Odborný článek

ProInflow: časopis pro informační vědy
Vol. 11, No. 1, 2019, str. 3–21
DOI: https://doi.org/10.5817/ProIn2019-1-2

http://www.phil.muni.cz/journals/index.php/proinflow

ANALÝZA ČESKÉHO WEBOVÉHO ARCHIVU: PROVENIENCE, AUTENTICITA A TECHNICKÉ PARAMETRY

ANALYSIS OF THE CZECH WEB ARCHIVE: PROVENANCE, AUTHENTICITY AND TECHNICAL PARAMETERS

Jaroslav Kvasnica, Andrea Prokopová, Zdenko Vozár, Zuzana Kvašová

Webarchiv Národní knihovny ČR

Abstrakt

Účel – Článek poskytuje přehled možných vstupních kritérií při archivaci webových stránek webovými archivy a popisuje, jaký dopad může mít jejich nastavení na výsledná archivní data v rovině obsahové, formátové a technické. Nastavení vstupních parametrů při webové archivaci přímo determinuje výslednou podobu archivního obsahu a v případě realizování výzkumu nad těmito daty badatelé potřebují znát jejich provenienci. Bez těchto znalostí není pro badatele možné pracovat s archivními daty jako s reprezentativními.

Design/metodologie/přístup – Stěžejní metodou pro zpracování článku byla datová analýza indexu, tj. seznamu všech digitálních objektů českého webového archivu (Webarchivu) Národní knihovny ČR, a vstupních proměnných při tvorbě archivních dat. Konkrétně byla zkoumána jejich provenience, autenticita nebo obsah. V neposlední řadě pak i technická stránka věci, kterou je například nastavení sklízečů. Analýza vychází z praxe a proběhla nad reálně sklizenými daty.

Výsledky – V článku jsou shrnuty faktory, které ovlivňují výslednou podobu archivních dat. Zaprvé jsou to faktory, které mají dopad na sběr dat, což jsou technická nastavení, strategie výběru zdrojů, tzv. Collection policy, a legislativa. Zadruhé se jedná o nakládání s archivními daty, a to zejména o pravidla pro jejich mazání a omezování přístupu k obsahu. V článku je dále popsána analýza indexu webového archivu, která přinesla kvantifikovaný pohled na archiv a ukázala počty digitálních objektů, procentuální zastoupení souborových formátů, složení domén a vývoj archivu v čase.

Originalita/hodnota – Největším přínosem článku je ucelený náhled na data uložená ve Webarchivu, jakým způsobem vznikají a co jejich vznik ovlivňuje. Toto je stěžejní pro všechny potenciální badatele, kteří mají zájem pracovat s daty Webarchivu a kteří potřebují znát provenienci dat pro svůj výzkum.

Klíčová slova: archivace webu, Webarchiv, big data, vytěžování dat, datová analýza, digitální archivace, webové zdroje, metody webové archivace

Abstract

Purpose – The article provides an overview of possible input criteria when archiving web pages through web archives and describes what impact their settings may have on the resulting archive data in the content, format, and technical plane. Setting the input parameters for web archiving directly determines the resulting form of archive content, and if research is done over these data, researchers need to know the source of the data. Without this knowledge, it is not possible for researchers to use archival data as representative source.

Design/Methodology/Approach – The basic method for article processing was data analysis of the index, i.e. the list of all digital objects of the Czech Web Archive (the Webarchiv) of the National Library of the Czech Republic, and the input variables in the creation of archival data. Specifically, their provenance, authenticity, or content was investigated. Furthermore, the technical side of

the archiving concerns, for example, the setting of the harvesters. The analysis is based on experience and was performed with the actual harvested data.

Results – The article summarises the factors that influence the resulting form of archive data. First, there are factors that directly affect data collection, such as technical settings, resource Collection policy, and legislation. Second, there are factors concerning the handling of archive data, in particular rules for deleting and limiting access to content. The article also describes web archive index analysis that brought a quantified view of the archive and showed the numbers of digital objects, layout of file formats, domain composition, and archive development over time.

Originality/Value – The greatest benefit of the article is a comprehensive overview of the data stored in the Webarchiv, how they are created and what affects their creation. This is crucial for all potential researchers who are interested in working with Webarchiv data and who need to know the source of the data for their research.

Keywords: web archiving, Webarchiv, big data, data mining, data analysis, digital archiving, web resources, web archiving methods

Úvod

I když internet nebyl primárně určen k archivaci (ať už s ohledem na strukturu, rozsah nebo technické parametry), jsou veškeré snahy o zachycení jeho obsahu v současné době nutností. Jde totiž o jeden z pramenů široké škály informací o současné společnosti. S postupným přesunem stále více lidských aktivit do online prostředí je proto nutné sledovat a uchovat významné, ale i na první pohled zdánlivě nepodstatné informace. Například weby vládních institucí, blogy nebo třeba reklamy. Všechny tyto informace mohou mít pro budoucí výzkumníky nevyčíslitelnou hodnotu. Dalším důvodem, proč archivovat web je bezesporu krátký poločas rozpadu informací. Výzkumy uvádí, že životnost webové stránky je okolo 100 dní. Poté jsou informace upraveny, aktualizovány, mazány nebo přesunuty jinam, a mohou tak být nenávratně ztraceny (Shein, 2016).

Rostoucí tendence v oblasti archivace webu jsou jasným důkazem důležitosti této problematiky. Jednou z prvních institucí, která se začala soustavně věnovat archivaci webu, je nezisková organizace Internet Archive. Jde o digitální knihovnu, která obsahuje archivní kopie webových stránek, digitalizované knihy, audio a video záznamy, obrázky i softwarové programy. Organizace byla založena roku 1996, je určena výzkumníkům i široké veřejnosti a jejím hlavním cílem je zprostředkovat přístup k celkovému vědění, které se vyskytuje nebo vyskytovalo na internetu (Gomes & Costa, 2014).

V rámci České republiky archivaci webových stránek zajišťuje webový archiv Národní knihovny České republiky (Webarchiv). Od roku 2000 Webarchiv plní funkci digitálního archivu webových stránek, který vznikl za účelem shromažďování, ochrany, zpřístupnění a dlouhodobého uchování informací pro budoucí generace. Obsahem archivu jsou dokumenty s bohemikálním charakterem a webové stránky s českou doménou.

Autenticita a provenience dat

Pokud bychom se detailněji věnovali definici webové archivace, tak jde o archivaci digitální. Pro archivaci webu je zásadní předmět archivace a způsob akvizice. Předmětem archivace jsou výhradně dokumenty, které jsou volně přístupné na internetu. Akvizice těchto digitálních objektů probíhá formou jejich automatického získávání z internetových serverů, které se nazývá sklízení (harvesting). Vzhledem k rozsáhlosti webu je při jeho archivaci klíčová otázka výběru (Cubr, 2010).

Kritéria výběru budou podrobněji popsána v následující kapitole – Strategie výběru zdrojů (Collection policy). V této části se budeme věnovat rovněž důležitým vlastnostem, a to provenienci a autenticitě sklizených dat.

Pokud mluvíme o provenienci dat v archivnictví obecně, řešíme původ dokumentů, informací a souvisejících metadat. V souvislosti s tím je nutné zmínit provenienční princip, který spočívá v respektování původu, zachování celistvosti a organické struktury dokumentů a informací (Hartig, 2009). Totéž platí i v oblasti webové archivace. Pro výzkumníky je původ dat velmi důležitý. Pokud mají dostatek metadat, může být zdroj dat, se kterým pracují, považován za relevantní, a zvyšuje tak reprezentativní hodnotu celého výzkumu (Brügger & Schroeder, 2017).

