Vědecké využití dat z webových archivů

Research use of web archived data

Mgr. Jaroslav Kvasnica, Mgr. Barbora Rudišinová, Bc. Rudolf Kreibich / Národní knihovna České republiky (National Library of the Czech Republic), Klementinum 190, 110 00 Praha

Resumé:
Stále větší část komunikace i autorských děl (ať už článků, fotografií nebo jen krátkých příspěvků na blogu) se přesouvá z tradiční tištěné produkce do webového prostředí. Tento obsah je však velmi různorodý a proměnlivý, množství webového obsahu každý den vzniká, mění se, ale také zaniká. Z hlediska výzkumných a badatelských zájmů je takový obsah zcela unikátní a významný. V případě, že by tento obsah zmizel, nebude možné studovat podstatnou část nedávné historie. Získáváním a dlouhodobým uchováváním webového obsahu se zabývají webové archivy.
V dnešní době mají webové archivy uložené obrovské množství dat a hledají způsob, jak je nabídnout svým uživatelům. V článku je popsáno, jakým způsobem webové archivy data uchovávají a jaká metadata se v těchto archivech nacházejí. Potřebami a požadavky badatelů na webové archivy se zabývalo již několik výzkumů, článek shrnuje aktuální výsledky těchto prací a na jejich základě doporučuje postup pro navázání spolupráce s badateli.

Klíčová slova: webová archivace, badatelé, velká data, big data, metadata, analýza dat, WARC

Summary:
Major part of our communication and media production has moved from traditional print media into digital universe. Digital content on the web is diverse and fluid; it emerges, changes and disappears every day. Such content is unique and valuable from academic perspective but as it disappears over time, we are losing the ability to study recent history. Web archives are now taking the responsibility to capture and preserve such content for future research.
Web archives preserve vast amounts of data captured over the years and one of the main goals now is to improve research usability of their collections. This article describes the way the web archives store web content and related metadata and summarizes several recent studies that have dealt with research requirements for web archived data. Based on conclusion of these studies, it suggests further actions to establish cooperation with research community.

Keywords: web archiving, researchers, big data, metadata, data analysis, WARC

Úvod

Od vzniku webu na počátku 90. let 20. století se webové prostředí stává stále významnějším zdrojem pro studium nedávné historie a současných politických, sociálních a kulturních jevů. Je zcela odlišné od jiných typů médií a publikačních platforem, k jeho charakteristickým rysům patří neustálá proměnlivost, propojení obsahu odkazy nebo množství různých formátů (obrázky, video, aplikace atd.), a v neposlední řadě také to, že jeho obsah tvoří různorodá množina osob i společností. V současnosti hraje internetové prostředí významnou roli nejen v oblasti masové a osobní komunikace, ale i jako platforma pro vývoj nových služeb. Internet už dávno není dostupný pouze ze stolních počítačů, prostřednictvím mobilních zařízení a volně přístupných datových služeb může být dnes množství lidí „online“ v jakoukoliv denní i noční dobu. Stále větší část veřejného života se odehrává na internetu, ať už to jsou politické či sociální kampaně, výměna názorů prostřednictvím diskusních fór, názory publikované na osobních blozích, částečně veřejné informace sdílené na sociálních sítích.

V současnosti jsou již vědci a výzkumníci zvyklí běžně pracovat s online publikovanými materiály, jako jsou elektronické články, databáze nebo různé typy aplikací. Studie z roku 2014 provedená na Harvard Law School zkoumala webové odkazy v odborných právnických časopisech a ukázala, že více než 70 % těchto citací neodkazuje na správný obsah (ZITTRAIN 2014).

Z důvodu přesunu podstatné části komunikace i sociálních aktivit na web nabývá prostředí internetu, které bylo oproti vědeckým informacím zatím poměrně opomíjeným zdrojem, na významu pro výzkumníky a badatele. Web nabízí nejen množství jednotlivých informací a zdrojů, ale lze na něj nahlížet i jako na soubor velkých datových setů, na základě kterých je možné vytvářet analýzy a využít je pro výzkum. Proměnlivá povaha webu, kdy jednotlivé webové zdroje mizí, přesouvají se nebo jsou nejrůzněji upravovány a doplňovány, přináší problém právě pro výzkumníky a badatele, kteří chtějí pracovat s webovými zdroji staršího data. S tímto problémem se pak musejí obrátit na webové archivy, jež se snaží unikátní a dynamický webový obsah zachytit a uchovat v čase.

U archivovaných webových zdrojů můžeme najít určité specifické rysy a odchylky od online zdrojů. Webová archivace je vždy archivací v *reálném čase.* Aby mohl být obsah, který se dnes nachází na webu, zaznamenán, musí být archivován co nejdříve, než se změní nebo zanikne. Poté už by nebylo možné ho jakkoliv získat. Archivovaný webový obsah není pouze kopií toho, co bylo v minulosti online, je to *unikátní verze*, která však obsahuje „mezery“ *v obsahu*.

