Národní knihovna
České republiky
National Library
of the Czech Republic

Zpráva ze zahraniční služební cesty

Jméno a příjmení účastníka cesty	**Marek Melichar**
Pracoviště – dle organizační struktury	ODODD
Pracoviště – zařazení	**1.5.2.**
Důvod cesty	**IIPC fórum – účast na Preservation Working Group**
Místo – město	**HAAG**
Místo – země	**Nizozemí**
Datum (od-do)	**10-12. 5. 2011**
Podrobný časový harmonogram	10. 5. Cesta do Haagu 11. 5. 9:30 až 17:00 Jednání v Haagu 12. 5. 9:30 až 13:30 Jednání v Haagu pak cesta do Prahy
Spolucestující z NK	0
Finanční zajištění	136
Cíle cesty	Seznámení s činností Preservation WG IIPC
Plnění cílů cesty (konkrétně)	Návrhy zapojení NK do činnosti Preservation WG IIPC – viz. přiložená zpráva
Program a další podrobnější informace	10. 5. - 11:50 Odlet z Prahy, 16:30 Příjezd do Haagu 11. 5. - 9:30 Zahájení jednání (prezentace v General Assembly) 10:00 Jednání pracovní skupin 13:30 Pokračování jednání pracovních skupin 12. 5. - 9:30 Zahájení jednání (Prezentace v General Assembly) 10:00 jednání pracovních skupin. 13:00 Ukončení 19:00 Odlet z Amsterdam Schiphol
Přivezené materiály	
Datum předložení zprávy	25. 5. 2011
Podpis předkladatele zprávy	
Podpis nadřízeného	
Vloženo na Intranet	
Přijato v mezinárodním oddělení	

❖ ZPRÁVA Z JEDNÁNÍ PRACOVNÍ SKUPINY PRESERVATION GROUP, IIPC, 11. A 12. 5. 2011, HAAGUE

Marek Melichar , 13. 5. 2011.

OBSAH

INFORMACE ČINNOSTI SKUPINY IIPC WG PRO PRESERVATION V MINULÉM ROCE:

JHOVE 2

- Nástroje pro extrakci technických MD a charakterizaci – BNF ve spolupráci ve firmou ATOS Origin vyprodukovala Jhove2 modul pro validaci gzip komprese a formátu ARC
- Nastroj, umí extrahovat technická metadata a validuje také objekty uložené uvnitř ARCu
- Jhove 2 zatím nemůže zcela nahradit Jhove 1, zatím nepodporuje některé důležité formáty jako PDF nebo JP2. Vývoj většiny modulů dělá CDL – a nemá už peníze na další vývoj, jako jsou moduly pro WARC nebo html. Bude se jednat o tom, aby z financi IIPC byl znovu osloven ATOS origin. Na stránkách Jhove 2 jsou vyjmenovány formáty, které ještě stihnou udělat, dál se neví.
- Jhove 1 – umí validovat HTML soubory, zatím neumí ani jhove 2

NÁVRH PRO NK

Měli bychom udělat testovací projekt, který by poskytl zpětnou vazbu vývojovému týmu a umožnil nam ziskat s jhove2 zkušenosti. ta by byla užitečná zdaleka nejen pro data z webarchivu. Tzn. Nainstalovat jhove2, konfigurovat modul pro ARCy, připravit testovací balík dat – v řádech x desítek gb – a provést dokumentovanou validaci a různým nastavením

Bude třeba přidělit IT – infrastrukturu a vypracovat dokumentaci (anglicky) o tom, co jsme udělali. Pokud bychom byli schopni proces validace monitorovat z hlediska náročnosti na výkon procesu a času bylo by to výborné

MIGRACE ARC > WARC

- Debata o migraci ARC do WARCu – je to preservation action-přibude místo pro metadata, problém – zatím není standard jak metadata do WARCů plnit.

NÁVRH PRO NK

Vyzkoušet migraci ARC > WARC a zjistit, jestli WARC tools přidávají do těla WARCu nějaká metadata. Porovnat ARC a WARC a podle toho se rozhodnout, co dál, jestli ma cenu v našem kontextu migrovat nebo ne. Bude vyžadovat plán, infrastrukturu a zapojení IT, ODO, i WA.