Obecně se webové archivy stále častěji zapojují do diskuze o strategii sběru a způsobech přístupu k datům, které by byly pro výzkumníky co nejpřívětivější. Díky tomu se zde otevírají nové možnosti spolupráce mezi výzkumníky a webovým archivem. Ten je totiž bohatým zdrojem dat pro budoucí výzkumy. Aby mohl být tento potenciál plně využit, je nutné vytvořit standardy a nástroje, které vzájemnou spolupráci umožní (Brügger & Schroeder, 2017), což je hlavním cílem projektu Vývoj centralizovaného rozhraní pro vytěžování velkých dat z webových archivů, který je financovaném z programu na podporu aplikovaného výzkumu Ministerstva kultury ČR NAKI II. Projekt je realizován v součinnosti tří institucí: Národní knihovny ČR, Západočeské univerzity v Plzni a Sociologického ústavu AV ČR. Cílem projektu je zpřístupnit data z Webarchivu pro badatele a podpořit jejich výzkumné záměry.

Druhou vlastností archivovaných webových stránek, které je nutné věnovat pozornost, je jejich autenticita. Autenticita elektronických archiválií, mezi které můžeme řadit i webové stránky, znamená, že: „dokument je tím, za jaký se vydává, nebyl zfalšován ani porušen“ (Cubr, 2017). Je také nutné zmínit, že digitální technologie s sebou nesou mnohem větší riziko záměrného i neúmyslného ohrožení autenticity. Ta je u elektronických dokumentů ohrožena kdykoli jsou přenášeny prostorem (odesílání mezi systémy a aplikacemi) nebo časem (během ukládání informací, při aktualizaci nebo náhradě softwaru nebo hardwaru, při jejich zpracování nebo komunikaci). Při posuzování autenticity archivních kopií webových stránek je tedy nutné zohlednit způsob, jakým byly právě tyto dokumenty spravovány v okamžiku jejich vytvoření.

Webový archiv rovněž musí zavést a dodržovat postupy, které zajistí, že autenticita dokumentů nebude ohrožena například vytvářením kopií. Dále také vytvářet dokumentaci, která popisuje způsob uchování záznamů v průběhu času a způsob vzniku kopií (Cubr, 2017).

Strategie výběru zdrojů (Collection policy)

Každý archiv, i ten webový, by měl být budován s určitou vizí, a tato vize přímo ovlivňuje výslednou podobu archivních dat. V případě Webarchivu jde o uchování národního kulturního dědictví v podobě webových dokumentů bohemikálního charakteru. V obecné rovině naplňuje Webarchiv poslání Národní knihovny ČR, kdy bohemikálním charakterem webu jsou myšleny zdroje, které se k území dnešní České republiky vztahují teritoriálně, autorsky, jazykově nebo obsahově.

Cílem webové archivace je výběr, uchovávání a zpřístupnění dat uživatelům čili budování komplexního fondu digitálních zdrojů. Webarchiv se v tomto případě řídí dokumentem Strategie budování sbírky Webarchivu (ang. Collection policy), který hlouběji popisuje hlavní cíle Webarchivu, typy sklizní, kritéria výběru dokumentů i definici bohemikálního webu.

Ke stanoveným cílům patří pravidelné sklízení webových zdrojů, zpřístupnění tohoto fondu, zajištění dlouhodobého uchování a trvalého přístupu ke všem archivovaným zdrojům a v neposlední řadě také kontinuální tvorba a organizace fondu za účelem zajištění vyhledávání.

Webarchiv využívá k akvizici zdrojů tři druhy přístupů. Celoplošné sklizně jsou zaměřeny na archivaci všech webových stránek zveřejněných na doméně .cz, jejichž seznam je dodáván agenturou CZ.NIC. Tento druh sklizní je prováděn jednou až dvakrát ročně, přičemž z kapacitních důvodů jsou archivované stránky sklízeny do menší hloubky než další dva typy sklizní. Jak již bylo zmíněno v úvodu, cílem je zachycení obrazu českého internetu v daném čase.

Výběrové sklizně probíhají nad vybranými zdroji, přičemž důraz je kladen na zachycení zdroje a jeho změn v celém rozsahu. Jedná se o hodnotné zdroje napříč tématy, které jsou vybírány kurátory Webarchivu s cílem vytvořit veřejně přístupný vzorek českého kulturního dědictví obsaženého na internetu. Budování této kolekce využívá metody Konspektu, tj. přidělení a rozřazení dle předmětových kategorií. Zdroje vybrané do výběrových sklizní jsou posuzovány především na základě svého obsahu. Jde o zdroje s kulturní, vědeckou či historickou hodnotou, které disponují originálním a unikátním obsahem a mají dlouhodobou badatelskou hodnotu. Podstatným kritériem je přístupnost zdroje. Je nutné, aby byl celý zdroj, popřípadě jeho podstatná část obsahově přístupná. Dalším aspektem je technická povaha zdroje, tedy aby webovou stránku bylo možné sklidit v co nejvěrohodnější podobě.

Posledním typem sklizně jsou tematické kolekce, které jsou vytvářeny za účelem zachycení určité události nebo tématu, mající v prostředí internetu širší ohlas. Jejich cílem je uchovat důležité informace o mimořádných událostech, jako jsou například volby nebo povodně. (Kvasnica, 2015). Jedná se o zachycení tématu v online zdrojích, typicky např. významných ekonomicko-společensko-vědních událostí. Monitorování a sklizeň zdrojů probíhá v několika fázích. Před samotnou událostí, v jejím průběhu a po ukončení. Kromě událostí vybíraných kurátory Webarchivu se jedná i o události, jejichž sklízení je koordinováno mezinárodním konsorciem IIPC (International Internet Preservation Consortium, tedy Mezinárodní konsorcium pro uchovávání internetu).

Technické parametry sklízení

Vliv na podobu archivních dat Webarchivu mají jak konfigurace sklízení, které odpovídají strategii archivace Webarchivu, tak i reálný rozsah a možnost škálování hardwarových prostředků, kterými disponuje Národní knihovna ČR, dále programové prostředky, kterými je samotný sběr uskutečňován, a formáty, ve kterých jsou soubory archivovány a indexovány. Možnosti zobrazovače by teoreticky na kvalitu dat vliv mít neměly, avšak strategie sklízení se zčásti přizpůsobuje zobrazitelnosti sklízených dat tak, aby se maximalizovala hodnota prostředků vložených do získání, uchování, vyhledání a zobrazení informace.

Programové vybavení

Webarchiv ke sklízení využívá program, tzv. headless browser, který umožňuje procházet a sklízet webové stránky bez vizuální interpretace (na rozdíl od běžného prohlížeče) a balíčkovat je do kontejnerů specializovaného formátu WARC (nebo jeho předchůdce ARC). Naším sklízečem (ang. crawler) je Heritrix verze 3.2, který je vytvořený v Javě, konkrétně v prostředí Java Runtime Enviroment (min. verze JRE 1.6) s virtuálním strojem. Heritrix je "open-source, extensible, web-scale, archival-quality web crawler project" (Osborne, 2018). Program je volně šiřitelný a modifikovatelný pod svobodnou licencí Apache Licence 2.0 a byl vyvinut společností Internet Archive. Heritrix díky svému využití v profesionální komunitě je dnes standardním řešením umožňujícím sklízení webů pro potřeby webových archivů. Vzhledem ke svým bohatým konfiguračním nastavením umožňuje zasahovat do samotné procesovací úrovně jeho jednotlivých modulů, čímž je také vysoce kustomizovatelný pro většinu úkolů a strategií webových archivů, bez nutnosti zásahů do kódu.