Akvizice webových zdrojů

Při archivaci webu je nutné přijmout mnohá rozhodnutí, která determinují, jakou podobu bude mít obsah, který je nakonec v archivu uložen.

Z technického hlediska je to volba použitého softwaru pro archivaci. Každý současný software pro archivaci s sebou nese jistá technická omezení, která předurčují, jaké části obsahu budou chybět, tzn., u kterých použitý software nedokáže vytvořit kopii. Ale i volba hardware, zejména úložné kapacity a výpočetní výkon, zásadně ovlivňuje konečnou podobu dat uložených v archivu. Je zřejmé, že úložné kapacity webového archivu podmiňují počet archivních kopií, které je archiv schopen uložit. Síla výpočetního výkonu pak ovlivňuje rychlost a efektivitu archivace.

S tím souvisí volba rozsahu vlastní archivace u jednotlivých webových zdrojů, kterou lze omezit buď horizontálně, nebo vertikálně. U horizontálního omezení mluvíme o počtu stažených navazujících stránek, které jsou kontextem pro původní webový zdroj, např. u citací v článku je potřeba stáhnout i webové zdroje, na které je odkazováno, aby byl do budoucna zachován celý kontext. Vertikální omezení archivace pak definuje, kolik digitálních objektů je umožněno archivovat na jednu doménu, zde mluvíme o tzv. hloubce sklízení. Dalším technickým faktorem, který ovlivňuje podobu webového obsahu, je také platforma a zařízení, na němž obsah zobrazujeme. Na různých zařízeních, v různých prohlížečích nebo s různými aplikacemi se může výsledná podoba obsahu lišit.

Webové archivy se snaží reagovat na neustálý vývoj technologií, které se objevují na webu, jako jsou například dynamické prvky stránek, javascript, propojení se sociálními sítěmi, integrovaná videa či různé aplikace. Z důvodu rychlého vývoje webových technologií jsou ovšem webové archivy obvykle pozadu za tímto vývojem, a to opět vede k nedostatkům v obsahu archivovaného webu oproti jeho originální online podobě (BRÜGGER a FINNEMANN 2013).

Z obsahového hlediska je to definice teritoriality, která je typická pro evropské webové archivy, jež jsou většinou součástí národních knihoven. Provádějí tzv. celoplošnou archivaci na základě národního charakteru webů, obvykle z technických důvodů definují teritorialitu národní doménou (např. „.cz“), ale opomíjejí národní zdroje na jiných či me-

zinárodních doménách (např. „.eu“, „.org“), tedy kompletní národní webový obsah (např. v České republice se jedná o bohemikální obsah).

S tím, jaké zdroje budou archivovány, obecně souvisí to, jaký typ sklizně je realizován. Již byla zmíněna *celoplošná sklizeň*[1], která představuje archivaci webových zdrojů v co největším možném rozsahu, bez jejich předchozího manuálního výběru. Dalším typem sklizně jsou pak sklizně výběrové, u kterých jsou zdroje vybírány a hodnoceny kurátory na základě předem určených kritérií. *Výběrové sklizně* bývají zpravidla svým rozsahem menší, ale zároveň jsou webové zdroje sklízeny více do hloubky.

V případě archivovaného webového obsahu se tedy nejedná o identickou kopii originálních online webových zdrojů, ale jde spíše o obsah, znovu vytvořený na základě tohoto originálního online obsahu.

Webový archiv je *multičasový a multiprostorový*. Na rozdíl od online webu, kdy v daném čase existuje pouze jedna verze obsahu, webový archiv obsahuje množství verzí webového zdroje v různých časových bodech. U některých rozsáhlých webových zdrojů může jejich sklízení trvat delší dobu, a tak může dojít ke změnám v obsahu v průběhu archivace (typicky u velkých internetových deníků se stává, že během jejich sklízení vycházejí nové články nebo jsou aktualizovány ty staré), proto je výsledná archivní verze mozaikou různých částí obsahu (i co se rozsahu týká). To vše samozřejmě přináší komplikace badatelům při výzkumu, proto je klíčovou součástí samotných dat jejich metadatový popis. Díky metadatovému popisu mohou badatelé najít a definovat takovou množinu dat, která je nejvíce relevantní pro jejich výzkum.

Kontejnerový formát WARC

Webové archivy pro ukládání archivních kopií webových stránek využívají specializovaných kontejnerových formátů, které umožňují agregaci jednotlivých fragmentů staženého webového obsahu. K tomuto řešení musely webové archivy přistoupit zejména kvůli složité manipulaci s enormním množstvím malých souborů, ze kterých jsou jednotlivé webové stránky složeny. Webové stránky se běžně skládají ze stovek až tisíců malých souborů: skriptů, obrázků, videí apod.

