to mělo vypadat.