Jednou z výrazných nevýhod aplikace Heritrix je zastarávání modulů, např. pro archivaci sociálních sítí, jejich vývoj není schopen držet krok s vývojem dynamicky poskytovaného obsahu na internetu.

Pro zobrazování obsahu uživatelům Webarchiv používá aplikaci Wayback Machine, která slouží k prohlížení archivních dat a kterou v roce 2001 vyvinul Internet Archive. Wayback Machine vede index archivních jednotek a umožňuje uživateli přímý přístup k jednotlivým časovým verzím dat.

Rozsah, systémové prostředky a škálovatelnost

Sklizně jsou realizovány s ohledem na požadavky kurátorů, které však musí být sladěny s reálnými kapacitními možnostmi Národní knihovny ČR, možnostmi nastavení konfigurace a scénářů sklízení, což samozřejmě ovlivňuje výslednou podobu dat. Plánování sklizní tak probíhá dle přidělené kapacity úložiště pro jednotlivé přírůstkové ročníky, které je svým současným (konec roku 2018) celkovým rozsahem 385 TB nezanedbatelnou položkou v režii Národní knihovny ČR. Průběh sklizní je rovněž regulován a plánován s ohledem na technický stav úložiště a přidělenou šířku síťového pásma, či jiné probíhající úkoly (např. migrace dat).

Data jsou agregována na diskových polích a zároveň archivována na magnetické pásky (typ LTO 5). Roční přidělené kapacity pro sklízení se pohybují průměrně mezi 20 až 40 TB, a tomu se přizpůsobuje objem jednotlivých sklizní. Jejich konfigurace bývá stejná, ale délka trvání se může pohybovat od týdne až do 25 dní, přičemž standardní doba je cca 20 dní. Většinou jsou spouštěny mezi 20. až 30. dnem měsíce. Rozsah sklizně se pohybuje měsíčně v průměru kolem 1,5 TB, přičemž díky deduplikační redukci (viz níže) dosahuje na začátku roku až 3,5 TB a na konci méně než 1 TB. Jedna celoplošná sklizeň představuje cca 8 TB až 20 TB dle poskytnutých systémových prostředků.

Termín konání celoplošné sklizně bývá variabilní, obvykle v polovině kalendářního roku a pak v průběhu prosince. Sklizeň je tedy realizována ideálně dvakrát do roka, minimálně jednou, přičemž probíhá v kooperaci nebo alespoň s informovaným souhlasem ostatních systémových techniků. Celoplošná sklizeň zahrnuje výhradně weby na doméně .cz, která Webarchivu poskytuje její správce CZ.NIC. Při poslední sklizni v prosinci roku 2018 to bylo 1 303 776 domén druhého řádu. Pro takový objem je třeba do procesu sklízení zapojit více sklízečů na vícero strojích tak, aby se dostatečně rychle stihla sklidit česká doména do požadované hloubky. Škálování je důležitou součástí plánování sklizní. Sklízení je možné provozovat pomocí jednoho nebo více sklízečů, na jednom nebo více strojích, jak statických (fyzických i virtuálních), tak i pomocí virtualizovaných prostředků nasazených dle potřeby a vždy s unikátní konfigurací.

Virtualizace umožňuje konsolidovat prostředky organizace a plánovaně je poskytovat v rozsáhlé míře jednotlivým naplánovaným úkolům s prioritou, nebo nevyužívané prostředky přidělovat na vyžádání. Oproti tomu výhodou statického nastavení je jistota přidělených prostředků k hospodaření.

Webarchiv používá vždy jeden stroj pro jeden sklízeč. Momentálně má trvale k dispozici jeden virtualizovaný stroj s jedním sklízečem, pomocí kterého realizuje většinu výběrových sklizní, v případě potřeby postupně přidává jeden až dva stroje s dalšími sklízeči. Celkově má možnost disponovat až s deseti statickými stroji. Historicky největší sklizeň v rámci projektu Národní digitální knihovny byla realizována až 20 stroji, ale už deset strojů postačuje pro základní potřeby celoplošné sklizně.

Konfigurace sklizně

Jak bylo výše zmíněno, Heritrix umožňuje rozmanité konfigurace pro sklízení webů, které jsou ukládány v konfiguračních souborech. Počáteční inicializace sklízeče Heritrix probíhá v souboru start.sh, kde je také definována maximální možná provozní paměť. Heritrix se po počáteční inicializaci řídí konfigurací v dalších souborech. Takovým souborem je negative-surts.txt, v němž je uložen seznam adres, které nesmí sklízeč navštívit. Seznam je přikládán ke každé konfiguraci sklizně a Webarchiv ho má společný pro všechny sklizně, aby byla zachována kontinuita (více v kapitole věnované blokaci obsahu).

Primární konfigurační soubor se jmenuje Crawler-beans.cxml a obsahuje všechna důležitá nastavení pro sklizeň a její moduly. Ačkoliv může každá sklizeň mít rozličná nastavení, tak Webarchiv dosud zachovává kvůli udržení konzistence dat pro všechny sklizně stejná pravidla.

V tomto souboru je i možnost nastavení redukce duplikace. Ta probíhá pomocí porovnávání sklizených dat s archivními. Webarchiv má pro porovnávání dat nastavený 1-2letý cyklus, což zaručuje zabezpečení dostatku duplicitních dat pro případ jejich ztráty a zároveň snižuje riziko zahlcení archivu stejnými daty.

Robot exclusion standard

Další z konfigurací, které mají signifikantní dopad na výslednou podobu archivních kopií webových stránek, je práce s Robot exclusion standardem, známém spíš jako robots.txt, podle názvu souboru, ve kterém je uložen. Pomocí tohoto protokolu vývojáři webových stránek dávají pokyny robotům, jak se mají na jejich stránce chovat ("About /robots.txt", 2007). Roboty je myšlen software, který se automatizovaně pohybuje na internetu a provádí nějakou činnost, typicky to mohou být indexační roboti vyhledávačů, ale také sklízeče webových archivů. Webové archivy po celém světě již dlouhou dobu diskutují o tom, jak k této problematice přistupovat.

Americká organizace Internet Archive, která je největším webových archivem na světě a hybnou silou v dalším vývoji oboru, dlouhou dobu tento protokol respektovala. Nicméně v roce 2017 začala od této praxe upouštět, protože se ukázalo, že vývojáři protokol využívají zejména pro instruování indexačních robotů vyhledávačů, aby mimo jiné i vylepšili svou pozici ve výsledcích, a o archivních sklízečích nepřemýšlejí. Internet Archive došel zkoumáním svých dat ke zjištění, že soubory robots.txt, které jsou primárně zaměřeny na indexační roboty vyhledávačů, nemusí nutně dobře sloužit k archivním účelům (Graham, 2017).

Webarchiv již od svého vzniku robot.txt nerespektuje, a to ze stejných důvodů, kvůli kterým k tomuto řešení přistoupil i Internet Archive. To, jak vypadá web pro roboty, neodpovídá tomu, jak vypadá pro uživatele, a snahou Webarchivu je zachovat otisk webu v jeho co nejvěrnější podobě.