Kontejnerové formáty spojují několik digitálních zdrojů do jednoho agregovaného archivního souboru spolu se souvisejícími informacemi (BAILEY a LACALLE, 2015). Tyto formáty jsou určeny k tomu, aby umožnily všechny soubory uložit, manipulovat s nimi jako s jedním celkem, a zároveň zachovaly za pomoci metadat jejich vzájemnou provázanost. Díky tomu je možné webovou stránku zpětně z fragmentů rekonstruovat a opětovně ji zobrazovat v původní formě uživateli. Pro badatele, který chce využít ke svému výzkumu data z webového archivu, představují metadata možnost prvotní analýzy archivu, aniž by musel pracovat s enormním množstvím dat jako takových.

Princip uložení dat a metadat v kontejnerovém formátu spočívá v tom, že každému datovému objektu předchází hlavička s metadaty. Datovým objektem mohou být buď samotné soubory stažené z webových stránek, nebo metadatové záznamy. Těchto hlaviček je celkem osm druhů, jejich výčet a využití jsou standardizovány právě v normě ISO týkající se formátu WARC (ISO 28500:2009 2009).

1. „warcinfo“ – Hlavička „warcinfo“ slouží k popisu celého obsahu, který následuje po této hlavičce a je ohraničen dalším výskytem hlavičky „warcinfo“. Nejčastěji se využívá pro popis celého kontejneru.

[1] Sklizně jsou procesy automatického stahování a sběru dat z vybraných webových zdrojů (vytváření kopií).

2. „request“ – Uvozuje datový objekt, který obsahuje kompletní dotaz na soubor, jenž je odesílán na server. Slouží payloadu, ve kterém jsou informace o tom, jak vypadal požadavek o stažení webového obsahu, který crawler[2] poslal na cílový server.
3. „resource“ – Po této hlavičce následuje samotný obsah stažený ze serveru, např. HTML soubor nebo obrázek.
4. „response“ – Využívá se pro popis odpovědi serveru při stahování dat. K této hlavičce je totiž uložena odpověď serveru, na který byl zaslán požadavek. V případě, že je požadavek přijat pozitivně (server odpovídá, na URL se nacházejí data), následuje samotný webový obsah.
5. „metadata“ – Tato hlavička slouží pro metadatový popis, který není možné přiřadit k žádné jiné hlavičce.
6. „revisit“ – Využívá se pro obsah, který již byl dříve archivován, v současné době je záznam označovaný jako revisit určen zejména k odkazování na duplikáty.
7. „conversion“ – Hlavička pro alternativní verze obsahu, který je vytvořen například při formátové migraci.
8. „continuation“ – Pro rozdělené datové objekty, k rozdělení dochází při dosažení velikostního limitu kontejneru.

Příklad hlavičky

Následující příklad obsahuje hlavičku typu typ „request“ z testovací sklizně, která byla provedena pro účely tohoto článku. Každý typ hlavičky má určitý počet shodných povinných elementů. V příkladu jsou zvýrazněny tučně a obsahují informace o typu obsahu (WARC-Type), času vzniku (WARC-Date) a velikosti obsahu (Content-Length) a jedinečný identifikátor (WARC-Record-ID), jehož pomocí lze na obsah odkazovat.

V hlavičce jsou kromě povinných elementů navíc další, jako je URL, na které byl poslán požadavek (WARC-Target-URI), a také kontrolní hash pro payload (WARC-Block-Digest), který slouží pro budoucí ověření autenticity nebo poškození obsahu.

Samotný payload začíná na řádku GET /HTTP/1.1; z tohoto řádku lze určit, že crawler poslal požadavek GET pomocí protokolu HTTP 1.1, což je standardní požadavek pro získání informací v prostředí internetu.

Příklad hlavičky „request“ a jejího payloadu:

WARC/1.0
WARC-Type: request
WARC-Record-ID: <urn:uuid:27087c01-98ae-4aca-b52c-55f33dbcb8b0>
WARC-Date: 2016-08-03T12:18:35Z
WARC-Target-URI: http://webarchiv.cz/
WARC-Concurrent-To: <urn:uuid:26300322-d4d0-4a81-8851-086f811b0a24>
WARC-Block-Digest: sha1:a514938cef4b9c3b3a88403c4ccdedd3863a74db
Content-Type: application/http;msgtype=request
Content-Length: 451
GET / HTTP/1.1
Accept: text/html,application/xhtml+xml,application/xml;q=0.9,image/webp
User-agent: Mozilla/5.0 (Macintosh; Intel Mac OS X 10_11_6) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/51.0.2704.103 Safari/537.36
(…)
Host: webarchiv.cz
Accept-language: cs-CZ,cs;q=0.8,sk;q=0.6

[2] Crawler (česky sklízeč) je nástroj pro automatické stahování a sběr dat. Pomocí crawleru jsou realizovány sklizně.

