VÝVOJ CENTRALIZOVANÉHO ROZHRANÍ PRO VYTĚŽOVÁNÍ VELKÝCH DAT Z WEBOVÝCH ARCHIVŮ

ZÁKLADNÍ METODICKÁ VÝCHODISKA VÝZKUMNÉHO PROJEKTU ZABÝVAJÍCÍHO SE DATOVÝMI ZDROJI ČESKÉHO WEBU

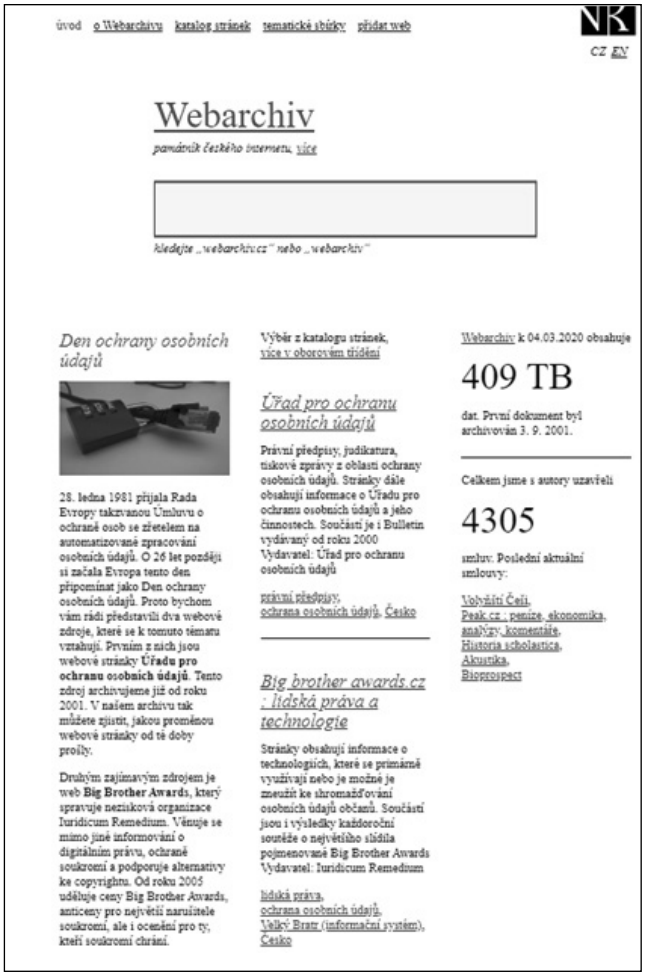
Web jako rozsáhlá a dynamická komunikační platforma představuje jedinečný zdroj, který je klíčový pro současná i budoucí vědecká bádání. Cílem webových archivů je informace publikované v prostředí internetu především získávat a dlouhodobě uchovávat. Zároveň hledají způsoby, jak nasbíraná data zpřístupňovat a smysluplně poskytovat odborné i široké veřejnosti k dalšímu užití. Webový obsah má řadu specifik, s nimiž se webové archivy musí potýkat. Charakterizuje ho proměnlivost (data jsou rychle aktualizována, přesouvána, mazána), propojení obsahu odkazy, velké množství formátů i absence závazných standardů pro tvorbu webových stránek, která znesnadňuje jejich sběr. Na web lze nahlížet jako na nositele kulturního dědictví, zároveň je možné k němu přistupovat jako k souboru velkých datových setů, nad kterými je možné vytvářet analýzy a využít je pro výzkum. Zajímavé jsou nejen pro knihovníky a archiváře, ale i pro profesionály v oblasti digital humanities, datové vědce, vývojáře a další badatele. Vzhledem k tomu, že archivovat veškerý internetový obsah je nemožné, archivy se musí zabývat otázkami, jak obsah určený k archivaci vybírat, jak s daty nakládat v rámci platné legislativy, kdo jsou jejich uživatelé, a stanovit si politiku sběru (způsob profilování sbírek), která je klíčová pro výsledný charakter archivu. Zásadní jsou i vstupní technické parametry při archivaci webových stránek, jejichž nastavení (například respektování nebo ignorování protokolu robots.txt nebo nastavení crawlerů) má na archivní data dopad, nebo pravidla pro nakládání s nimi, zejména pro jejich mazání a omezování přístupu k obsahu. Při posuzování autenticity archivních kopií je proto nutné zohlednit způsob, jakým byly vytvořeny.[1]

