Národní knihovna
České republiky
National Library
of the Czech Republic

Zpráva ze zahraniční služební cesty

Jméno a příjmení účastníka cesty	**Libor Coufal**
Pracoviště – dle organizační struktury	OAW 8.1.1
Pracoviště – zařazení	**knihovník**
Důvod cesty	**Jednání pracovní skupiny IIPC Preservation a konference The future of the past of the web**
Místo – město	**Londýn**
Místo – země	**Velká Británie**
Datum (od-do)	**6. – 7. 10. 2011**
Podrobný časový harmonogram	6.10. dopoledne – přílet, jednání PWG 7.10. konference, odlet
Spolucestující z NK	J. Hutař, R. Kreibich
Finanční zajištění	VaV.0121
Cíle cesty	- účast na jednání pracovní skupiny Preservation (PWG) - účast na konferenci
Plnění cílů cesty (konkrétně)	Viz zápis v příloze
Program a další podrobnější informace	Viz zápis v příloze
Přivezené materiály	
Datum předložení zprávy	12.10.2011
Podpis předkladatele zprávy	
Podpis nadřízeného	
Vloženo na Intranet	
Přijato v mezinárodním oddělení	

Čtvrtek 6.10.2011

Jednání skupiny Preservation Working Group

Clement Oury (BnF) – Arc to Warc conversion mapping

BnF připravila dokument, který obsahuje návrh mapování jednotlivých polí při konverzi z formátu Arc na formát Warc. Dokument popisuje jednotlivá pole v Arc a Warc, údaje, které se do nich zapisují a jak jsou tato pole mapována mezi Arc/Warc. Dokument byl podrobně projednán. Konečným cílem by mělo být vytvoření určitého standardu, kterého by sloužil jako vodítko pro instituce při migraci z Arc na Warc. Dokument bude postupně dále upřesňován. (Pozn.: NK ČR v současnosti také pracuje na této problematice v rámci institucionálního výzkumu – naše závěry budou použity pro srovnání; dále se ukazuje jako nutné vzít v úvahu strukturu a možnosti zapisování do Warců při sklízení v budoucnu – struktura těchto „původních“ a konvertovaných Warců by si měla co nejvíce odpovídat).

Amanda Spencer (TNA) – informace o Pronom/Droid

TNA pracuje na nových verzích nástrojů Pronom a Droid

Pronom
- měl by být konvertován do formátu LinkedData
- připravují požadavky, poté najmou externího programátora (příští rok)

Droid
- proběhne veřejná konzultace k dalšímu vývoji, výsledky by se měly promítnout do příští verze Droid

Steve Knight (NLNZ) – zdá se, že některé požadavky/potřeby TNA jsou v rozporu s tím, co potřebují ostatní instituce – bylo by vhodné více koordinovat a najít společné zájmy, které by se měly vzít v potaz při vývoji

Barbara Sierman (KB) a Sven Schlarb (ONB) – informace o evropském projektu SCAPE
- 2012 – 2016, 16 partnerů
- 5 podprojektů (Platform, Preservation and watch, Preservation components, Take up, Testbed)
- k dispozici první výsledky (zpráva Evaluation of characterisation tools - Part 1: Identification dostupný ze stránek projektu Open Planets http://www.openplanetsfoundation.org/system/files/SCAPE_PC_WP1_identification21092011_0.pdf a zpráva Identification and selection of large-scale migration tools and services http://www.scape-project.eu/wp-content/uploads/2011/09/SCAPE_D10.1_KEEPS_V1.0.pdf)

Pátek 7.10.2011

Konference The future of the past of the web

Podrobný program konference je na http://www.dpconline.org/events/details/35-future_past_web?xref=35

Výběr z nejzajímavějších příspěvků:

H. Van der – mluvil o projektu Memento; tento projekt byl představen už na dřívějších akcích IIPC, naposledy na valném shromáždění IIPC v Haagu v květnu t.r. V současnosti byly vytvořeny další aplikace pro Firefox, Android, mcurl a platforma pro Wikimedia. V druhé části příspěvku přednesl svou vizi o nutnosti archivovat akademické a vědecké publikace na webu. Tato nutnost vyplývá z povahy těchto publikací, jejichž obsah, kontext a reference se, na rozdíl od tradičních tištěných publikací, často mění. Jeho vize řešení tohoto problému vychází z rozšíření metodologie použité v projektu Memento na tyto typy publikací, tak aby jejich uživatelé byli schopni využívat webové vědecké publikace v původním kontextu tak, jak vypadal při jejich vzniku.