NEJDŮLEŽITĚJŠÍ TÉMATA PRESERVATION GROUP

PRESERVATION METADATA PRO WEBARCHIV

BNF pracuje na implementaci PREMISu pro WA – bude navrhovat update PREMISu. BNF pošle ke komentáři návrh metadatového schématu pro WA – část bude mimo schéma PREMISu zatím. Během pár měsíců začne BNF toto schéma používat. Editorial Comitee PREMISu bude vyzvána k updatu. Schéma umožní zaznamenat „event" sklizně, nastavení crawleru, seedy, ale také zapsat do xml seznam všech objektu z ARCu nebo WARCU. (také počítají s jinými balícími formáty jako ZIP, TAR, RAR), to pouze volitelně.

NÁVRH PRO NK

Oponovat jejich metadatové schéma. Tak dalece, jak budeme schopni konfrontovat jejich návrh polí s našimi představami pro IOP.

JAK MÁ VYPADAT INFORMAČNÍ BALÍČEK PRO WEB ARCHIV

Mezi členy pracovní skupiny není shoda o tom, co má byt základní jednotkou pro archivaci. Většina archivuje a přidává metadata především na úroveň „harvesting instance" – čili sklizně, další metadata mají pro jednotlivé balíky, a jiná metadata pro collections. Velikosti balík, technologie sklizeni i organizace sklízení se mezi institucemi dost liší, někdo má. jako my, malé ARCy se smíšeným obsahem, LOC.gov má WARCy velikosti GB a vice, s metadaty třeba 4GB.

NÁVRH PRO NK

NK – znovu oživit debatu o tom, jak by mel vypadat AIP pro WA, naším cílem, je jako všech instituci, nakonec vložit data do sdíleného repositáře, udělat reálný ingest a přidat MD. Dohoda o tom, jak AIP balit a kam přidat jaká MD je třeba před implementaci LTP systému v NK. Bylo by dobře oživit use case připravovaný pro IOP – ingest dat z WA.

BUDOVÁNÍ DATABÁZE RIZIK, SOFTWARU A FORMÁTŮ

- Probíhající úkol – nástroje jako PRONOM nejsou pro WA právě ideální je třeba je doplnit. Skupina pracuje na databázi rizik a všech souvisejících technologií (formátů, rendering aplications a browserů, dokumentace k harvestorům atd.)

NÁVRH PRO NK

Měli bychom vyzkoušet projít jejich nástroj na risk assessment. Je to podobné jako DRAMBORA. Udělat pracovní skupinu (ODO + WA + IT) a vyzkoušet si hodnocení jejich nastrojem. Poslat zpětnou vazbu, opakovat za rok. Nástroj bude dostupný na novém webu IIPC.

http://www.ignaciogc.com/netpreserve/risks.php

(zatím tady, bude na novem webu IIPC, pak budou lepe reguovany pristupy)

PREZENTACE KB O NOVÉM E-DEPOTU

E-depot jako základní infrastruktura celé knihovny:

- jsou v něm všechna data (journals, digitized masters and papers i WA)

- ne všechno bude dostupné stejně rychle – digitized master za 20s z pasky

- spojují archiv, digitální knihovnu a katalog do jednoho kusu.

- všechno modulární, z různých komponentů, co vyvinou zveřejní pro ostatní

- Národní archiv NL má od Tessellu SDB, KB se velmi líbí systém workflow, tak budou vypadat i jejich workflow

- trvalý vývoj...počítají s tím, že systém bude TRVALE vyvíjen, protože to má být systém pro trvalé uložení a zpřístupnění

- před tím než začali s novým e-depotem si doma pěkně zametli – sloučili všechny digitální projekty a lepe je provázali, aby se nedělaly zbytečnosti nebo věci , nesmyslné, nespojitelné s jinými komponenty. Vše směřuje k naplnění jasné definované strategie. Chtějí mít jeden integrovaný celek, ne projekty kolem každého SW a dig knihovny, a výzkum – ale instituci, který má mission a cíle, a jednotlivé projekty je pomáhají naplnit.

Během měsíce zveřejní svoje definitivní požadavky na nové e-depot a budou k dispozici v AJ vsem.

e-depot – jejich verze transformačního modulu bude zpracovávat 28 kanálů vstupujících dat s různým nastavením ingestu, další musí byt možné kdykoli přidat. Vše postaveno na rule based workflows, workflow budou řídit vše, včetně accessu třeba.

Nástroje pro charakterizaci a podobné služby - jsou volně použitelné, zapojitelné do workflow všude v E-depotu – jakýkoli nástroj je možné zapojit v archivu, accessu nebo ingestu, kdekoli, nástroj/service je tam ale jen jednou. Týká se všech nástrojů i pro validaci MD například-

Mají nějaké komponenty systemu nad OAIS, především tzv. „Proces data store“ – dokumentace procesu v celém e-depotu pro reporting, management IT, cost analyses, služby vydavatelům atd atd.