ČESKÝ WEBOVÝ ARCHIV

Český webový archiv (Webarchiv)[2] je součástí Národní knihovny ČR a soustředí se zejména na archivaci národního webu, tzn. na dokumenty, které mají bohemikální charakter - k území České republiky se vztahují teritoriálně, autorsky, jazykově nebo obsahově. Sklízí obsah, který je volně dostupný, a aktuálně spravuje více než 400 TB dat. Jeho akviziční politiku tvoří tři linie - jednou až dvakrát do roka provádí tzv. celoplošné sklizně, kdy v daném čase vytvoří otisk webového obsahu na doméně .cz. Kromě toho provádí sklizně výběrové (pravidelná archivace webů zvolených na základě selekčních kritérií se snahou vytvořit vzorek webu napříč všemi oblastmi lidského vědění, opírá se o metodu Konspektu) a tematické (zaměřené na aktuální celospolečenská témata, zdroje určité instituce nebo typ dokumentů)[3]. K výzvám, kterými se zabývají i zahraniční archivy, patří hledání způsobů archivace sociálních médií. Vzhledem k stávajícím legislativním omezením, která umožňují zpřístupnit mimo budovu Národní knihovny ČR jen malou část archivu (buď stránky vystavené pod licencí Creative Commons nebo stránky, s jejichž vydavateli byla uzavřena licenční smlouva) a neumožňují poskytování dat pro vědecké účely, hledá Webarchiv cesty, jak zveřejňovat tu část archivu, která není autorsky chráněna, zejména metadata.[4]

POPIS ZÁKLADNÍCH PROCESŮ WEBOVÝCH SKLIZNÍ

Proces archivace webového obsahu začíná výběrem a sklízením webových stránek. Sklízeč neboli crawler je robot, který webové stránky navštěvuje podle seznamu URL adres, které mají být sklizeny. Frekvence sklízení se u různých zdrojů liší. Může probíhat pravidelně – několikrát za rok, jednou za měsíc, ale třeba i jen jednorázově. Webarchiv například při realizaci tzv. celoplošných sklizní používá URL adresy, které mu poskytuje provozovatel národní domény, společnost CZ.NIC. Jedná se zhruba o 1 400 000 adres, ale jejich počet se s každou sklizní trochu liší – adresy mizí a objevují se nové.

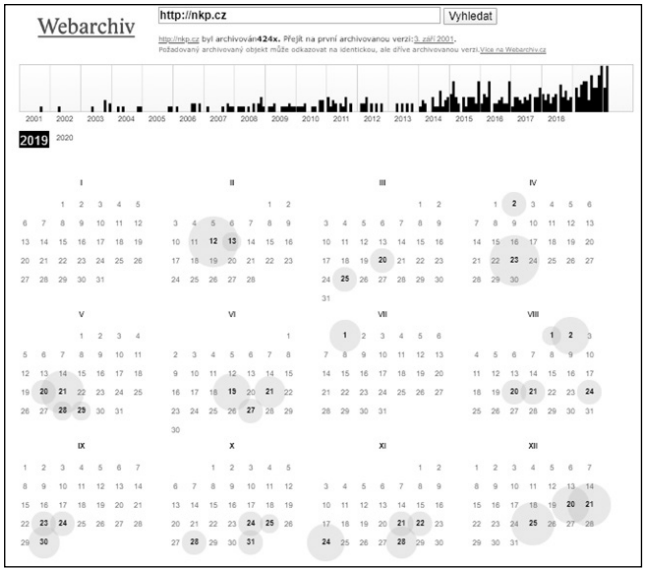


Obr. č. 1: Vstupní rozhraní českého webového archivu.[5]