Maureen Pennock a Lewis Crawford z BL představili nové možnosti práce s britským webovým archivem – 3D vizualizace archivovaných verzí stránek, generování tag cloudů z nejčastěji se vyskytujících klíčových slov, n-gramy – srovnání četnosti výskytu 2 klíčových slov nebo frází v čase – vše založeno za aplikaci Big Sheets od IBM. Vyhledávání v archivu umožňuje také klustrovat výsledky podle 3 typů – dokumenty, obrázky a média.

Celkově se celá konference nesla v duchu budoucího využití archivů – možností práce s obrovskými objemy dat a data miningu.

Národní knihovna
České republiky
National Library
of the Czech Republic

Zpráva ze zahraniční služební cesty

Jméno a příjmení účastníka cesty	**Libor Coufal**
Pracoviště – dle organizační struktury	OAW 8.1.1
Pracoviště – zařazení	**vedoucí oddělení**
Důvod cesty	**Účast na konferenci iPres 2011 a souvisejících workshopech a turoriálech**
Místo – město	**Singapur**
Místo – země	**Singapur**
Datum (od-do)	**30.10.-6.11.2011**
Podrobný časový harmonogram	30.10.-31.10. – let Praha-Frankfurt-Singapur 1.11. – turoriály Premis a Archiving websites 2.-4.11. – konference iPres 4.11. odp. – workshop Web Analytics 5.-6.11. – let Singapur-Mnichov-Praha
Spolucestující z NK	J. Hutař, M. Melichar
Finanční zajištění	VZ137
Cíle cesty	Seznámení s novými trendy v oblasti archivace webu a digitálního uchování, účast na turoriálu Premis a workshopu Web Analytics
Plnění cílů cesty (konkrétně)	Viz podrobný zápis níže a sborník
Program a další podrobnější informace	http://ipres2011.sg/pages/programme-overview
Přivezené materiály	e-sborník (usb, uložen na sps – složka se zprávami ze SC)
Datum předložení zprávy	20.11.2011
Podpis předkladatele zprávy	
Podpis nadřízeného	
Vloženo na Intranet	
Přijato v mezinárodním oddělení	

Tutorial Premis
Sébastien Peyrard (BnF) – úvod do PREMIS
Peter McKinney (NLNZ) – implementace PREMIS v Národní knihovně Nového Zélandu
Haliza Jailani (NLB Singapur) – implementace PREMIS v NLB Singapuru

Evaluation of Danish migration project (Danish National Archives)

- DNA rozhoduje, jaké formáty přijímají
- Všechny digitální objekty jsou plně standardizované
- Probíhá 4letý projekt na digitální uchování
- 2005-8 – 30 člověkoroků na migraci, 190 tis. USD na investice, celkem 2,6 mil. USD, 15 specialistů

Zpráva je pracovníkem do mezinárodního oddělení předložena nejpozději při vyúčtování cesty do 2 týdnů po jejím ukončení. Bez cestovní zprávy nebude provedeno vyúčtování. Při výjezdu více pracovníků na tutéž služební cestu s týmž programem lze odevzdat společnou cestovní zprávu.

- Standardizace záznamů – umožní automatizaci budoucích procesů
- Část starých magnetických pásek se nedala přečíst, museli je obnovovat, velké náklady – cca 9 tis. USD na pásku!
- Standard pro uchování – AIP vychází ze SIP (je stejný)
- Problém při nakládání se zastaralými formáty – databáze, COBOL,...
- Migrace velmi starých dat může být velmi nákladná
- Standardization is the key!

Developing robust migration workflow for preserving and curating hand-held media (Angela Dappert, British Library)

- migrace z „přenosných" nosičů na online média – stabilizace
- strategie – disk image – není „bit faithful" – jiná framentace, vymazané soubory jsou „ztraceny" X file extraction
- testovali různé typy automatizovaných podavačů CD – LIFO, FIFO – zajímavé praktické zkušenosti, které by se daly využít

Preserving web archives (panel)
Sébastien Peyrard (BnF)

- 200 TB dat z webového archivu, 1,5 mil ARC
- Ukládají v LTP systému SPAR, kapacita 16 PB
- Charakterizace a validace časově náročná
- Vyvinuli modul pro JHOVE2 pro formát ARC – pouze validace konterjnéru, nikoliv vlastního obsahu!
- Bude se vyvíjet podobný modul pro WARC
- Zapisují metadata na úrovni AIP
- Vytvořili speciální metadatový formát ContainerMD

Peter McKinney (NLNZ)

