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. ODiF 2.4.1 OAW	
Pracoviště - zařazení	knihovník	
Důvod cesty	konference GA IIPC 2015	
Místo - město	Palo Alto, Kalifornie	
Místo - země	**USA**	
Datum (od-do)	**25. 4. 2015 - 2. 5. 2015**	
Podrobný časový harmonogram	25. 4. 2015 - odlet z Prahy 27. 4. - 1. 5. 2015 - konference 2. 5. 2015 - návrat do Prahy	
Spolucestující z NK	Zuzana Kvašová, Rudolf Kreibich	
Finanční zajištění	VaV 0137	
Cíle cesty	Cílem cesty byla účast generálním shromáždění konsorcia IIPC, jehož je NK ČR členem. Součástí shromáždění byla konference a workshopy. Cílem bylo seznámení se s aktuálními trendy v oboru webového archivnictví.	
Plnění cílů cesty (konkrétně)	Cíle byly splněny. Byly získány poznatky, které budou moci být využity pro činnost a rozvoj českého webového archivu.	
Program a další podrobnější informace	Viz. níže	
Přivezené materiály	—	
Datum předložení zprávy	13. 5. 2015	
Podpis předkladatele zprávy		
Podpis nadřízeného	Datum: 13.5.2015	Podpis:
Vloženo na Intranet	Datum:	Podpis:
Přijato v mezinárodním oddělení	Datum: 20.5.15	Podpis:

Program valného shromáždění a konference na adrese http://netpreserve.org/general-assembly/2015/overview

Generální shromáždění probíhalo formou konference, workshopů a diskuze. Níže jsou vybrané témata a přednášky, které jsou pro náš webový archiv nejvíce přínosné.

Tématické sklizně

Tématické sklizně (archives of events) jsou speciální sklizně vytvářené pro nejrůznější události, od kterých se očekává, že jsou reflektovány na internetu.
Příklad jedné z takových sklizní
https://archive-it.org/organizations/156

Problémem je potřebná manuální příprava URL, která zpomaluje reakci webových archivů na aktuální události. Snahou ulehčení práce kurátorům je sbírání semínek ze social media na základě - klíčových slov, tagů a z nich extrahovat URL. Nevýhodou tohoto přístupu je příliš mnoho šumu mnoho.

Dále byly představeny modely pro vytváření takových sklizní - viz obrázky níže.

Iniciace pracovní skupiny pro společné API

Mezi hlavní cíle od vzniku IIPC patří vývoj snadno použitelných a kvalitních aplikací s otevřeným kódem. Za tu dobu si řada institucí osvojila různé produkční mody vývoje otevřeného SW. Je na čase zhodnotit společné zkušenosti a vybrat vhodný režim vývoje projektů IIPC. Ještě zásadnějším krokem je jasná kodifikace okruhů problémů, které různé nástroje řeší. Každý okruh problémů je možné považovat jako komponentu společného API. Pokud budou vyjasněné jednotlivé komponenty a problémy které řeší, je možné definovat rozhraní mezi komponentami. Takto se ušetří spousta práce v budoucím vývoji. Pro WA to znamená připojit se k pracovní skupině pro vytvoření API. Takto budeme moci koordinovat náš vývoj s vývojem v rámci IIPC.