Na začátku sklizně crawler přichází na stránku a sklidí ji. Pokud jsou na ní další odkazy, navštíví je, také je sklidí a stejným způsobem pokračuje u všech adres ze seznamu. Webarchiv ke sklízení používá nástroj Heritrix (verze 3.4), který je vytvořený v programovacím jazyce Java. Program je volně šiřitelný a modifikovatelný pod licencí Apache Licence 2.0 a vyvinula ho společnost Internet Archive (americká iniciativa, která se snaží archivovat internet v co nejširší míře a má sklizenu i řadu českých webů[6]). Stránky, které sklízí, ukládá do specializovaného kontejnerového formátu WARC[7]. Jde o formát, do kterého se během sklizně ukládají všechny typy souborů ze sklízených webových stránek. Vychází ze staršího formátu ARC, který obsahoval méně metadat. Do těchto kontejnerových formátů je možné ukládat velké objemy dat a operovat s nimi. Veškeré soubory jsou bezztrátově komprimovány. Základní strukturu WARC souboru tvoří hlavička s metadaty a vlastní sklizený obsah. Metadata jsou uložena přímo v kontejneru a vytváří je crawler během sklizně. Metadata nejsou zakonzervována, ale je do nich zanesena každá změna, která se s datovým objektem děje. Velikost jednoho WARC souboru byla stanovena na 1GB. Pokud v jednom kontejneru dojde místo, bude se část stránek ukládat do jiného kontejneru. Proto kontejner obsahuje mnoho fragmentů různých domén. Sklizená, uložená a zaindexovaná data je možné zobrazit. K tomu je nutné využít specializovaný software. V dnešní době je asi nejpoužívanější Wayback Machine[8]. Program je open-source, pochází opět z dílny Internet Archive a slouží k zpřístupnění webových stránek uživatelům. Princip je jednoduchý. Spojuje webový obsah z kontejnerů zpět do webových stránek a umožňuje tak procházení jednotlivými archivními kopiemi napříč časem. Z WARC souboru lze těžit data dle zadaných požadavků. Přístup je možný přes příkazový řádek (pomocí programu Spark), další možností je přístup přes uživatelsky přívětivější rozhraní, které by mělo být výsledkem zmiňovaného projektu.

PŘÍNOS PROJEKTU VÝVOJ CENTRALIZOVANÉHO ROZHRANÍ PRO VYTĚŽOVÁNÍ VELKÝCH DAT Z WEBOVÝCH ARCHIVŮ[9]

Jednou z významných ambicí výzkumného projektu "Vývoj centralizovaného rozhraní pro vytěžování velkých dat z webových archivů[10]" je najít funkční řešení, jak efektivně zpracovávat data českého webového archivu za účelem jejich legálního zpřístupnění široké odborné veřejnosti, a tím posunout potenciál těchto dlouhodobě shromažďovaných a dosud z velké části nezpracovaných dat nad dosavadní rámec použití. Na realizaci projektu se podílejí tři typologicky odlišní partneři - Národní knihovna ČR jakožto instituce odpovědná za sklízení a dlouhodobou archivaci českých webových zdrojů, Sociologický ústav Akademie věd ČR, v. v. i. zastupující progresivní pracoviště zabývající se výzkumem v sociálně vědní oblasti, jež pro své moderní vědecké postupy potřebuje i data ukládána ve webových archivech, a konečně pak přední odborníci Fakulty aplikovaných věd Západočeské univerzity v Plzni mající neocenitelné zkušenosti ve strojovém zpracovávání velkých datových souborů včetně automatického rozpoznávání informací obsažených ve video či audio souborech. Na realizaci technologické části řešení se podílí i externí partner, jímž je společnost InQool, a. s. V návaznosti na zapojené řešitele projektu a smělou ambici dosáhnout nových metod v oblasti propojení a interpretace webových dat lze realizaci projektu rozdělit na tři části – datovou, procesní a badatelskou. Datová část řešení projektu je zaměřena na přípravu dat pro výzkumné analýzy v oblastech relevantních dat pro jasně definované a předem speci-



