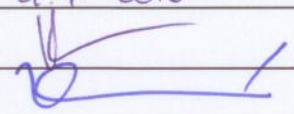
Národní knihovna
České republiky
National Library
of the Czech Republic

Zpráva ze zahraniční služební cesty

Jméno a příjmení účastníka cesty	**Jaroslav Kvasnica**
Pracoviště – dle organizační struktury	2.4.1 - OAW
Pracoviště – zařazení	knihovník
Důvod cesty	**účast na valném shromáždění konsorcia IIPC a konference Web Archiving**
Místo – město	**Reykjavík**
Místo – země	**Island**
Datum (od-do)	**9.4. - 16.4.**
Podrobný časový harmonogram	9.4. - odlet z Prahy 11.-12.4. - valné shromáždění IIPC 13.-15.4. - konference Web Archiving 16.4. - návrat do ČR
Spolucestující z NK	Barbora Rudišinová
Finanční zajištění	VaV 0136
Cíle cesty	Cílem byla účast na shromáždění a seznámení se s aktuální činností konsorcia a se současnými trendy v oblasti archivace webu.
Plnění cílů cesty (konkrétně)	Cíle byly splněny. Byly získány poznatky, které budou moci být využity pro činnost a rozvoj webové archivace digitálních dokumentů v NK ČR.
Program a další podrobnější informace	Viz níže
Přivezené materiály	-
Datum předložení zprávy	29. 4. 2016
Podpis předkladatele zprávy	
Podpis nadřízeného	
Vloženo na Intranet	
Přijato v mezinárodním oddělení	

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.

Podrobnější informace o programu a jednotlivých přednáškách je k dispozici na webu IIPC: http://netpreserve.org/general-assembly/2016/overview

11. - 12. 4. Valné shromáždění konsorcia International Internet Preservation Consortium (IIPC)

11. 4. První den valného shromáždění

Valné shromáždění IIPC bylo otevřené pouze členům konsorcia. První den byl věnován zejména organizačním a administrativním záležitostem konsorcia. V rámci programu byli představeni znovu zvolení členové řídícího výboru a jednotliví členové konsorcia stručně prezentovali aktuální dění ve svých organizacích. Odpolední program pak byl věnován novinkám v činnosti jednotlivých pracovních skupin konsorcia a setkávání jednotlivých pracovních skupin v rámci konsorcia.

Novinky u členů:

- Francouzská národní knihovna
  - zavedla nové služby vytvořené na míru badatelům (text & data mining)
  - vytvořila tzv. web archive labs, ve kterých testuje nejrůznější nástroje, které pak zavádí do praxe
- OIA Germany
  - firma představila nové webové systémy a služby pro webovou archivaci určené pouze pro instituce typu knihoven. http://www.oia-owa.de/
- INA Theque (Fr)
  - představení služby a workflow pro vysokofrekvenční harvesting (několikrát denně), využití je zejména pro TV, video, sociální media a zpravodajství
- Portugalský webový archiv
  - nakoupil kompletně nové uložiště v řádu PB
  - v minulém roce pro své uživatele představil nové uživatelské rozhraní zpřístupňující aplikace
  - jejich archiv nyní obsahuje 2 600 milionů souborů
- Nizozemská národní knihovna
  - začala pracovat na sklízení digitálního zpravodajství na denní bázi
  - na svém úložišti začali s přípravou na zavedení normy OAIS
- Stanford university libraries
  - vyvíjí a využívají Fedoru ve workflow webového archivu
  - používají Blacklight katalog a pro vyhledávání SearchWorks
- Netarchive.dk
  - vydali novou verzi nástroje NetarchiveSuite v. 5.1, která již pracuje s Heritrix 3

12. 4. - Druhý den valného shromáždění

Druhý den shromáždění byl rozdělen do tří paralelních bloků na základě tematického zaměření dle pracovní skupiny. Zúčastnili jsme se bloku pracovní skupiny CDG věnovaného vytváření sbírek ve spolupráci - **collaborative collections**, kde byly prezentovány jednotlivé společné sbírky za minulý rok. Jedná se o tematicky zaměřené kolekce webových zdrojů, které jsou dodány spolupracujícími institucemi a poté v rámci služby Archive-It zaarchivovány. V loňském roce proběhla takováto sbírka na téma připomínky první světové války. Tuto sbírku koordinovala Národní knihovna Francie. Druhou loňskou tematickou sbírkou ve spolupráci byla sbírka zaměřená na uprchlickou krizi, které jsme se také zúčastnili dodáním zpravodajských webových zdrojů z českého prostředí. Byly diskutovány možnosti