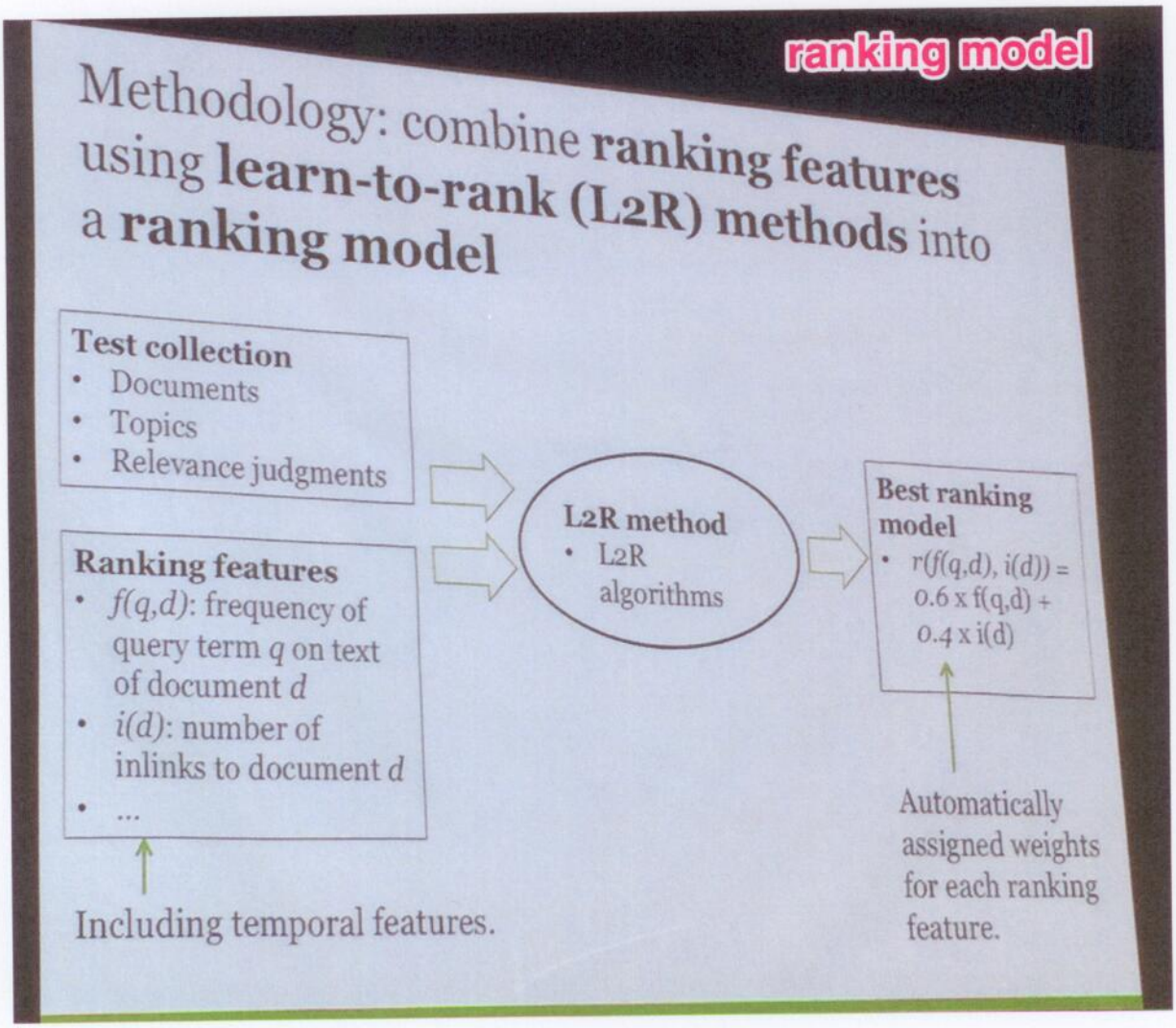
Časová koherence webových archivů

Studie Old Dominion Unverisity prokázala velkou časovou nekoherenci při zobrazování obsahu webových archivů. Při přehrávání archivovaných HTML dokumentů se projevilo, že prvky z nichž je web složen, mohou pocházet z různě starých sklizní, které od sebe může dělit i několik měsíců až let. Uživatele tak vidí podobu dokumentu, která nikdy neexistovala a je proto závádějící. Taková nekoherence snižuje relevanci webového archivu jako informačního zdroje. Projekt Memento navrhuje v rámci své služby agregovat objekty z několika webových archivů, aby čtenář dosáhl co nejlepší časové koheze jednotlivých objektů. V příštích letech z toho plynou dva náročné úkoly pro WA. Připravit podmínky pro analýzu časové koheze zobrazovaných prvků ve webovém archivu, obohatit GUI Wayback o možnost zobrazení časovosti jednotlivých prvků.

Fulltextové vyhledávání ve webových archivech

Webový archiv z Portugalska prezentoval svůj výzkum fulltextového vyhledávání nad webovými archivy. Běžné algoritmy fulltextového vyhledávání nejsou dostatečné z důvodu časové duplicity záznamů.
K vytvoření fulltextového vyhledávače předcházelo testování na uživatelích.