Obr. č. 2: Zobrazení archivních kopií webové stránky Národní knihovny – aplikace Wayback Machine

fikované badatelské záměry. Pro efektivní způsob práce s takto velkým souborem různorodých dat bude vytvořen integrovaný fasetový a full-textový vyhledávač, který jednotlivým badatelům umožní jednoznačně definovat tu část dat, kterou potřebují pro své výzkumné aktivity. V návaznosti na tyto pre-analytické postupy bude vyvinuta speciální exportní aplikace umožňující extrakci nadefinovaných setů dat. Tato část vývoje pak již přímo navazuje na další rovinu realizace projektu - procesní část. Ta je spojená s vytvořením a následným využitím nových softwarových analytických nástrojů pro strojové zpracování požadovaných setů dat. Jedná se zejména o automatické přiřazení metadat dokumentům s využitím metod rozpoznávání řeči a metod sémantické analýzy textu. Posledním a možná nejdůležitějším pilířem realizace projektu je samotný výzkum aplikovaný na přímých zdrojích webového archivu. Zájem badatelské obce o archivovaná webová data roste s možnostmi práce s nimi. Nutně tak bude docházet k postupné změně náhledu na rozvoj dosud využívané technologické infrastruktury, která již nebude sloužit pouze k uchování primárních zdrojů, ale i k možnostem jejich dalšího analytického zpracování. Pro tuto sondu do využití webových zdrojů byly zvoleny inovativní postupy sociologického výzkumu, pro který tyto zdroje nabízejí neocenitelný zdroj informací. Díky zapojení vědecké komunity dojde k vytvoření obecných tezí sociologického výzkumu, které konkrétně aplikují zpracovaná data, ale zároveň definují, jak mají vypadat požadované výstupy pro potřeby odborných sociologických studií.

Realizace projektu byla rovněž rozdělena do tří časových etap – vstupní, analytické a implementační. V rámci první z nich proběhla vstupní hloubková analýza dat z webového archivu za účelem zjištění současného stavu, prozkoumání dat a úpravy parametrů pro další výzkumné práce. Součástí byla i definice výzkumných otázek sociálně vědního výzkumu včetně stanovení základních metodologických postupů. V rámci druhé fáze realizace projektu, která bude ukončena v prosinci 2020, dojde k extrakci a analytickému zpracování vzorových setů dat, vytvoření strukturovaného indexu a vývoji nástrojů pro sémantickou analýzu textových souborů. V posledních dvou letech pak dojde k vývoji centralizovaného rozhraní pro vytěžování velkých dat z webových archivů, vytvoření grafického uživatelského rozhraní pro vyhledávání a na závěr pak k integraci vytvořených aplikací v předchozích etapách do jednoho systému.

Mimo tyto jednoznačně definované cíle má projekt ambici posunout oblast webové archivace i v dalších rovinách – zejména právní a technické. Ve spolupráci s expertkou na autorské právo tak došlo k vytvoření právní analýzy popisující postavení webového archivu a možnosti využití dat v aktuálním geopolitickém kontextu, včetně vyhlídky na v blízké budoucnosti implementovanou směrnici o autorském právu na jednotném digitálním trhu (EU Directive on copyright in the Digital Single Market)[11]. Tato směrnice může být pro využití digitálních dat zlomová, zejména s ohledem na možnost zavedení výjimek pro textový a data mining pro výzkumné organizace a instituce uchovávající kulturní dědictví. V technické rovině pak v rámci realizace projektu dochází k implementaci nových technologií, které mohou být pro knihovny do budoucna nosné. Zde je třeba uvést zejména využití frameworku APACHE-HADOOP[12] pro podporu clusterování dat. Důležitým krokem mezi přímou extrakcí dat a metadat z archivních balíčků (WARC-ů) prostřednictvím formátu pro přenos dat (JSON) a vytvořením databáze je navržení tzv. intermediary formátu. Dochází v něm k očištění archivních dat od sekundárních informací před zpracováním do databázového modelu. Jeden intermediary formát odpovídá jednomu WARC-u. Po jeho vložení do databáze a po indexaci bude možné pokročilejší vyhledávání či export dat do datasetů.

ZÁVĚR

Všechny výše uvedené cíle projektu spolu úzce souvisí a vzájemně se ovlivňují. Bez jasného badatelského záměru není možné vytvořit nebo vybrat vhodný nástroj pro analýzu dat. A bez zpřístupnění a analýzy dat není možné naplnit případný badatelský záměr. Tento

vstupní koncept je v prostředí využití datových zdrojů webových archivů do jisté míry unikátní, což je již od první fáze realizace projektu oceňováno i na mezinárodním fóru. Poster s názvem "Web Archiving and Social Sciences: Challenges, Problems and Solutions" shrnující základní výzkumný směr realizace projektu představený na významné sociologické mezinárodní konferenci ESRA 2019 dokonce získal ocenění jako nejlepší poster konference, což svědčí o komplexním a inovativním přístupu k řešení dané problematiky.