Veškerý obsah, který je uložen v kontejnerech WARC, je členěn přesně tak, jak bylo ukázáno na příkladu. Ať se jedná o jakýkoliv druh webového obsahu, vždy mu předchází záznam komunikace se serverem, a to včetně záznamu chybových požadavků nebo záporných odpovědí serveru. Tato metadata původně vznikla pro potřeby webových archivů. Díky těmto metadatům je také možné zpětně zjistit, kdy nastala případná chyba, proč není stránka správně v archivu uložená, a také, jakým způsobem chyba vznikla. Díky zaznamenanému času vzniku metadat, jejich velikosti nebo komunikaci klient-server pak mohou tato metadata sekundárně sloužit badatelům, kteří chtějí s webovými archivy pracovat.

Datové sety pro badatele

Velkou nevýhodou dat z webových archivů je to, že jsou příliš „velká“. Badatelé mnohdy nemají k dispozici potřebnou výpočetní infrastrukturu, na které by mohli data analyzovat, ale dokonce ani si je přenést z archivu. Z toho důvodu jsou generovány a vytvářeny speciální derivované datové sety určené pro badatele, neboť se ukazuje, že badatelé nepotřebují vždy veškerý obsah, který webové archivy nabízejí; např. pro lingvistickou analýzu textu nejsou potřeba audiovizuální materiály, pro hypertextovou analýzu stačí jen metadata apod.

Datové sety mohou být generovány jak z metadat, tak i ze samotných dat. Tyto sety pak představují jen tu část, která je nutná k saturaci konkrétní informační potřeby uživatele. Uživatelé si tyto sety mohou stáhnout a zpracovat na vlastní běžné výpočetní infrastruktuře pomocí volně dostupných nástrojů.

Na půdě Internet Archive byly definovány tři základní datové sety, které poskytují prvotní vhled do webových archivů a zároveň ukazují, jakým způsobem k definici datových setů přistupovat. Díky volně dostupné specifikaci jednotlivých datových setů a možnosti generovat a zpracovávat je volně dostupnými nástroji můžeme považovat tyto tři typy za určitý standard pro výzkum ve webových archivech.

Základní datový set pro výzkum dat z webových archivů

WAT je akronym pro *Web Archive Transformation*. Datový set WAT obsahuje základní informace o uložených digitálních objektech, jako je čas jejich stažení, velikost nebo např. formát, který sklízeči sdělil HTTP server. Vedle těchto údajů set WAT obsahuje informace získané vytěžením samotných digitálních objektů, jako jsou jméno autora nebo název dokumentu uvedené v metadatové části zdroje, případně všechny odkazy (URL) nalezené v digitálním objektu.

Pro vytváření datových setů WAT se používá volně dostupný nástroj *archive-metadata-extractor*, který kromě uložení metadat zpracovává i samotný digitální objekt a extrahuje informace např. z metadatové hlavičky HTML nebo odkazy z těla dokumentu. Zjednodušeně je možné říci, že WAT je archivním kontejnerem WARC redukovaným o archivní digitální obsah. Výsledkem je textový soubor obsahující metadata, která jsou organizována do struktur, a díky tomu nový soubor zabírá jen zhruba 5 % – 20 % z původního souboru WARC (BAILEY 2016a).

Datové sety WAT, díky své jednoduchosti, relativně malé velikosti a jasně definované struktuře, jsou snadno zpracovatelné a použitelné pro badatele, a to i pomocí běžně dostupných nástrojů.

Příklad části datového setu WAT:

{
„Envelope“: {
„Format“: „WARC“,
„Payload-Metadata“: {}
“WARC-Header-Length”: “298”,
“WARC-Header-Metadata”: {}
}

Datový set pro analýzu odkazování archivovaných dat v čase

Druhý datový set specifikovaný v Internet Archive nese název *Longitudinal Graph Analysis* (LGA). Obsahuje data potřebná k vytvoření grafu odkazů v čase. Díky setu je tak možné sledovat vývoj odkazování mezi doménami v čase, a to jak v šíři celého archivu, tak jen mezi vybranými doménami. Dataset LGA je tvořen dvěma soubory.

Každé URL je nejdříve zapsáno do souboru ID-Map, kde navíc získá jedinečný identifikátor a tzv. SURT[3] zápis URL (BRAGG a ODELL 2010). Poté je vytvořen soubor ID-GRAPH, který obsahuje časové razítko, jedinečné ID zdrojového URL a všechna URL, které byly z daného URL odkazovány.

Příklad zápisu datového setu LGA:
ID-Map:

{„url“:“https://www.youtube.com/watch?v=--FDzShdFjw&gl=US&hl=en“,“surt_url“:“com,youtube)/watch?gl=us&hl=en&v=–fdzshdfjw“,“id“:294869}

ID-Graph:

{“timestamp“:“20150206180648“,“id“:294870,“outlink_ids“:[62596,110007,129599,145417,148627,215031,277534,277535,277668,277678,277679,737423,737436,737459,737476,737483,737503,737514,803348,803349,803373,803374,803490,803565,803586,803590]}

Datový set se jmennými entitami

Název datového setu WANE je akronymem pro *Web Archive Named Entities*. Set obsahuje seznam lidí, míst a organizací zmíněných v digitálním objektu. Extrakce entit se realizuje nástrojem Stanford Named Entity Recognizer (Stanford NER). Výsledkem je soubor, ve kterém jsou data strukturována obdobně jako u WAT. Výsledná velikost souborů WANE odpovídá zhruba 1 % velikosti analyzovaných dat.