Hloubka sklízení

Hloubka sklízení určuje, kolik položek může sklízeč stáhnout z jednoho webu. Položky chápeme jako vše, co stránka obsahuje a co má své URL. Nejde jen o vizuální a skliditelný obsah, ale i o styly stránek, skripty a další podpůrné technické soubory. V případě, že se jedná o příliš mnoho odkazů vedoucích mimo stránku, jsou zde nastavení, která umožňují jejich ignorování.

Pro pravidelné výběrové měsíční sklizně se nastavuje hloubka sklízení pro každý jeden sklízený web paušálně, a to na 15 000 položek. Výjimku mohou tvořit tzv. big budget stránky, kde jich může být povoleno až 60 000, a naopak u tzv. small budget stránek dochází k redukci na 5600 položek. U celoplošné sklizně, kde postačuje základní zachycení charakteristiky webu, je to jenom 5000 položek.

Další nastavení jsou ta, která umožňují blokovat nežádoucí obsah: u Webarchivu jsou to především fóra, kalendáře, případně tzv. local traps (místní pasti) v podobě podsekcí s komentáři, diskusemi, antispamovými moduly apod. Tato filtrace probíhá pomocí regulárních výrazů, které sklízený obsah filtrují na základě url.

Těchto pravidel může být daleko více a souhrnně se nazývají Decide Rules, neboli rozhodovací pravidla. Tato pravidla jsou uložena u sklizně v záznamu činnosti (dále logy, z ang. logs). Uložení archivních dat je strukturované v adresářích rozdělených po jednotlivých letech. Každý adresář obsahuje podadresáře typů sklizní a ty obsahují adresáře jednotlivých sklizní, v nichž jsou uloženy ARC/WARC kontejnery. Až od roku 2012 adresář obsahuje i logy, nastavení dané sklizně a indexy. Pro starší data jsou logy uloženy buď mimo hlavní strukturu, nebo se nezachovaly vůbec.

V tomto souboru se také nastavuje škálování sklizně, tedy jsou v něm uložena pravidla pro distribuci webů mezi jednotlivými sklízeči, v případě že je sklizeň realizována více než jedním.

ARC/WARC

ARC a WARC jsou ústřední formáty webové archivace, ve kterých jsou uloženy archivní kopie webových stránek. Implementace staršího ARC formátu byla vyvinuta v rámci Internet Archive v roce 1996 Mikem Burnerem a Brewsterem Kahlem (Kahle & Burner, 1996). Nahradil ho novější WARC, který vznikl ve spolupráci s konsorciem IIPC, a v roce 2009 se stal normou ISO 28500. V roce 2017 byl formát povýšen na WARC 1.1, odpovídající normě ISO 28500:2017.

Zjednodušeně jde o kontejnerové formáty umožňující pomocí jednoduché struktury uskladňovat obrovské objemy dat a operovat s nimi. V zásadě mají hlavní identifikační hlavičku, za kterou následují záznamy, podobně jako v kartotéce. Ty mají standardizovaný formát, ale zároveň se do jejich „obalu" vejde leccos. Nicméně způsob jak zapisovat do kontejnerů může být také konfigurovatelný v crawler-beans.cxml. Každý WARC má v rámci politiky Webarchivu své jméno následující jmennou konvenci. Počáteční část názvu souboru umožňuje zvenku, bez dalšího rozbalování a čtení, identifikovat operátorovi příslušnost kontejneru ke sklizni.

Zobrazení dat uložených ve WARCích předchází proces indexace, která připraví formální seznam záznamů ve formátů CDX, více o indexu v kap. Datová analýza indexu webového archivu.

Pravidla pro omezování archivace a přístupu k obsahu

Webarchiv se snaží udělat maximum proto, aby jeho archivní obsah mohli uživatelé považovat za autentický. V předchozí kapitole bylo popsáno jedno z takových opatření, a to jasně definovaná a transparentní akviziční politika. Ale snad i důležitější opatření představují nastavená pravidla pro další manipulaci s archivním obsahem. Uživatelé musí vědět, že s obsahem archivních kopií není nijak manipulováno, a zároveň musí znát pravidla pro řešení nejrůznějších problémů, ke kterým může docházet.

V této části článku se nejedná o technickou problematiku, jako je fyzická ochrana digitálního obsahu nebo technické problémy, které také mohou omezovat archivaci nebo přístup k obsahu (havárie úložiště, dočasná nedostupnost serverů apod.), ale jde o zákonná, morální nebo etická omezení a o interní pravidla instituce, jakým způsobem zachází s problémovým obsahem.

Důležitou informaci pro badatele představují pravidla pro případné vyřazení webových stránek z archivace, resp. z archivu, nebo pravidla omezování přístupu k webovým stránkám při archivaci, resp. k obsahu v archivu. Základní otázkou tedy je, v jakých případech může dojít k nějakému omezení, co takové omezení vyvolá a jaký to má dopad na uživatele.

V první řadě je třeba upozornit na to, že Webarchiv se snaží jakýmkoli omezením archivace obsahu vyvarovat a předcházet. To znamená, že se snaží co nejméně do tohoto procesu zasahovat a zároveň je jeho akviziční politika pro archivní obsah velmi liberální, tak aby nedošlo k poškození autenticity archivu a aby mohl uživateli nabídnout co nejvěrnější obraz českého internetu.

Webarchiv aktivně nevyhledává a neomezuje problémový obsah, protože to není kapacitně možné, a je tak odkázán na zpětnou vazbu uživatelů. Problémovým obsahem je pak zejména myšlen nelegální obsah, tak jak jej definuje česká a evropská legislativa. Jde o (ne výlučně) materiály spojené s terorismem, projevy nenávistné povahy (hate speech), se sexuálním zneužíváním dětí nebo spojené s obchodováním s lidmi. Ale může se jednat také o porušování práv duševního vlastnictví, nezákonné obchodní praktiky („Tackling Illegal Content Online“, 2017) a případná další porušení zákona.

Nelegální obsah ovšem není jediný z problémů, kterými se musejí pracovníci Webarchivu při své práci zabývat, mohou to být také nejrůznější potíže buď technického rázu (např. sklízeč navštíví nezabezpečené administrativní rozhraní webu), nebo způsobené zásahy na straně vydavatele (přesun domény, vlastníka webu apod.), na které uživatelé narazí.

Mazání obsahu

Webarchiv za žádných okolností již archivovaný obsah nemaže. A to ani u obsahu, který porušuje zákon. O tento typ obsahu pak mívá zájem Policie ČR, soudy nebo jiné podobné instituce, se kterými Webarchiv spolupracuje. Hlavním důvodem je právě zachování autenticity archivních kopií stránek a jejich kontextu. Je potřeba si uvědomit, s jakým typem dokumentu Webarchiv pracuje. Přestože je často web připodobňován k největší světové knihovně, tak tato analogie je neopodstatněná, protože webové stránky nerespektují tradiční uspořádání fondů. Články, zprávy, informační bulletiny a další „izolované publikované položky“ jsou běžnou součástí webových stránek, ale jejich struktura – vztah mezi dokumenty – je také důležitá (Corey Davis, 2014).

Webové stránky jsou navzájem propojeny hypertextovými odkazy a velmi často se stává, že jediná stránka nenese žádnou nebo nekompletní informaci bez dalších dokumentů, na které odkazuje. To znamená, že webové dokumenty na úrovni stránky, ale také na úrovni webu, téměř nikdy samotné nedávají smysl, ale jsou vmíchané ve větší sítě dokumentů (Masanès, 2005).