V květnu 2012 první ingest –nejdříve migrují journals, převod starých dat, pak přepnou všechny dodavatele obsahu (několik desítek velkých vydavatelů odborných časopisů) na nový E-depot. 2013 začnou do E-depotu dávat data z digitalizace, v létě 2013 websites – single systém, single point of Access.

Systém E-depot se dotkne všeho v knihovně, každý dokument jim projede.

Hlavní jejich snaha je vyhnout se „company lock-in" a také „black box solution" – to už měli dost dlouho od IBM. Momentálně dělají RFI na storage management systém, je pravděpodobné, že tohle možná bude komerční řešení – chtějí mít poměrně složité uložiště, s řadou pravidel a policies pro různé typy dat. Do E-depotu nepůjdou jen data archivní, ale i data krátkodobého významu.

Systém jako celek obsahuje jak databázi pro zpřístupnění (asi i indexy atd), tak archivní databázi a navíc ještě onu „process data store" databázi pro management a CRM.

NÁVRH PRO NK

Sledovat jak budou pokračovat, v blízké době zveřejni (mail konference, web atd.)

1. *Requirements na svůj systém*
2. *Data model - !*
3. *Svoje představu o storage management systému*
4. *Sledovat OPF – dají tam všechno, co půjde, z toho co vyvinou*
5. *Ředitel NK by tam měl jet, a měl by mluvit s Hilde a lidmi nad ní a kolem ní.*

DALŠÍ ZAJÍMAVÉ PREZENTACE:

- Implementace SORLU do hledání ve WA /dobré nápady – filtrování podle typu dokumentu – formáty pdf, ppt, filtry podle času sklizně nebo full text- nevím kolik toho dovede náš Nutchwax/
- Projekty SCAPE a KEEP – obzvlášť SCAPE je mamuti projekt pokračovaní Planets. Budou dělat 1000 člověku-měsíců – vývoj z vetší části – v Planets se ukázalo, ze nástroje, které mají jsou omezené na menší pocty dat, nelze je použít na miliony objektů. Scape je asi největší projekt na DP v EU ever.
- ISO standard pro WA –digitální strategii to nepůjde. Koordinovaně říct co chceme dělat v čem je přínos (WA digitalizace), marketing a komunikace se stakeholdery. Kvalita – kolik obsahu WA v CR už není online? Lze to zjisti?
- COST – jak vyjádřit náklady na archivaci WA, jak měřit
- Recollection tool loc gov – SEE WEB
- Quality assurence v kontextu WA..

NÁVRHY PRO NK

Marketing pro WA – různé definovat cílové skupiny a na ně zaměřit komunikaci-

Popularizace WA u badatelů (politologové, novináři, studenti, lingviste atd.) – obsah, lepší pohledy na archiv, lepší vyhledávání, filtry, highlights, special collections, text dumps analýzy google NGgram atd. Stanovit cíle – počet UV pro rok 2012, sledovat live kolik toho má NK jak jedina, kolik už není on-line…

Popularizace WA mezi techniky a IT komunitami – analýza obsahu WA – formáty, objemy, technologie sklizeni, emulace a zpřístupnění

Změna technologie sklízení WA

Zvážit – proč nesklízet větší dokumenty (na 100MB, videa atd)

Implementovat efektivní deduplikaci do workflow – uvolněné místo je až 80 % při opakovaném targeted harvestu. Jinak alespoň 30%.

NK CR Obecně

NLCR by si měla udělat jasnou strategii, jak a co chce dělat v digitálním světě: Vytvořit srozumitelnou collection policy, digital strategy,– jasně říct, že WA chceme dělat a dáme na to peníze, resp. budeme je hledat…..stejně tak ostatní projekty, bez jasné strategie a plánu se budeme potácet v chaosu daleko ze KB a BNF, BL atd.. Například rozhodnutí, že nesklízíme WA některé dokumenty (videa z webu) měli bychom mít jasně argumentované ve collection strategii – že se nás to netýká – a nebo týká? Kdo jiný sklidí ten nadměrný obsah?

Podle mě je třeba zřídit koordinátora pro přechod knihovny do digitálního světa, který ale nebude mít jen poradenskou roli, ale i exekutivní: Člověk s vizí, který udělá strategii přizpůsobení digitálnímu světu a pak ji bude postupně prosazovat (Národní knihovna v pohybu), navzdory zájmům různých skupin v NK, firem, který z ní žijí, navzdory organizační strnulosti atd. Inspirovat se organizačními změnami v KB – všechny projekty směřují k naplnění jasně formulované strategie….

-