K vytvoření datových setů WAT a LGA stačí jen relativně jednoduchá extrakce informací ze strukturované části zdrojových dat, případně jen správné rozpoznání URL v dokumentech. Pro potřeby datového setu WANE je nutný klasifikátor, který v nestrukturovaných textech odhalí jednotlivé entity a správně je rozřadí. Nástroj Stanford NER obsahuje standardně několik klasifikátorů, které rozpoznávají entity v anglickém a německém textu. Některé z klasifikátorů umí dokonce najít entity pro čas, peníze apod.

Bohužel při zpracování českého textu se kvalita klasifikace devalvuje pod únosnou mez. Je proto potřeba použít vhodný klasifikátor pro rozčlenění dokumentů podle jazyků a pro každý jazyk pak použít specializovaný klasifikátor.

Pro český jazyk se jako ideální jeví nástroj NameTag, který využívá *Czech Named Entity Corpus 2.0 (CNEC 2.0)*, jenž dosáhl výsledky srovnatelné s výsledky klasifikátorů

[3] SURT je zkratka pro Sort-friendly URI reordering Transform, což je transformace aplikovaná na URI, která umožňuje, aby reprezentace URI lépe odpovídala přirozené hierarchii doménových jmen.

pro anglický jazyk (STRAKOVÁ et al. 2013). CNEC 2.0 obsahuje bohatou klasifikaci českých výrazů, jež rozlišuje různé formy časového vyjádření, člení jména osob na křestní, příjmení nebo jména mytologická. Dále rozpoznává entity, jako jsou filmy, knihy, normy, výrobky, jednotky měření nebo měny. Rozlišuje mezi veřejnými, vládními a soukromými institucemi. Rozpoznává různé rozhlasové a televizní stanice, e-mailové adresy a odkazy. Místní údaje člení na kontinenty, ulice, řeky apod. Identifikuje také adresy nebo telefonní čísla. Oproti standardnímu záznamu WANE však při použití CNEC 2.0 narůstá jak bitová velikost výsledného datového setu, tak i jeho komplexnost.

Jelikož současný WANE definuje jen místa, osoby a instituce, jako nejvhodnější postup se nám jeví úprava záznamu WANE podle taxonomie CNEC 2.0, kdy vytvoříme novou úroveň členění. Otázkou zůstává, zda výsledný datový set stále nazývat WANE, či již v názvu vzniklého souboru nezmínit užitou ontologii entit, např. cnec2-wane. Tuto otázku bude vhodné diskutovat v rámci konsorcia IIPC a najít jednotný způsob pojmenování setů se jmennými entitami.

Příklad záznamu pro místo ve WANE:
„locations“:[„Miami“,“Virginia“,“Fort Lauderdale“,“Wash.“,“Va.“,“Blacksburg“,“St. Louis“,“Clayton“,“-Fla.“,“Chapel Hill“,“Michigan State“,“Austin“,“North Carolina“,“Michigan“],

Předpokladaný zápis ve WANE z klasifikátoru CNEC 2.0

“Geographical_names”:{“streets, squares”:[“Václavské náměstí”,”Kaprova”],”castles/chateus”:[“Hradčany”,”Karlštejn”]}

Tyto tři datové sety rozhodně nejsou dostačující pro pokrytí veškerých informačních potřeb, které mohou mít potenciální výzkumníci. Poskytují však základní vhled do datových archivů, přinášejí případy užití dat webových archivů a zároveň slouží i jako příklad toho, jak by datový set měl vypadat.

Badatelé a jejich potřeby

S narůstajícími nároky na webové archivy ze strany badatelů (výzkumníků nebo vědců), kteří požadují plný přístup k samotným datům archivů a ne pouze přístup na uživatelské úrovni, se webové archivy musely začít zabývat těmito konkrétními potřebami vědců. Tak se objevily problémy s formou doručení dat, s hardwarovými a softwarovými požadavky a nároky na počítačovou gramotnost.

Jako hlavní problém se ukázalo, že badatelé nejsou schopni definovat, jakou část webového archivu chtějí. Badatelé často chtějí všechna data, i když je ke svému výzkumu nepotřebují (BAILEY 2016b). Tím vzniká další problém, tentokrát spojený s technickou vybaveností badatelů, kteří nemají dostatečné kapacity k převzetí dat v objemu desítek až stovek terabajtů. Na druhou stranu však ani webové archivy neumějí svá data nabídnout tak, aby s nimi mohli badatelé jednoduše pracovat. Nemají svá data pro toto využití dostatečně zpracovaná a nemají pro své uživatele potřebné nástroje.