S tímto souvisí i technická náročnost takového zásahu do archivních dat, protože veškeré odkazy a propojení, které budou vést na smazanou stránku, se ztratí. Navíc webová stránka není jeden celistvý dokument, ale je tvořena sadou digitálních objektů (obrázky, text, skripty atd.). Tyto digitální dokumenty

mohou být uloženy v různých archivních kontejnerech a zároveň jeden digitální objekt může využívat více stránek, např. logo firmy je jeden digitální objekt, který je využíván napříč všemi stránkami na doméně. Pokud by došlo ke smazání nějaké webové stránky, tak všechny archivní kontejnery, kde se nalézají digitální objekty patřící k této stránce, musejí být otevřeny, nově vytvořeny a přelinkovány, s tím rizikem, že hrozí ztráta informací i u jiných webových stránek.

Webarchiv řeší tyto problémy znepřístupněním obsahu. Pokud Webarchiv přijde na závadný obsah, tak vždy přistupuje k jeho blokaci. Blokace má nastavená jasná pravidla a může probíhat na několika úrovních.

Blokování obsahu

V současné době Webarchiv eviduje dva druhy černých listin (blacklistů) a jednu bílou listinu (whitelist). První blacklist slouží pro sklízeče, stránky z tohoto seznamu sklízeč nenavštíví, a tedy nejsou archivovány. Druhý blacklist je pro zobrazovací aplikaci, na tomto seznamu najdeme stránky, které archivovány jsou, ale mají kompletně omezený přístup k archivnímu obsahu. Whitelist pak obsahuje seznam webových stránek, které jsou volně dostupné veřejnosti online, všechny ostatní jsou dostupné pouze z referenčního centra NK ČR.

Pracovníci Webarchivu vždy posuzují každý problémový materiál individuálně a dle závažnosti problému volí vhodný postup. Postup při omezování obsahu pak vypadá následovně:

Prvním stupněm je vyřazení zdroje z whitelistu (pokud je v něm stránka uvedena), tím dojde k zamezení zobrazování webové stránky pro všechny uživatele veřejně online, nicméně stránka je stále přístupná z referenčního centra NK ČR a nadále je i archivována. Tento případ může nastat např. při vypovězení licenční smlouvy ze strany vydavatele, při změně vlastníka domény apod. Jde tedy o problémy s licencí, která umožňuje zpřístupnit webové stránky celé veřejnosti.

Druhým stupněm je zařazení stránky na blacklist sklízeče. Stránka již není archivována, ale stále je dostupná z referenčního centra NK ČR a uživatelé si ji mohou přijít i nadále prohlédnout. Velmi často je jedná o případy, kdy sklízecí robot omylem pronikne do špatně zabezpečené stránky, např. administrační části periodika nebo internetového obchodu.

Třetím stupněm je zařazení stránky na blacklist zobrazovací aplikace. Toto je nejextrémnější řešení, protože takové stránky již nejsou přístupné pro uživatele ani v referenčním centru, ale nejsou běžně dostupné ani pro pracovníky Webarchivu. Ve většině případů je stránka zařazena i na blacklist sklízeče. K tomuto kroku se přistupuje právě při porušení zákona, nejčastěji jde o nahlášení poškození autorských práv ze strany autora.

Datová analýza indexu webového archivu

Předchozí kapitoly článku se věnovaly proměnným, které ovlivňují výslednou podobu dat. Jak ale reálně vypadají data, která se ve Webarchivu nacházejí? Na tuto otázku neexistuje jednoznačná odpověď, jelikož

se ke konci roku 2017 jednalo o více než 300 TB dat. A s takovým množstvím dat už není jednoduché jakkoliv manipulovat a jednoduše je popsat. Proto se jako nejlepší řešení nabízela analýza indexu webového archivu. Ten totiž přináší nejjednodušší a zároveň kompletní náhled na něj, protože se zjednodušeně jedná o seznam veškerého archivovaného obsahu.

V současné době je celosvětově nejvíce využívaný pro ukládání indexu souborový formát s příponou .cdx. Jeho oblíbenost souvisí s tím, že se jedná o formát, který je využíván pro vyhledávání v zobrazovací aplikaci Wayback Machine. Stejně jako tuto aplikaci, tak i tento formát vyvinul Internet Archive a dlouhodobě usiluje o to, aby se z CDX souborového formátu stala mezinárodní norma, podobně jako tomu je u souborového kontejnerového formátu WARC, což by ještě zvýšilo interoperabilitu mezi jednotlivými webovými archivy.

CDX je jednoduchá textová tabulka (podobně jako soubory .csv), v níž jednomu řádku odpovídá jedna URL (digitální objekt). CDX formát se skládá z různých polí, která popisují každý záznam, seřazený dle URL a data zachycení. Index se používá k vyhledávání a zobrazování konkrétních záznamů požadovaných koncovým uživatelem (Blumenthal, 2018).

V našem případě nám pomohli s analýzou indexu kolegové z Katedry kybernetiky Západočeské univerzity v Plzni, kteří jsou také našimi partnery ve výše zmiňovaném projektu. Pomocí analýzy indexu jsme získali konkrétní představu o struktuře archivu, což je jeden z prvních důležitých kroků v projektu, ale zároveň pomůže i potenciálním uživatelům webového archivu formovat představu o struktuře, velikosti a rozsahu archivních dat.

V současné době index Webarchivu dosahuje velikosti přes 4TB a každý měsíc narůstá s novými nasbíranými archivními daty. Analýza probíhala na indexu s archivními daty od počátku archivu do července 2017. Webarchiv využívá pro vytvoření indexu standardní nástroje přiložené k aplikaci Wayback Machine.

V následující tabulce jsou vidět základní souhrnné údaje vypočítané z indexu.

obsahuje záznamy od:	3.9.2001
obsahuje záznamy do:	9.7.2017
celkem řádků:	8 642 394 282
validních objektů:	8 640 501 571
validních text/html objektů:	4 911 913 105
počet různých domén I. řádu:	1 934
počet různých domén II. řádu:	2 177 229
počet různých subdomén:	4 147 240

Tab. 1 Základní statistiky indexu

Z tabulky č. 1 je patrné, že první archivní data, kterými Webarchiv disponuje, jsou ze září roku 2001. Webarchiv vznikl jako projekt v roce 2000 a s první archivací začal právě o rok později. Nejstarší data jsou uložena ještě ve starším formátu ARC, který pak nahradil vyspělejší WARC.

Dalším z důležitých údajů je celkový počet řádků, který se rovná počtu unikátních digitálních objektů. Digitálním objektem v tomto kontextu je myšlen počítačový soubor, který má vlastní unikátní URL a je uložen v archivu. Může se stát, že dojde k chybě a URL v indexu je nekompletní, nebo řádek indexu nemá správný počet sloupců. Proto byly při analýze nevalidní objekty vyfiltrovány, a vyplynulo, že Webarchiv obsahoval do 9. 7. 2017 více než 8,6 miliardy (8 642 394 282) unikátních digitálních objektů, z čehož bylo 189 271 objektů nevalidních. Tyto objekty byly z dalších analýz vyřazeny. Z celkového počtu validních digitálních objektů je více než polovina souborů (4 911 913 105) ve formátu HTML, který je dodnes základním jazykem pro tvorbu webových stránek.

Přestože je Webarchiv českým webovým archivem a většinou se zaměřuje na doménu .cz, i když ne výhradně, analýza ukázala, že obsahuje 1 934 unikátních domén prvního řádu (např. .cz, .com). Při analýze byla využita metoda srovnání záznamů z indexu se seznamem domén prvního řádu ve veřejné databázi publicsuffix.org.