- výběrové + celoplošné sklizně, celkem 20 TB
- ukládání všech požadovaných metadat by byl problém z hlediska velkosti
- jak určit vhodný rozsah metadat, který bude „dostačující" a nezatíží kapacity?
- Používají WebCurator Tool (WCT)
- Workflow pro výběrové sklizně včetně katalogizace

Aaron Binns (IA)

- nejstarší webový archiv – od 1996
- 1,6 mld. archivovaných URL
- Každý den přírůstek 3TB, ročně 1PB!
- Snaží se mít vše na rychlých discích – rychlé a levné
- Průběžně obměňují/zvyšují kapacitu
- Uložení v několika lokalitách

Workshop Web Analytics
Aaron Binns (IA)

- představení nástrojů pro analýzu dat z webových archivů – velké objemy dat
- co je v archivu, co schází?
- Jaká je struktura archivu?
- Pochopení obsahu archivu
- Prokázání užitečnosti archivu
- Lepší uživatelské rozhraní – zvýšení hodnoty archivu
- Umožnění výzkumu

Používaná platforma:
SW – Apache Hadoop (framework pro distribuované výpočty) + Apache Pig
Datové formáty – WARC, CDX, WAT

Zpráva je pracovníkem do mezinárodního oddělení předložena nejpozději při vyúčtování cesty do 2 týdnů po jejím ukončení. Bez cestovní zprávy nebude provedeno vyúčtování. Při výjezdu více pracovníků na tutéž služební cestu s týmž programem lze odevzdat společnou cestovní zprávu.

Zpráva ze zahraniční služební cesty

Jméno a příjmení účastníka cesty	Rudolf Kreibich	
Pracoviště – dle organizační struktury	8.1 ODF	
Pracoviště – zařazení		
Důvod cesty	Konference LAWA	
Místo – město	Paříž	
Místo – země	Francie	
Datum (od-do)	14-16.11.2011	
Podrobný časový harmonogram	14.11 let Praha > Paříž 15.11. účast na LAWA konferenci 16.11 odlet Paříž > Praha	
Spolucestující z NK		
Finanční zajištění		
Cíle cesty	Účast na konferenci LAWA	
Plnění cílů cesty (konkrétně)	cíle naplněny	
Program a další podrobnější informace	viz níže	
Přivezené materiály		
Datum předložení zprávy		
Podpis předkladatele zprávy		
Podpis nadřízeného	Datum:	Podpis:
Vloženo na Intranet	Datum:	Podpis:
Přijato v mezinárodním oddělení	Datum:	Podpis:

RDF Ontologie – Pierre Senellart (Telecom ParisTech, Webdam project, aplikace Paris)
Současné ontologie (dbpedia, freebase, yaho, uniprot) mají podobné či překrývající se data. Je snaha ontologie normalizovat na úrovni entit i vztahů mezi nimi a vytvořit tak kombinovanou ontologie z více bází. Mapování se provádí pomocí pravděpodobnostního modelu.

Experiment s propojením YAGO + Dbpedie se stále optimalizuje. Problém s dočasnými vztahy např. prezidenství. Ontologie se neustále vyvíjí, jak pracovat s dynamickým obsahem.

http://webdam.inria.fr/wordpress/
http://webdam.inria.fr/wordpress/?p=893
http://wiki.dbpedia.org/About
http://www.mpi-inf.mpg.de/yago-naga/yago/

Pokud budem někdy provádět seménatickou analýzu textů, bude dobré zkontrolovat současné ontologie např. na bázi wikipedie či NTK PSH.

INTERNET MEMORY FOUNDATION

Zpráva je pracovníkem do mezinárodního oddělení předložena nejpozději při vyúčtování cesty do 2 týdnů po jejím ukončení. Bez cestovní zprávy nebude provedeno vyúčtování. Při výjezdu více pracovníků na tutéž služební cestu s týmž programem lze odevzdat společnou cestovní zprávu.

http://internetmemory.org/en/
Prezentovali hlavně infrastrukturu: z 200TBS na 1 PB = large-scale crawling – HDFS/HBASE
Sociální sítě sklízejí ve spolupráci s Hanzo Archives
řeší vizualizace běhu hadoop.

DATA PRIVACY OF SEARCH QUERIES, Vicenc Torra
Problematika nabízení search-query dat. Zatím u nás jen otázka, jak lidé u nás hledají v archivu, zatím search logy jiných institucí nearchivujeme.
Základ je anonymizovat uživatele. Ovšem otázka jak toto řešit na sociálních sítích. Zpřístupnit už anonymizovaná data, přičemž uchovat shluk hledání, tj. aby se nevědělo kdo hledal, ale že někdo hledat to a to.
Stejně postupovat při vztazích na sociálních sítích.