Webové archivy se nyní snaží klást větší důraz na zvýšení míry komunikace a navázání spolupráce s badateli, na vývoj technologií, budování lokálních počítačových kapacit a na školení a rozvoj dovedností pro práci s datovými sety (TRUMAN 2016). Avšak bez znalosti svých potenciálních uživatelů a jejich potřeb nelze vyvinout odpovídající nástroje. Proto se začínají objevovat výzkumy zaměřené na toto téma. Např. Jefferson Bailey z Internet Archive se pokusil definovat první typologii zájmů uživatelů. Díky ní si mohou tvůrci webových archivů udělat představu o tom, co badatele zajímá, a případně začít pracovat na nástrojích a způsobech, jak jim svá data poskytnout.

Typologie zájmu badatelů (BAILEY 2016b):

1. dokumentární – badatelé se zájmem o klasifikaci nebo evidenci webových stránek. Spadají sem i výzkumy týkající se *legálnosti obsahu na webových stránkách* apod.;
2. sociální a političtí badatelé – sociologické a politologické výzkumy. Práce s vládními daty nebo open daty;
3. web Science – výzkum internetových technologií a protokolů;
4. Digital Humanities – do této kategorie jsou zařazeni novodobí historici a další humanitní badatelé, kteří pracují s digitálními daty;
5. počítačové vědy (Computer Science) – získávání informací, zpracování dat a indexování, infrastruktura a nástroje;
6. datová analytika (Data Analysts) – nejrůznější aktivity typu data mining, zpracovávání jazyka a výzkum trendů na internetu.

Tato typologie přináší základní pohled na možné oblasti výzkumu nad daty z webových archivů. Badatelé z různých oblastí přicházejí s jinými požadavky – nejen na to, jaká data potřebují, ale také, jakým způsobem či v jakém formátu chtějí mít k datům přístup. Potřeba různých způsobů doručení dat k badateli je ovlivněna faktory, jako jsou druh nástrojů využívaných v oboru a úroveň počítačové gramotnosti. Například sociologové využívají ke svému bádání jiné nástroje než informační či IT specialisté a zároveň můžeme předpokládat, že i jejich počítačová gramotnost se liší. Zatímco IT specialisté si sami mohou napsat skripty, které jim pomohou data zpracovat, pak například pro některé humanitní vědce je potřeba vytvořit konkrétní nástroje, které budou moci jednoduše použít.

Výzkumy potřeb badatelů

Na téma využití webových archivů badateli a na téma informačních potřeb badatelů vztahujících se k webovým archivům bylo provedeno již několik výzkumů. Pro tento článek jsme vybrali a shrnuli relevantní poznatky ze dvou aktuálních průzkumů, které vzešly z řad členů konsorcia IIPC, jehož je český webový archiv také členem, a na jejichž základě spolupracující webové archivy připravují své další kroky v této oblasti.

Poznatky francouzského výzkumu realizovaného Národní knihovnou Francie prezentovali ve svém článku Peter Stirling, Philippe Chevallier a Gildas Illien (2012). Jednalo se o kvalitativní studii provedenou v letech 2010 a 2011, kdy byly dotazovány tři typy uživatelů, včetně skupiny výzkumníků. Tato skupina se skládala z pěti odborníků z různých oblastí bádání (historie, filozofie, sociologie a informační technologie).

Kromě běžné praxe využití informací v článcích publikovaných online a s tím souvisejícím problémem jejich citování, kdy jsou články na webu často přemístěny nebo URL adresy zanikají, výzkumníci vyjádřili také potřebu využití webových zdrojů pro dokazování nebo ilustraci určitých sociologických či historických jevů a s tím spojené potíže, jako nutnost odůvodnění a dokumentování využití vybraných webových stránek ve vědecké práci. V komunitě vědců jsou data získaná z webových zdrojů stále pokládána za nedůvěryhodná a pomíjivá. V této souvislosti vyslovili dotazovaní badatelé potřebu častého sběru dat z webových zdrojů, které je zajímají.

Podstatnými tématy pro výzkumníky z hlediska archivace webu je význam archivace pomíjivých a proměnlivých dat a také význam typu obsahu (např. byla vyjádřena potřeba archivace blogů), legitimita využití dat z archivovaných elektronických zdrojů a etické otázky s tím spojené (např. osobní údaje), jakož i způsob nahlížení na archivaci webu s důrazem na zachycení jeho současné dynamické povahy.

Výzkumníci především souhlasili s aktuálním postupem Národní knihovny Francie při archivaci, tedy s kombinací automatizovaných sklizní se širokým záběrem (např.

národní domény) a menších výběrových sbírek. Projevili ale také přání, aby archivace více reagovala na nové trendy na webu (populární stránky, obsah ze sociálních sítí atd.). Dalším požadavkem badatelů bylo zveřejnění a dokumentace kritérií pro výběr zdrojů do webového archivu. V neposlední řadě také výzkumníci zdůraznili potřebu spolupráce a zapojení badatelské společnosti do vytváření metodologie a rozhodování o výběru zdrojů k archivaci.