Počet domén druhého řádu (např. nkp.cz), kterých bylo více než 2 miliony (2 147 240), byl počítán v kombinaci s doménou prvního řádu, jedná se tedy spíš o počet unikátních kombinací domén prvního a druhého řádu. A konečně počet unikátních domén třetího řádu (aleph.nkp.cz) je více než 4 miliony (4 147 240).

Tabulka č. 2 ukazuje srovnání licencovaného obsahu s nelicencovaným. Licencovaným obsahem se myslí webové stránky, se kterými máme uzavřenou licenční smlouvu, nebo jsou vystaveny pod svobodnou licencí a zároveň jsou zařazeny do výběrových sklizní. Srovnání licencovaného obsahu s nelicencovaným je nejen důležité pro další vývoj projektu, ale také pro vytvoření dalších možností pro přístup badatelské komunity k archivu. Vzhledem k tomu, že Webarchiv má licenci ke zpřístupnění těchto dat, analýza měla zjistit, jak moc se licencovaná data odlišují a zda by se dala využít jako reprezentativní vzorek pro výzkumné účely.

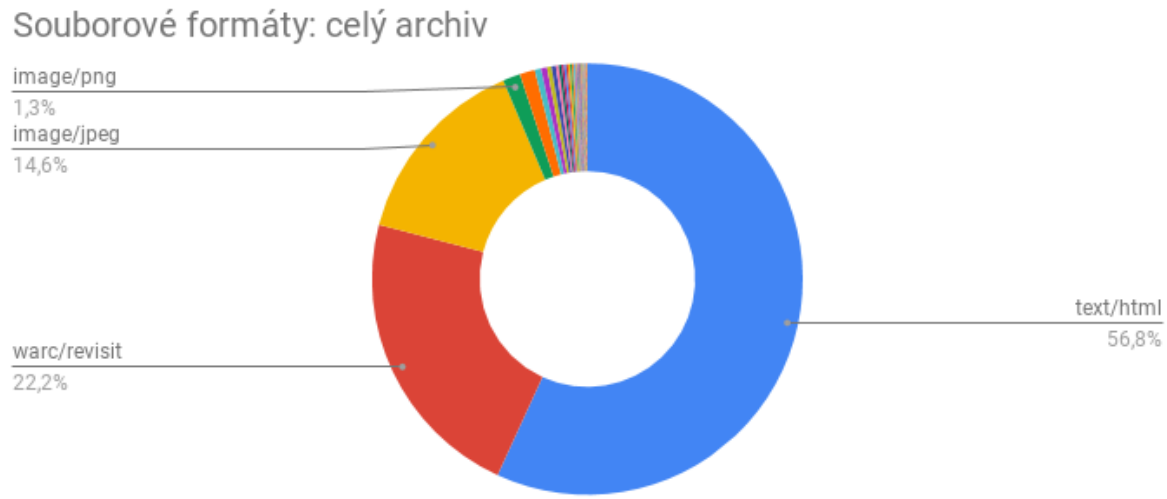
Index	**celkem**	**podíl v celém archivu**
validních objektů:	437 150 778	**8,17 %**
validních text/html objektů:	258 729 862	**7,71 %**
počet různých domén I. řádu:	25	**1,40 %**
počet různých domén II. řádu:	4 269	**0,20 %**
počet různých subdomén:	5 324	**0,13 %**

Tab. 2 Srovnání licencovaného a nelicencovaného obsahu

V tabulce č. 2 je vidět značné snížení počtu domén prvního řádu, jehož příčinou je kurátorský přístup k výběru webů do výběrových sklizní a zaměření na bohemikální zdroje. Dále je zde daleko nižší počet domén druhého řádu, který ale odpovídá počtu uzavřených licenčních smluv. Naopak vyšší frekvence archivace je příčinou toho, že ačkoliv se jedná pouze o 0,2 % počtu domén druhého řádu, je počet archivovaných validních objektů daleko vyšší, konkrétně 7,71 %.

Souborové formáty

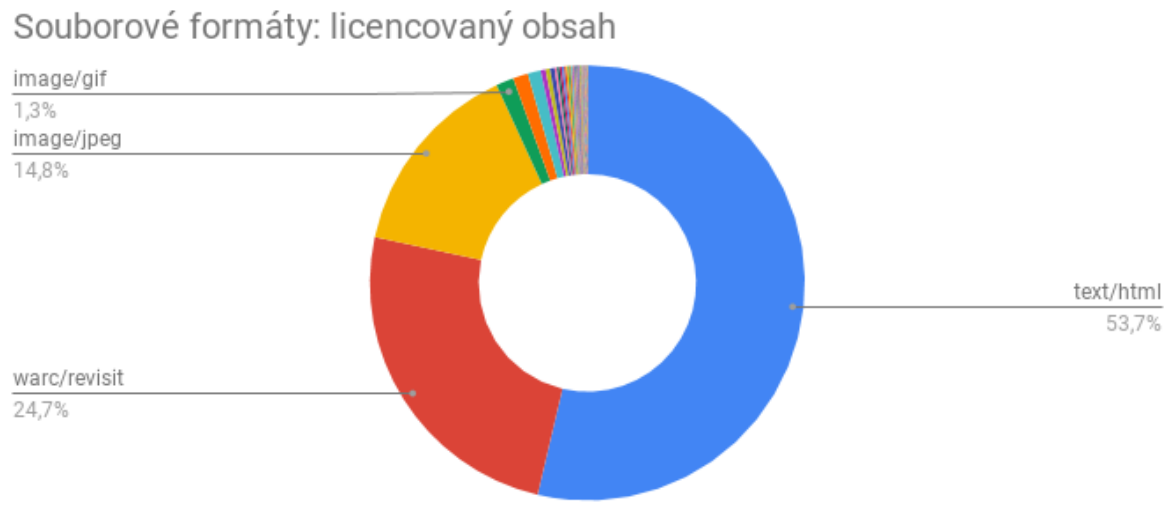
Následující graf (graf č. 1) znázorňuje zastoupení souborových formátů v archivu. Jde o porovnání identifikátorů souborových formátů na internetu, tzv. MIME typů. Tyto identifikátory se ukládají při archivaci a jsou také uloženy v indexu. Identifikátory MIME type nemusejí vždy odpovídat realitě, protože soubory na internetu mohou deklarovat jiný typ, než kterým opravdu jsou, a proto toto srovnání nemusí být přesné. Pro identifikaci souborových formátů existují sofistikovanější metody, které jsme popsali v článku Formátová analýza sklizených dat v rámci projektu Webarchiv NK ČR (ProInflow, 2013). Tyto metody jsou výpočetně velmi náročné a je vhodné je využít pro účely, které přesnost vyžadují, např. pro dlouhodobou ochranu dat. Nicméně jako základní pohled na povahu dat je metoda porovnání MIME typů plně dostačující.



Graf 1 Souborové formáty: celý archiv

Obdobně jako v roce 2013 se ukazuje, že drtivou většinu webu tvoří pouze několik málo formátů. Na grafech můžeme vidět, že 93,6 % archivu tvoří pouze tři souborové formáty. Více než polovinu představují HTML soubory (56,8 %) a 14,6 % obrázky ve formátu JPEG. Warc/revisit záznamy, které dosahují 22,2 %, jsou záznamy, které vznikají při tzv. deduplikaci, což je metoda odkazování na identická, v minulosti již archivovaná data. Díky této metodě Webarchiv dosahuje velké úspory úložných kapacit a místo opakovaného ukládání identického souboru je vytvořen odkaz na jeho první archivní kopii. Tento odkaz je uložen v souboru warc/revisit.