FOREST, Marilena Oita
a Pierre Senellart
Otázka jak se vypořádat s obrovským množství uživateli generovaným obsahem (blogy, tweety, zprávičky, wiki články, diskuzní příspěvky atd) a jak v něm najít to podstatné.
vytvořit WEB CHANNEL: agregované feedy z různých (i sociálních) sítí.
Standardizovaná struktura: Titul, popis, čas vydání
URL:TITLE:DESCRIPTION
Jednotlivé uzly (nodes) shlukovány pomocí tag paths (Strukturální vlastnost feedu). Feedy analyzovat na základě překvapivosti a statistické sémantické hustoty. Neb co je překvapivé s sémantické husté, tak je relevantní.
Stejné aplikovat i na samostatné stránky, tj. nelimitovat se jen na feedy.
Tj. kurátoři již při sklizni, nikoliv lidsky v post procesu.

TEXT, ENTITY, TIME ANALYSIS – Marc Spaniol
Webový archivy jsou zlatým dolem pro časovou analýzu: veřejená prohášení, výrobky, instituce a společnosti, technologie, vývoj pojmů a frází.
Uživatelé: Sociologové, politologové, market výzkumníci.

Identifikovat relevantní události a intervaly. Získat pojmy a fráze. Identifikovat, odlišit a sledovat entity. Shlukovat informace podle důležitosti. Postupně analyzovat data na větší úrovni, až po celý web.

Problém jak odlišit jména. – se řeší hlavně předchozím zaznamenaným vztahem. Oni řeší pomocí graphu a dotazování ontologie. Stejně tak řeší i časový údaj o události. Byli schopní z věty strojově zjistit že se jedná o hráče, který hraje v určitém klubu a že v nějaký čas odešel. Tj. namapovali událost do ontologie.

http://www.lawa-project.eu/index.php/news/aida_an_online_tool_for_accurate_disambiguation_of_named_entities_in_text_a
aplikace: http://www.mpi-inf.mpg.de/yago-naga/aida/

YAHOO RESEARCH: TIME EXPLORER

Zpráva je pracovníkem do mezinárodního oddělení předložena nejpozději při vyúčtování cesty do 2 týdnů po jejím ukončení. Bez cestovní zprávy nebude provedeno vyúčtování. Při výjezdu více pracovníků na tutéž služební cestu s týmž programem lze odevzdat společnou cestovní zprávu.

Zajímali se o hledání na základě času, tj. hledání textů s prognózami.
Jejich aplikace Time Explorer: http://fbmya01.barcelonamedia.org:8080/future – žel mi to nikdy neběžela..
Diversity aware hledání: zdroj, místo, čas, jazyk, nároz, rank. Pozná entity i vztahy.
Zdroje jako NY times, málo diverzifikovaný na rozdíl od webarchivních kolekcí.
Workflow: výtah entit, TimeML, Sentiment, ontologické mapování na YAGO, anlýza obrázů (sentiment, rozpoznání tváře).
Solr: na úrovni vět a dokumentů.
Jen pro anglické texty.
Objevli např. Vztah mezi pojmy války v Jugoslávii a Husajnem – tehdy se o Mileševičovy mluvilo ve spojitosti s Husajnem.

Degeneracy based community evaluation, Michalis Vazirgiannis

WWW je graph, Společnské sítě a citační indexy jsou vztahové graphy. Takové vztahy mohou být propojené (WWW) či potvrzené (sítě důvery, sociální sítě atd.). Neustále se mění tvarem i velikostí.
Hledání komunit a jejich vyhodnocování v graphech. Různé přístupy (huby, autority, centrálita, mezičlínovost, strukturální podobnost apod.)

Řešili k-core analýzu, rankovali autory v citacích.. Např. knihy co měli 119 spoluautorů.
http://graphdegeneracy.org/

UK Webarchive
Chybí jím kontext k obrázkům
Chtěli by profilovat webové stránky, jestli odpovídá kurátorskému profilu.
Používají geografické entity (PSČ), ale mají jich 15 miliónů.
Vědí co mají, ale nevědí jestli je to užitečné.

Zpráva je pracovníkem do mezinárodního oddělení předložena nejpozději při vyúčtování cesty do 2 týdnů po jejím ukončení. Bez cestovní zprávy nebude provedeno vyúčtování. Při výjezdu více pracovníků na tutéž služební cestu s týmž programem lze odevzdat společnou cestovní zprávu.