Poznamky:

[1] KVASNICA, Jaroslav, Andrea PROKOPOVÁ, Zdenko VOZÁR a Zuzana KVAŠOVÁ. Analýza českého webového archivu: Provenience, autenticita a technické parametry. ProInflow: Časopis pro informační vědy [online]. 2019, 11(1), 3 – 21 [cit. 2020-02-24]. ISSN 1804 – 2406. Dostupné z: http://www.phil.muni.cz/journals/index.php/proinflow/article/view/2019-1-2

[2] Webarchiv: památník českého internetu [online]. Praha: Oddělení archivace webu Národní knihovna ČR [cit. 2020-02-24]. Dostupné z: https://www.webarchiv.cz/cs/.

[3] Strategie budování sbírky Webarchivu (Collection policy) [online]. Praha: Oddělení archivace webu Národní knihovna ČR [cit. 2020-02-24]. Dostupné z: https://www.webarchiv.cz/cs/dokumenty.

[4] KVASNICA, Jaroslav, Barbora RUDIŠINOVÁ a Rudolf KREIBICH. Vědecké využití dat z webových archivů. Knihovna: knihovnická revue [online]. 2016, 27(2), 24-34 [cit. 2019-11-20]. Dostupné z: https://knihovnarevue.nkp.cz/archiv/dokumenty/2016-2/Kvasnica.pdf.

[5] https://webarchiv.cz/cs/.

[6] INTERNATIONAL INTERNET PRESERVATION CONSORTIUM - IIPC [online]. International Internet Preservation Consortium, c2019 [cit. 2020-02-24]. Dostupné z: http://netpreserve.org/.

[7] ISO 28500:2009. Information and documentation — WARC file format. 1 st ed., 2009. 28 s.

[8] https://wayback.webarchiv.cz/wayback/*/www.nkp.cz

[9] Projekt evidovaný pod číslem DG18P02OVV016 je financován ze zdrojů dotačního mechanismu Ministerstva kultury NAKI II (více https://www.mkcr.cz/verejna-soutez-61.html). Doba realizace projektu je pětiletá, od roku 2018 do konce roku 2022.

[10] Základní informace o realizaci projektu jsou dostupné na webu https://www.webarchiv.cz/vyvoj/

[11] Více zde https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52016PC0593

[12] Více https://hadoop.apache.org/

Mgr. Marie Haškovcová
marie.haskovcova@nkp.cz
Bc. Andrea Prokopová
andrea.prokopova@nkp.cz
Mgr. Tomáš Foltýn
Tomas.Foltyn@nkp.cz
(Národná knižnica Českej republiky)

EODOPEN – MEDZINÁRODNÝ PROJEKT DIGITALIZÁCIE A SPRÍSTUPNENIA DOKUMENTOV Z 20. A 21. STOROČIA

EODOPEN – medzinárodný projekt digitalizácie a sprístupnenia dokumentov z 20. a 21. storočia. Autori článku a zároveň účastníci projektu uvádzajú základné informácie o európskom projekte EODOPEN (eBooks-On-Demand-Network Opening Publications for European Netizens), ktorý je spolufinancovaný Európskou komisiou v rámci programu Kreatívna Európa a podprogramu Kultúra. Centrum vedecko-technických informácií SR je platným členom medzinárodného konzorcia knižníc, ktoré rozvíja moderné digitalizačné služby pre rôzne používateľské komunity. Projekt EODOPEN sa zameriava na dokumenty 20. a 21. storočia, pričom kladie osobitný dôraz na spôsoby vysporiadania autorských práv.

Za posledné desaťročie sme naprieč Európou zdigitalizovali veľké množstvo dokumentov. Primárne sme sa zameriavali na knihy, ktoré bolo možné rýchlo a bezproblémovo sprístupniť prostredníctvom digitálnych knižníc či špecializovaných portálov. Predovšetkým išlo o historické publikácie (rôzne typy kníh, novín a časo-