pokračování archivace této sbírky. Jako další společná sbírka je budována také kolekce webových adres mezinárodních a nevládních organizací na doméně .int.
Nastíněny byly také plánované sbírky, vytvářené ve spolupráci, na tento rok 2016. Plánované jsou zejména dvě sbírky, jedna k příležitosti letních olympijských her a druhá zaměřená na zpravodajství z celého světa. V ideálním případě by do této sbírky měly být získány webové zdroje nejvýznamnějších elektronických zpravodajských portálů ze všech zemí světa. Jako poslední byla prezentována činnost Portugalského webového archivu v oblasti vytváření tematických sbírek. Portugalský webový archiv upozornil na potřebu archivace vědeckých a výzkumných výstupů, zejména z projektů. Tato vědecká data jsou publikována především elektronicky a v rámci časově ohraničených projektů, což znamená, že po skončení projektu často tyto webové stránky s výstupy zanikají. V rámci prezentace bylo tedy apelováno na nutnost pokusit se tato cenná data z výzkumu a vědy získat a uchovat do budoucna.

13. - 15. 4. Konference Web Archiving

Keynote: Digital Salvage Operations – What's Worth Saving? – Hjálmar Gíslason
Úvodní přednáška se nesla v duchu "Hoarding is not a strategy", tedy hesla, že ne vše stojí za archivaci a hromadění co největšího objemu dat, není vhodnou strategií archivace digitálního obsahu. Jako příklad byl uveden videoportál youtube.com, kde je nahráno každou minutu 500 hodin videa, k čemuž by bylo potřeba 150 000 tisíc zaměstnanců, aby stíhali vše shlédnout a v tomto případě víme, že nestojí za archivaci vše. Problém je, ale jak vybereme co ano?

Panel: Rethinking Web Archiving – Developing Services for National Libraries – Helen Hockx-Yu
Helen Hockx-Yu pod záštitou instituce Internet Archive provádí výzkum webové archivace v národních knihovnách po celém světě. Výzkum ještě není u konce, nicméně Helen již prezentovala první výsledky. Výzkumu se prozatím zúčastnilo přes 30 národních knihoven z celkového počtu 246.
První zjištění tohoto výzkumu:

- u národních knihoven jsou rovnocenné přístupy in-house archivace nebo využití externích služeb
- většina dělá celoplošné + výběrové sklizně
- definování teritorality mimo národní doménu je komplikované a panuje zde nekonzistence
- mezi nástroji převládá Heritrix a Wayback Machine, ale s vlastními rozšířeními
- všechny webové archivy národních knihoven se potýkají s rozpočtem, zejména se objevuje trend krácení rozpočtových prostředků
- národní knihovny by rády archivovaly sociální media, ale nemají k tomu potřebné nástroje
- většina webových archivů má omezený přístup (copyright)
- webové archivy spojuje přání mít více uživatelů
- nedostatek uživatelů brzdí vývoj webových archivů (nedaří se obhájit rozpočtové prostředky)

Sekce: Emulation

- emulace jako strategie dlouhodobé ochrany webových zdrojů se začíná jevit jako vhodnější řešení než migrace a to zejména díky vývoji emulačních frameworků ve webovém prostředí (D. S. Rosenthal)
- I. Kremyer představil emulační aplikace oldweb.today, která funguje přímo ve webovém prohlížeči a narozdíl od ostatních emulačních frameworků je vytvořena přímo pro účely webových archivů.
    - oldweb.today emuluje různé webové prohlížeče z různých časových období
    - aplikace umožňuje emulaci HTTP protokolu, která je právě specifická pro webové emulátory - tímto může využívat předkonfigurovatelný proxy server pro zobrazování obsahu z archivu a tím se zamezí tzv. live leakům (situacím, kdy jsou některé části archivní verze zobbrazeny z živého webu, typicky se toto děje u reklamních skriptů.)
    - díky variabilitě aplikace a možnosti zapojení více browserů a operačních systémů se otvírá nový prostor pro obor software curation, kdy bude nutné určit, na jakém software archivní web funguje.
- Thomas Liebetraut představil největší výzvy, které představuje dlouhodobá ochrana webového obsahu a které by se mohly vyřešit díky emulačním frameworkům: autentické renderování (java, midi, flash), webové servery (deep web archiving), webové služby (aplikace, machine-to-machine communications)

Sekce: Policy and practices

V této sekci byly představeny iniciativy a přístupy národních knihoven v oblasti strategie archivace. První přednáška shrnula možnosti spolupráce institucí a organizací v Nizozemí, které se webovou archivací zabývají. V další z přednášek představila Sabine Schostag z Dánské národní knihovny jejich hledání vhodného způsobu dokumentace archivovaných zdrojů z výběrových sklizní. V poslední prezentaci tohoto bloku přestavili němečtí kolegové svůj projekt archivace elektronicky publikované literatury německých autorů. Na základě jejich workflow ukázali s jakými technickými i administrativními překážkami se potýkají při archivaci moderní "born-digital" literatury (nutnost získávání souhlasu s archivací autory, multimediální díla obsahující např. i video nebo audio materiál aj.).