K jednomu z nejnovějších výzkumů na téma povědomí o webových archivech a jejich využití akademickou veřejností patří průzkum Národní knihovny Nového Zélandu provedený v roce 2015 (RILEY a CROOKSTON). Účelem tohoto projektu bylo lepší pochopení možností využívání webových archivů vědci a také zjištění směru, kterým by se měl webový archiv dále ubírat. Výzkum ukázal, že méně než polovina dotazovaných výzkumníků (39 %, 113 z celkových 290 respondentů) má povědomí o tom, že existují mezinárodní iniciativy týkající se archivace webu, a ještě menší počet výzkumníků (23 %) ví, že Národní knihovna Nového Zélandu archivuje národní webové zdroje.

Výzkumníci by preferovali vyhledávání pomocí full-textu, zatímco vyhledávání pomocí URL adresy požadovaného zdroje v knihovním katalogu uvedli jako nejméně oblíbený způsob vyhledávání v archivu. Co se týká poskytování datových setů pro vlastní výzkum, dotazovaní vědci si nebyli zcela jistí, zda by tato data využili pro svůj výzkum (44 % výzkumníků by datové sety ve svém výzkumu využilo, ale 38 % se nedokázalo aktuálně rozhodnout).

Část výzkumu byla zaměřena na využívání webových archivů a archivovaných zdrojů ve výuce, v níž bylo zaznamenáno jako poměrně časté (34 % respondentů z těch, kteří pracují se zahraničními webovými archivy, je využívá také pro výuku). Nejčastějšími způsoby využití archivovaných zdrojů je jejich citování (poskytnutí odkazu), využití k ilustraci formou screenshotů nebo jako informační zdroje pro studenty.

Z hlediska obsahu označili dotazovaní výzkumníci jako nejpodstatnější zdroje pro výzkum vládní webové stránky (66 %). Významné množství respondentů také uvedlo, že by archivace obsahu sociálních sítí mohla být pro jejich činnost přínosná, za nejdůležitější typ sociálních médií pro výzkum uvedli sítě pro sdílení videí (např. Youtube, Vimeo) a diskusní fóra.

Velká část respondentů souhlasí s tím, že je důležité archivovat národní webové zdroje (77 %). Polovina všech dotázaných (51 %) také uvedla, že data z webového archivu budou významná pro jejich výzkum v horizontu pěti let. To kontrastuje s průzkumem Národní knihovny Francie z roku 2012, kdy dotázaní výzkumníci neměli aktuální potřebu využití dat z webových archivů, spíše předpokládali větší využití těchto dat svými studenty (STIRLING et al. 2012).

Shrnutí poznatků z výzkumů:

- Data z webového prostředí jsou stále vnímána vědeckou společností jako poměrně nespolehlivá a pomíjivá,
- mnoho vědců stále nemá povědomí o webových archivech, netuší, jak tato data využít pro svůj výzkum,
- většina vědců však souhlasí s významem a přínosem webové archivace v oblasti výzkumu,
- vědci stále spíše upřednostňují tradiční přístup k archivovaným datům (prohlížení jednotlivých zdrojů),
- silný požadavek na full-textové vyhledávání,
- vědci mají o datové sety většinou zájem, ale často nevědí, jak je pak využít, nemají představu o tom, jak velký vzorek vybrat, jak obsahově vymezit (mají pocit, že by bylo vhodné mít co nejvíce dat),
- vyjádřen zájem o archivaci sociálních sítí,

- potřeba vyřešení legislativních a etických otázek při využívání elektronických materiálů z webových archivů.

Závěr

Webové archivy byly donedávna přehlíženy ze strany badatelů, přestože jde o platformy, které se snaží zachovat v dnešní době nejvýznamnější, nejrozsáhlejší a nejrychleji rostoucí kulturní dědictví. S přibývajícím zájmem o archivovaný webový obsah však samotné archivy zjišťují, že nevědí, jak s badateli spolupracovat nebo komunikovat tak, aby to bylo prospěšné pro obě strany.

Poznatky z realizovaných průzkumů potřeb badatelů ukázaly některé jejich požadavky na webové archivy. Aby se mohly webové archivy lépe připravit na spolupráci s badatelskou komunitou, měly by se zaměřit na následující činnosti:

- Propagace a navázání spolupráce (badatelé nevědí o webových archivech),
- vysvětlování účelu archivace a jejího smyslu, zdůrazňování významu a relevance dat z webových archivů (vědci stále preferují klasické typy zdrojů),
- dokumentace archivních sbírek, vytvoření standardů pro popis a tvorba kvalitního popisu a metadat, aby bylo jasné, co se ve sbírkách nachází a proč (není to kompletní obraz internetu, takový, jaký byl na živém webu, některé části chybějí),
- pro vědeckou práci je nezbytné vědět, s jakými daty se pracuje, vhodné je vytvářet např. i krátké anotace k jednotlivým kolekcím,
- snažit se umožnit full-textové vyhledávání v archivu,
- vytvářet a poskytovat datové sety – přestože mají badatelé pocit, že by chtěli co největší objemy dat, je pro ně naopak třeba připravit nástroje, sestavovat menší kolekce zaměřené tematicky, časově ohraničené atd. (nutná dokumentace, proč byly právě tyto zdroje vybrány a jak byly vybrány, definovaná metadata a formáty),
- snažit se archivovat i méně běžné typy zdrojů (např. různé typy multimédií) a sociální sítě,
- vyřešit a jasně definovat oprávnění a podmínky využití dat z webových archivů, klást důraz na ochranu osobních údajů,
- navázat spolupráci s jinými webovými archivy, dosáhnout vzájemné součinnosti (web je svým charakterem mezinárodní, proto i výzkumy nad daty z webu budou mít převážně mezinárodní rozsah),
- požádat badatele o zpětnou vazbu při vytváření sbírek.

Pokud si chce lidská společnost i do budoucna zachovat schopnost psát a studovat svoji historii, bude k tomu neoddiskutovatelně potřebovat webové archivy. A webové archivy musí udělat všechno pro to, aby se s touto rolí úspěšně vypořádaly. Stejně jako se vědci humanitních oborů musí naučit pracovat s digitálními materiály, tak se webové archivy musí naučit spolupracovat s ostatními obory.

Tento článek vznikl díky podpoře poskytnuté Ministerstvem kultury ČR pro rozvoj Národní knihovny České republiky jako výzkumné organizace.

Použitá literatura

BAILEY, Jefferson, 2016a. *WAT Overview and Technical Details* [online]. 2016 [cit. 2016-08-10]. Dostupné z: *https://webarchive.jira.com/wiki/display/ARS/WAT+Overview+and+Technical+Details.*

BAILEY, Jefferson, 2016b. *Program Models for Research Services* [online]. 2016 [cit. 2016-08-10]. Dostupné z: *http://netpreserve.org/sites/default/files/WAC02_HEKLA_Jefferson_Bailey.pdf.*

BAILEY, Jefferson a Marie LACALLE. *Don't WARC Away: Preservation Metadata & Web Archives* [online]. 2015 [cit. 2016-08-10]. Dostupné z: *http://connect.ala.org/files/2015-06-27_ALCTS_PARS_PMIG_web_archives.pdf.*

BRAGG, Molly a Kate ODELL. *SURT Rules* [online]. 2010 [cit. 2016-08-10]. Dostupné z: *https://webarchive.jira.com/wiki/display/ARIH/SURT+Rules.*

BRÜGGER, Niels a Niels Ole FINNEMANN. The Web and digital humanities: Theoretical and methodological concerns. *Journal of Broadcasting & Electronic Media* [online]. 2013, s. 66-80 [cit. 2016-04-28]. ISSN 1550-6878. Dostupné z: *http://dx.doi.org/10.1080/08838151.2012.761699.*

ISO 28500:2009: Information and documentation -- WARC file format. Londýn, 2009.

RILEY, Harriet a Mark CROOKSTON. Awareness and use of the New Zealand web archive: a survey of New Zealand academics [online]. National Library of New Zealand, 2015 [cit. 2016-04-28]. Dostupné z: *http://apo.org.au/node/58430.*

STRAKOVÁ, Jana, Milan STRAKA a Jan HAJIČ. A New State-of-The-Art Czech Named Entity Recognizer [online]. s. 68 [cit. 2016-08-10]. DOI: 10.1007/978-3-642-40585-3_10. Dostupné z: *http://link.springer.com/10.1007/978-3-642-40585-3_10.*

STIRLING, Peter, Philippe CHEVALLIER a Gildas ILLIEN. Web Archives for Researchers: Representations, Expectations and Potential Uses. *D-Lib Magazine* [online]. 2012, 18(3/4) [cit. 2016-04-28]. DOI: 10.1045/march2012-stirling. ISSN 1082-9873. Dostupné z: *http://www.dlib.org/dlib/march12/stirling/03stirling.html.*

TRUMAN, Gail. Web Archiving Environmental Scan: Harvard Library Report. *Digital Access to Scholarship at Harvard* [online]. USA: Harvard Library, 2016 [cit. 2016-08-10]. Dostupné z: *http://nrs.harvard.edu/urn-3:HUL.InstRepos:25658314.*

ZITTRAIN, Jonathan, Kendra ALBERT a Lawrence LESSIG. Perma: Scoping and Addressing the Problem of Link and Reference Rot in Legal Citations. Legal Information Management [online]. 2014,**14**(02), 88–99 [cit. 2016-07-25]. DOI: 10.1017/S1472669614000255. ISSN 1472-6696. Dostupné z: *http://www.journals.cambridge.org/abstract_S1472669614000255.*

KVASNICA, Jaroslav. Vědecké využití dat z webových archivů. *Knihovna: knihovnická revue*. 2016, **27**(2), 23–34. ISSN 1801-3252.