Při porovnání licencovaného a nelicencovaného obsahu pak dochází u licencovaného k mírnému nárůstu warc/revisit záznamů, který je způsoben vyšší frekvencí sklízení, a tudíž i navýšením počtu duplicit (graf č. 2).



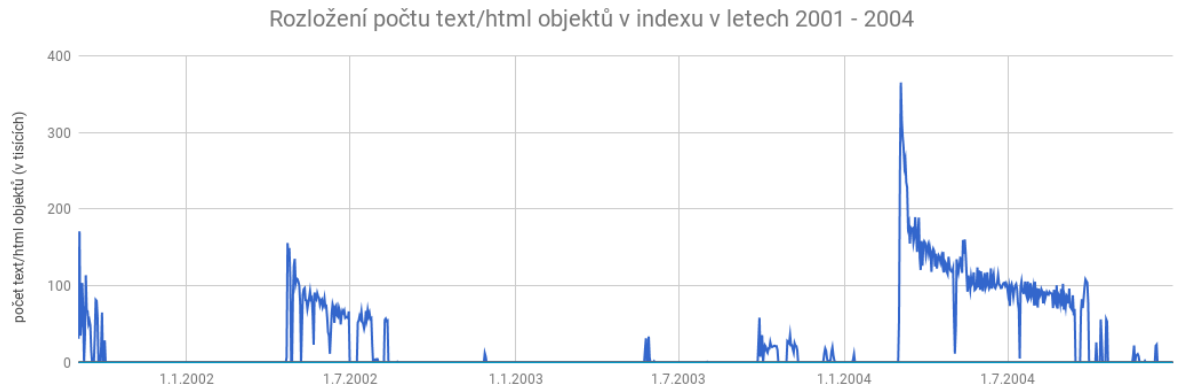
Graf 2 Souborové formáty: licencovaný obsah

Ačkoliv by se z uvedených údajů mohlo zdát, že web je prakticky homogenním prostředím, opak je pravdou. Následujících 6,4 % tvoří stovky až tisíce nejrůznějších souborových formátů, přesné číslo je velmi složité touto metodou zjistit, protože se zde objevuje velká chybovost (překlepy, špatně vygenerované MIME typy atd.), a zároveň není možné ověřit, zda se opravdu jedná o existující souborový formát. Nicméně zastoupení více než 1 % dosahují pouze formáty PDF a GIF. Na dalších místech jsou pak formáty, které tvoří webové stránky, jako jsou soubory s Javascriptovými skripty nebo soubory s kaskádovými styly CSS.

Rozložení textových dokumentů v čase

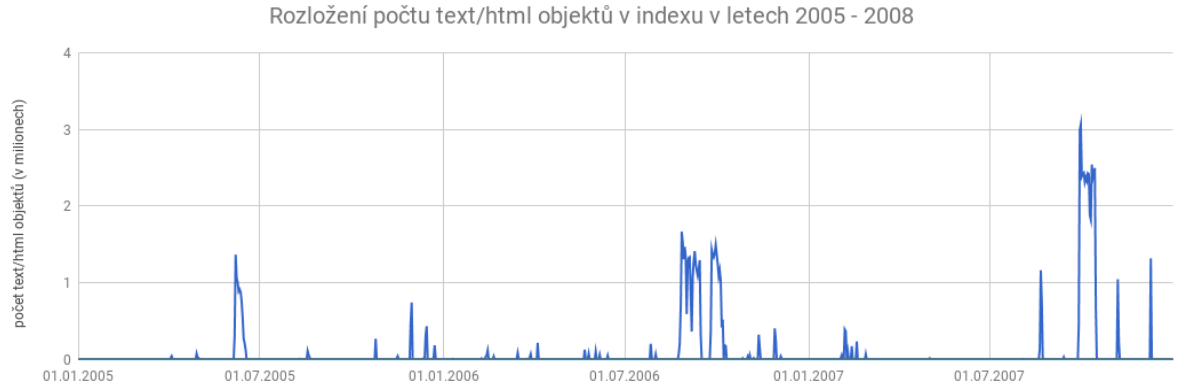
Na základě předchozích zjištění se analýza dále zaměřila na textovou část archivu. Text, resp. text/html, představuje více než polovinu archivu. Zároveň přináší nejvíce možností dalšího zpracování s využitím již existujících nástrojů. V dalších krocích pak Webarchiv může vytvořit nejen fulltextové vyhledávání, ale společně s badateli provádět hlubší analýzy, jako analýzu sentimentu, síť odkazů mezi weby nebo nejrůznější lingvistické analýzy.

Přestože následující grafy ukazují pouze rozložení textových dokumentů, i tak je možné si z tohoto vzorku vytvořit jasnou představu o nárůstu velikosti Webarchivu za dobu jeho existence. Grafy jsou rozděleny do jednotlivých období pro jejich lepší čitelnost. V prvním grafu je zachyceno rozmezí let 2001–2004 a je to jediné období, kde se počty textových objektů pohybují v řádech stovek tisíc. V tomto období Webarchiv začínal a vznikaly první experimentální sklizně.



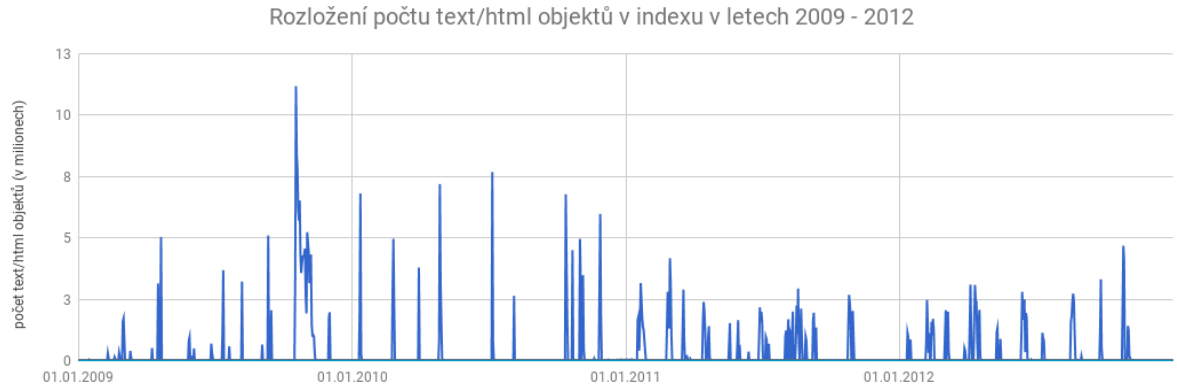
Graf 3 Rozložení počtu text/html objektů v indexu v letech 2001–2004

Od roku 2005 Webarchiv nastavuje pravidla pro archivaci podobná těm, jaká používá dnes a nastavuje jednotlivé typy sklizní: výběrové, tematické a celoplošné, viz kapitola Strategie výběru zdrojů (Collection policy). Za všemi vrcholy v grafech stojí samozřejmě celoplošné sklizně.



Graf 4 Rozložení počtu text/html objektů v indexu v letech 2005–2008

V následujících letech narůstá frekvence a pravidelnost sklizní a s ní i počet archivních objektů. Několikanásobně vzrostla i velikost výběrových sklizní.