Sekce: Quality Assurance Practices

První přednáška se zabývala praxí kontroly kvality v Britské knihovně, ve které nyní archivují webové zdroje na základě povinného výtisku el. dokumentů (legal deposit). Proces kontroly kvality zahrnuje posouzení různých kritérií jako jsou:

- zda byla data úspěšně stažena
- zda byl obsah zachycen a lze zpětně zobrazit
- chování - např. funkčnost odkazů
- vzhled stránky

Na základě kontroly kvality je posouzeno, zda byl webový zdroj zachycen v dostatečné kvalitě tak, aby tuto archivovanou verzi bylo možné uchovávat jako platnou archivní kopii. Dále byla prezentována změna v procesu kontroly kvality od období pouze výběrových sklizní, kdy tento povinný výtisk ještě nebyl v platnosti, po současnost, kdy jsou na základě povinného výtisku archivovány veškeré britské zdroje v rámci celoplošných sklizní a paralelně jsou také prováděny menší výběrově zaměřené sklizně (speciální kolekce, zpravodajské weby a klíčové zdroje).Vzhledem k nárůstu objemu archivace bylo nutné stanovit priority, zapojit do tohoto procesu i automatizované činnosti (Web Curator Tool QA Module, Kibana) a vytvořit zcela nové workflow. V druhé přednášce byla z pohledu webového archivu univerzitní knihovny vyjádřena důležitost automatizace alespoň části procesů při kontrole kvality pro úsporu práce i financí a byly také zmíněny nástroje, které k tomu mohou sloužit (Trello, Snagit, Zapier, IFTTT).

Výzkumná sekce

Standfordská univerzitní knihovna a Vládní archiv UK prováděly výzkum potřeb badatelů. Z výzkumu vyplynulo, že badatelé v dnešní době ještě nejsou schopni přesně definovat, jak velkou část webového archivu ke svému výzkumu potřebují. Nicméně je zřejmé, že badatelé potřebují přístup k datasetům, ne pouze k jednotlivým archivních kopiím stránek. Webové archivy musejí badatelům pomoci definovat, jaký dataset potřebují a musejí nabídnout badatelům nástroje, které umožní přístup k těmto datasetům.

Jefferson Bailey z Internet Archive ve své prezentaci sdílel své zkušenosti se spoluprací s konkrétními badateli. Z těchto zkušeností vyplynulo jednak několik technických problémů, které je třeba do budoucna vyřešit. Konkrétně se jedná o velikost dat a s tím související nároky na infrastrukturu, diverzita archivních dat a náklady na práci s velkým objemem dat. Co se týče koncepčních problémů, tak jde zejména o akvizici (jaká data archivovat), "lákadlo" neustále narůstajících objemů dat (více dat není lepší než méně dat) a původ dat (mnoho dat, ale z nich badatel nechce úplně všechny).

Shrnutí výzkumné sekce je, že badatelé často nevědí, co chtějí a většinou chtějí přístup ke všemu, i když to nepotřebují a nemají prostředky na zpracování tolika dat. Badatelé nepotřebují obrovské datasety, ale flexibiní "delivery" služby.

Archivace národní domény

- Estonská národní knihovna představila svůj nástroj Krool, který slouží ke správě sklízečů (až 75) a umožňuje individuální sklízení jednotlivých semínek
- Portugalský webový archiv referoval o své ztrátě části webového archivu a opatření, které po nehodě zavedli. Mezi tyto opatření patří: no blade systémy; záloha na páskách s náhodnými testy obnovení dat; záložní kopie na harddiscích; duplicitní použití monitorovacích služeb; QA pro vývoj software (dokumentace, testování správců atd.)
- Pilotní sklizeň .EU domény: 3,9 milionů domén; Portugalský archiv a projekt RESAW
- Dánský národní webový archiv představil nové workflow pro identifikaci semínek patřící k dánskému webu (nástroj Netakivet)
- Portugalský webový archiv provedl srovnávací testy Wayback Machines (WM 1.2.1, PyWB, oWM), nejlepší návratnost má PyWB, nejrychlejší zobrazování stránek oWM