Jejich technologie vyhledávání je založená na Lucene extensions. A vytvořili speciální rank pro čas na základě maching learning technik. Jejich search engine je celkově složený z několika běžných algoritmů. Viz obrázek



Heritrix

Heritrix vznikl v době statického webu, jeho architektura je navíc monolitická a není optimálně škálovatelný. V současnosti Heritrix velmi složitě zvládá sklízení webu postaveného na technologii JavaScript, streamované služby aj. Není snadné vytvořit distribuovanou sklizeň a sklízení na více strojích současně nese svá úskalí. Je strategickou otázkou zda-li by IIPC mělo iniciovat vývoj nového sklízeče nebo rozvíjet Heritrix současný. Z pohledu WA vnímáme stejná omezení, ale spíše než škálovatelnost je pro nás zásadní zvládnutí archivace webů postavených na současných technologiích.

Webarchiving pro veřejnost

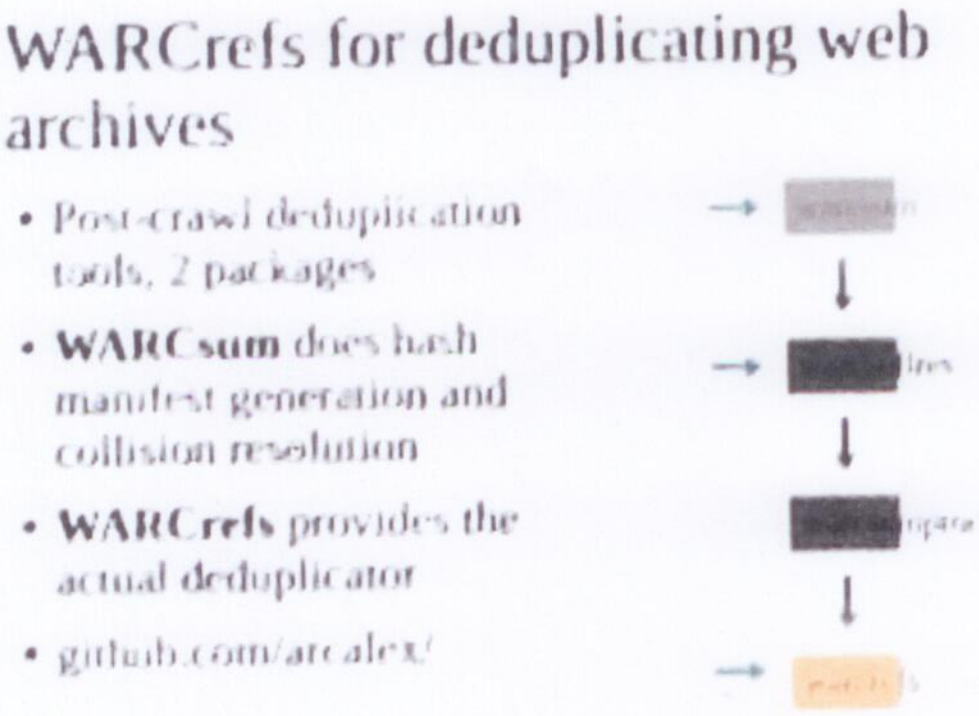
Ilya Kreymer vyvíjí webovou službu WebRecorder.io, která umožňuje veřejnosti zaznamenat průchod webem a výsledný WARC si stáhnout pro další přehrání ve stejné službě. Pro WA je zajímavé tři otázky. Za prvé jak se liší kvalita sklizně pomocí WebRecorder od sklizně pomocí Heritirx. Jaký bychom zvolili přístup, kdyby běžní uživatelé u nás WARC archivovali. Zda by nám provoz takové služby přinesl zajímavé informace o využívání webového archivu veřejnosti. URL projektu je https://github.com/webrecorder/webrecorder.