Graf 5 Rozložení počtu text/html objektů v indexu v letech 2009–2012

Další nárůst můžeme pozorovat i v následujících letech, kdy zejména celoplošné sklizně dosahují nejvyšších objemů. Výběrové sklizně se drží na podobných hodnotách, což je způsobeno zařazením deduplikace do sběru dat. V letech 2013 a 2014 lze pozorovat mírná stagnace v růstu, která byla způsobena nedostatkem úložného prostoru, který měl Webarchiv k dispozici.

Graf 6 Rozložení počtu text/html objektů v indexu v letech 2013–2017

K faktorům, které ovlivňují růst Webarchivu (a webových archivů obecně), patří samozřejmě rozvoj a rozšiřování internetu, který s sebou nese zvýšený počet webových stránek a větší množství uživatelů, kteří vytvářejí obsah, nebo rychlejší internetové připojení, které přináší nárůst velikosti souborů. Velkou roli hraje i rozpočet instituce provozující webový archiv, v našem případě Národní knihovny ČR.

Skokový nárůst, který můžeme pozorovat zhruba od roku 2009, způsobuje posílení infrastruktury, která umožnila Webarchivu sbírat více dat v kratším čase. Více dat samozřejmě vyžaduje více úložných kapacit a větší nároky na rozpočet. Od roku 2016 je pak vidět pravidelnost archivace a jen mírný růst u pravidelných výběrových sklizní, který je důsledkem větší orientace na tematické sklizně.

V roce 2015 je na grafu patrná anomálie, způsobená velkou sklizní, která byla provedena v rámci projektu Národní digitální knihovna. Tato celoplošná sklizeň byla největší, jakou kdy Webarchiv realizoval, a jak je vidět i na grafu, trvala několik měsíců. Je na ní možné dobře ilustrovat tradiční průběh tohoto typu sklizní, kdy na začátku je stahováno nejvíce dat a postupně pak tento počet upadá. Hned po této sklizni byla spuštěna další několikatýdenní celoplošná sklizeň s deduplikací vůči té předchozí a opět je vidět nejen tradiční průběh sklizně, ale také úspora, kterou deduplikace přináší.

Závěr

První část článku shrnuje teoretická stanoviska, která úzce souvisí s konceptem provenience a autenticity archivovaných dat. Tyto vlastnosti ovlivňuje nejen akviziční politika Webarchivu, která definuje formu sklizní, konkrétně celoplošných, výběrových a tematických, ale také konfigurace a rozsah technického a softwarového vybavení, kterým je archivace reálně zajišťována.

Se zachováním autenticity sklízených dokumentů souvisí pravidla pro manipulaci s archivem. Akviziční politika je nastavena tak, aby uživateli ukázala co nejpřesnější obraz internetu, a proto Webarchiv žádná archivovaná data nemaže. Tato činnost by byla jednak narušením autenticity a je rovněž technicky velmi náročná. Data je možné pouze znepřístupnit. S tím souvisí i možnosti blokace nežádoucího obsahu. Blokace probíhá prostřednictvím tzv. blacklistů a whitelistů, díky kterým je možné korigovat sklízení i zobrazení obsahu koncovým uživatelům.

Druhá část článku se věnuje datové analýze indexu Webarchivu, díky které je k dispozici nový pohled na archivovaná data. Analýza webového indexu proběhla pod záštitou Západočeské univerzity v Plzni a poskytla tak velmi podrobný přehled o archivovaných datech. Zahrnuje informace o celkové struktuře archivu, což je pro zmiňovaný projekt a jeho budoucí směřování stěžejní krok. Index obsahoval data od roku 2001, kdy proběhla první sklizeň, do července 2017. Jeho analýzou bylo například zjištěno, kolik digitálních objektů archiv obsahuje a v jakém formátu se vyskytují nejčastěji (dominují digitální objekty v HTML formátu), dále pak poměrné zastoupení domén prvního a druhého řádu, srovnání charakteru licencovaného a nelicencovaného obsahu, zastoupení souborových formátů nebo rozložení textových dokumentů v čase.

Veškeré poznatky z článku přinášejí unikátní vhled do jinak těžko přístupného Webarchivu umožňující jeho potenciálním uživatelům, zejména z badatelské komunity, udělat si o něm konkrétnější představu a případně i představu o tom, zdali a jak by se dala dále archivní data využít při realizaci budoucích výzkumných záměrů.

Poděkování

Děkujeme kolegům z Katedry kybernetiky ze Západočeské univerzity v Plzni, kteří pracovali na analýze indexu Webarchivu, jejíž výstupy jsou prezentovány v článku. Jmenovitě doc. Ing. Pavel Ircing, Ph.D., Ing. Jan Švec, Ph.D., Ing. Luboš Šmídl, Ph.D. a Ing. Jan Lehečka.

Článek byl realizován v rámci výstupů projektu „Vývoj centralizovaného rozhraní pro vytěžování velkých dat z webových archivů“ financovaném z programu na podporu aplikovaného výzkumu Ministerstva kultury ČR NAKI II.

Bibliografie

About /robots.txt. (2007). Dostupné z: http://www.robotstxt.org/robotstxt.html

Blumenthal, K. (2018). Access Archive-It's Wayback index with the CDX/C API. Dostupné z: https://support.archive-it.org/hc/en-us/articles/115001790023-Access-Archive-It-s-Wayback-index-with-the-CDX-C-API

Brügger, N., Schroeder, R. (2017). The Web as History. UCL Press. Dostupné z: http://discovery.ucl.ac.uk/1542998/

Corey Davis. (2014). Archiving the Web: A Case Study from the University of Victoria. Code4Lib Journal, Iss 26 (2014), (26).

Costa, M., Gomes, D., & Silva, M. (2017). The evolution of web archiving. International Journal on Digital Libraries, 18(3), 191–205. Dostupné z: https://doi.org/10.1007/s00799-016-0171-9

Cubr, L. (2010). Dlouhodobá ochrana digitálních dokumentů. Praha: Národní knihovna České republiky.

Cubr, L. (2017). Autenticita a digitální informace. Praha: Univerzita Karlova v Praze. Disertační práce.

Graham, M. (2017). Robots.txt meant for search engines don't work well for web archives. Dostupné z: https://blog.archive.org/2017/04/17/robots-txt-meant-for-search-engines-dont-work-well-for-web-archives/

Hartig, O. (2009). Provenance Information in the Web of Data. LDOW, 538.

Haškovcová, M., Holoubková, M., Kvasnica, J., & Hrdličková, M. (2017). Akvizice českých webových zdrojů. Acta Musei Nationalis Pragae (Historia), 71(3–4), 41–46.

Kahle, B., & Burner, M. (1996, September 15). Arc File Format. Dostupné z: https://archive.org/web/researcher/ArcFileFormat.php

Kvasnica, J. (2015). Budoucnost českého webového archivu. Inforum 2015. Praha: Národní knihovna České republiky.

Masanès, J. (2005). Web archiving methods and approaches: a comparative study. Library Trends, 54(1). Dostupné z: https://muse.jhu.edu/article/193226/summary

Osborne, A. (2018, July 4). Heritrix 3: Introduction. Dostupné z: https://github.com/internetarchive/heritrix3/wiki/Introduction

Shein, E. (2016). Preserving the Internet. Communications of the ACM, 59(1), 26–28. Dostupné z: https://doi.org/10.1145/2843553

Tackling Illegal Content Online. (2017). Dostupné z: https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:52017DC055