Zpráva ze zahraniční služební cesty

|  |  |  |
| --- | --- | --- |
| Jméno a příjmení účastníka cesty | Rudolf Kreibich | |
| Pracoviště – dle organizační struktury | 2.2.3 podpora aplikací NK/NDK | |
| Pracoviště – zařazení |  | |
| Důvod cesty | **konference o webové archivaci** | |
| Místo – město | **Londýn** | |
| Místo – země | **Velká Británie** | |
| Datum (od-do) | **13.-17.6.2017** | |
| Podrobný časový harmonogram | 13.6 let Praha > Londýn  14.6 – 16.7 účast na konferenci  17.6 odlet Londýn - Praha | |
| Spolucestující z NK |  | |
| Finanční zajištění | VaV | |
| Cíle cesty | účast na konferenci, setkání s kolegy z ostatních knihoven. | |
| Plnění cílů cesty (konkrétně) | cíle naplněny | |
| Program a další podrobnější informace | viz níže | |
| Přivezené materiály |  | |
| Datum předložení zprávy | 25.6.2017 | |
| Podpis předkladatele zprávy |  | |
| Podpis nadřízeného | Datum: | Podpis: |
| Vloženo na Intranet | Datum: | Podpis: |
| Přijato v mezinárodním oddělení | Datum: | Podpis: |

Týden webové archivace uspořádalo v červnu v Londýně šest významných institucí: British Library, School of Advanced Study, IIPC, RESAW *‒* A Research Infrastructure for the Study of Archived Web Materials, The National Archives a Archives Unleashed. Výsledkem bylo jedinečné setkání jak webových archivářů a vývojářů z celého světa, tak i výzkumníků, kteří běžně s webovými archivy pracují. Týden začal již v neděli třídenním workshopem *Archives Unleashed*, kde výzkumníci spolupracovali na vytvoření zajímavých výstupů, postavených na datech webových archivů. Obsahem jedné ze studií, prezentované na konferenci, která po workshopu následovala, bylo porovnání tří webových archivů a jejich pokrytí událostí kolem hnutí *Occupy Wall Street* v letech 2011 a 2012. Prvním ze dvou zásadních zjištění bylo, že webové archivy se příliš ve sběru dat nepřekrývají, proto s počtem webových archivů roste i pokrytí kolekce. Druhým poznatkem bylo, že je důležité archivovat URL adresy, které jsou sdíleny na sociálních sítích, jako je například Twitter, neboť často odkazují na důležité zdroje týkající se daného tématu.

Velká část příspěvků byla standardně věnována problematice webové archivace. Novým tématem se stala bezpečnost webových archivů. Sklízeče archivují obsah na webu tak, jak jej naleznou. Obsahuje-li daná adresa v okamžiku sklizně škodlivý kód, sklízeč jej poctivě za- archivuje. Většina škodlivého kódu by neměla fungo-vat, pokud běží na serveru, který nebyl napaden. Přesto je třeba bezpečnost webových archivů dále zkoumat, a z toho důvodu byl navržen repozitář github.com/harvard-lil/warcgames, kde si mohou vývojáři zkoušet mož- nosti archivačního nástroje Webrecorder, a hledat tak bezpečnostní chyby. Vedle technických příspěvků byl představen i recenzovaný časopis s otevřeným přístupem *Internet Histories*, zabývající se samotnou historií Internetu jako technologie.

Přítomnost velkého množství výzkumníků na konferenci stvrdila rostoucí důležitost webových archivů v roli primárního zdroje studia současné společnosti. Ještě před pěti lety byl výzkum ve webových archivech veden hrstkou badatelů. Dnes již univerzity vytvářejí vlastní webové archivy a začínají vyučovat metodologii práce s nimi. ebových archivů, z čehož máme skutečnou radost i motivaci do další práce.

automated watch

policy representations

automated planning

stací na PLATO Planets

preservation components

characterization components – srovnávací studie k nástrojům validace – viz výše

seznam open source nástrojů na large scale migrace

QA components

testbed

web content testbed

LSDR content testbed

science data content testbed

Sven Schlarb - testbeds

speciální platforma na sdílení workflow

poučili se s planets testbed

vývoj workflow jako shared social activity

taverna – workflow builder (university of manchaster)

workflow management v myExperiment

workflow registry – lze komentovat apod.

renderability test – zobrazit obsahy arců v prohlížečích (enginech) z browserů, pokud je problém, hlásí to = jeden z use casů (scénářů)

scenario – QA porovnání obrázků původního webu a warcu – při sklizni se udělá image webu a pak warcu – porovná se to

wiki.opf-labs.org

jeden ze scénářů bude i arc>warc

jhove2 umí kontejnery a v nich další kontejnery

dp4lib – vytvořili si vlastní nástroj na validaci (ddgo), kt. kombinuje jhove, jhove2 a droid – protože Jhove2 nemá některé moduly