Warcbase

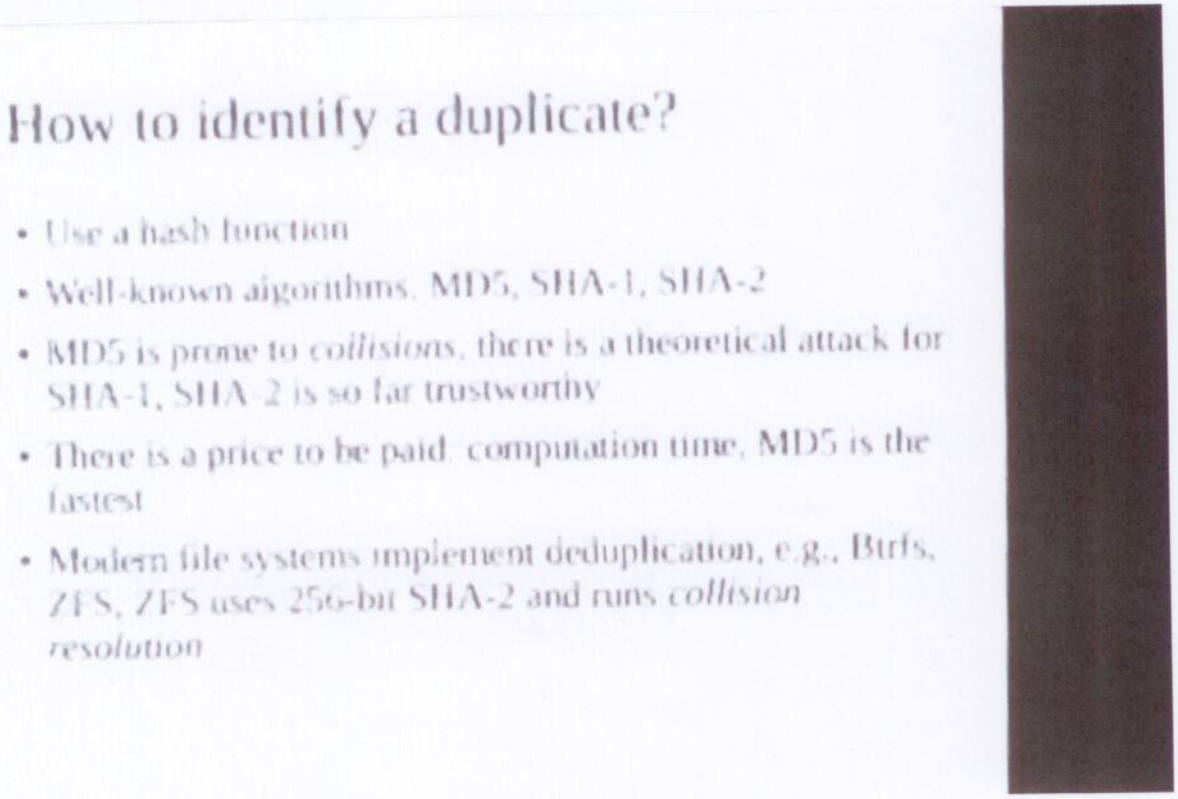
Jimmy Lin správně definuje, že využití webových archivů je nad rámcem fulltextového vyhledávání a procházení archivovaných dokumentů. Problém je, že akademičtí uživatelé nemají k dispozici nástroje, jak archiv efektivně vytěžit. Z druhé strany vývojáři zase nevědí jaké nástroje pro uživatele připravit. Jimmy Lin proto připravil platformu Warcbase (https://github.com/lintool/warcbase), které umožňuje data ingestovat do Hadoop ekosystému, nechat je zpracovat např. pomocí MapReduce nebo Spark a stejně tak, je i umožňuje zpřístupnit aplikací OpenWayback. Takového přístup je ideální pro použití ve WA NK. Úkolem pro WA v příštích letech je platformu si otestovat.

Deduplikace

Přednáška byla zaměřená na oblast deduplikace obsahu webových archivů. Byla zaměřena zejména na to, jakým způsobem poznat duplicitní obsah (viz obrázek) a jakou formou deduplikaci provádět.



Deduplikaci je možné provádět dvěma způsoby. Za prvé rovnou při sklizní (takto ji provádní český webový archiv) a nebo nad již uloženým obsah formou postprocesingu. Toto přináší výhodu nižší zátěže na crawler a údajně i větší bezpečnost při rekonstrukci obsahu.
Dále byl součástí přednášky představen nástroj WARCrefs tools, který slouží pro práci s deduplikovaným obsahem.



Generální shromáždění

Proběhlo třetí den akce na půdě Internet Archive, agenda byla prezentace ekonomických výsledků konsorcia, financování projektů a jejich výsledky. Návrhy na projekty do dalších let a diskuze nad dalším směřování konsorcia. Zároveň byla součástí exkurze po Internet Archive.