**SCAPE preservation policies – Barbara**

katalog policies

průzkum dostupných policies a aktivit, chtějí je rozkuchat na kousky a na tom stavět

zdají se všechny příliš obecné

musí být více podrobnější, jako podklad pro planning tool

žádná PP na webarchiving

na základě PP udělá decision making tool rozhodnutí o způsobu ochrany

DNB –Tobias Steinke

rozhodli se dělat selective harvesting, .de je příliš velké pro komplet harvest

neměli workflow pro pravidelné harvesty

chtěli tedy postavit interní workflow – s pomocí WCT nebo netarchiveSuite, ani jeden se jim nelíbil

možná budou dělat kompletní harvest jednou za 2 roky

internetFoundation jim nabídla řešení, kt. se jim líbí víc než WCT, NAsuite i heritrix

***Kreibich + Hutař – 7.10. konference DPC v BL***

H. van der Sompel: Memento – agregované časové brány, historie přesměrování se snaží vybudovat původní kontext. V současnosti existují pluginy do firefoxe, android, iphone, též je podpora na úrovni waybacku.

eternal now

current now – past representations nejsou dostupné, URI odkazuje na aktuální verzi

past verze v cms, v cache apod.

memento umožní pomocí de-referencing odkaz na původní dokument pod uri

memento je protokol, rozšíření http

archivní verze jsou tzv. „mementa“

browser použije timegate, URI, kt. ví o starých verzích, jejich uri je v http protokolu a browser se rozhodne, kt. verzi použije (stejně jako se rozhodne, jestli zobrazí informace v angličtině nebo francouzštině apod.)

funguje to přes plugin v browseru

firefox, android, iphone,

umí s tím pracovat i wayback

plugin na mediawiki

IA funguje jako timegate v případě, že není k dispozici jiná

linky s doi se také mění, resp. doi zůstává, odkazované zdroje se mohou měnit (ne zdroje, ale splash page), takže paper zůstává stejný, kontext se mění, timegate tedy musí být před doi a vědět o jednotlivých verzích

WA – Eric Meyer: ssrn.com. Společenské vědy se nezajímají o WA, hrozba zaprášeného archivu. Zkoumat používání archivů (logy z přístupu), připravit data pro vědeckou komunitu, algoritmické spouštění kolekcí (např. dle google trends, twitter trends – podívat se jak u nás vypadá monitoring médií, zda-li jde použít k spuštění kolekce)

BL – Lewis Crawford – 3d vizualizace frontpage posloužili k spatření patterns v design webů. Tagové cloudy z klíčových slov (použít nad doménami, nad tématy). N-gramy!!! (jak je to komplikované, stačíme na to sami?) Zkontrovat Hive – relační databázi na Big data.

dělají sbírky

vyhledávání ve fulltextu

ngram

rozdělení výsledků podle typů dokumentů – např. obrázky, média apod., lze podle toho hledat od počátku

vizualizace témat a webů v gephi seadragon

začínají využívat archiv na určité analýzy, obsahu apod. – vědí co se stalo kdy, kde – dle toho lze usuzovat na různé věci (ngram, mizení webů konkrétního oboru apod.)

cesta i u nás – nabídnout z WA nějaký užitek

nový web?

IA – Kris Carpenter – Zálohování logů, aby bylo možné zpětně rekonstruovat chování uživatelů na webu. Jak logy mapovat a jak je skladovat? HBASE? Obrázky nebo videa použití – v jaké frekvenci?

Wim Peters – Strategie vytváření kolekcí co nejrychleji až poté řešit zda-li prezervovat či ne. Tj. nejdříve vychytat trend a až poté evaluovat, zda data uchovávat. Komunitní vzpomínky – možná by šlo použít sběr logů z NKP, či univerzit a poté takto používané weby sklízet.. Model pro komunitní sítě: subjekt, místo, událost.

Hanzo – Individual data is dull -> big data is future. Tj. budoucnost výzkumu je v dataminingu nad webovým universem, už teď je se dělají úspěšné prognózy z chování na webu. My ale na rozdíl od firem máme v rukách hlubší minulost, můžeme sledovat vývoj!

**obecně zaměření na access – viz prezentace IA**

projekty univerzit a jednotlivých institucí – zapojit i naše univerzity a odborné instituce?

blog forever

vývoj řešení pro DP blogů, včetně zachování kontextu a pocitu používání blogů

harvest, data model pro obsah, metadata a semantiku

výstupem by měl být sw a sklízení a archivaci blogů, kt. by udržel autenticitu, integritu blogů

**Technologie**

**Zkvalitnění zpřístupnění**

Dle zkušeností komunity se Python wayback zdá jako mnohem lepší alternativa k rožšířenějšímu OpenWayback. Dokáže lépe prezentovat obsah uložený v archivních kontejnerech. Umí využít současného CDX indexu archivních objektů. Ale nabízí i vlastní indexátor kontejnerových formatů ARC i WARC, který by mohl vytvořit úplnější index než standardní cdx-indexer, který je součástí OpenWayback. Odkaz na repozitář: <https://github.com/ikreymer/pywb>

**Bezpečnost**

Webové archivy jsou při špatné konfiguraci napadnutelné v podobných vektorech jako ostatní webové aplikace. Jde hlavně o podvržení cookies a mezidoménové podvržení. Šikovným manipulováním cookies je možné i podvrhnout zobrazení. Obecná náprava je kontejnerizace webových aplikací a oddělení prezentační domény od té obsahové. Více je možné se dozvědět v repozitáři <https://github.com/harvard-lil/warcgames>, který umožňuje testovat útoky vůče webovému archive.